CHEMPROJECT, L'UTILISATION DE LA CHIMIOMÉTRIE PAR TOUS

Virginie Rossard[1] & Robert Sabatier[2] & Eric Latrille[1] & Cécile Trédaniel[3] & Jean-Michel Roger[14] & Fabien Gogé[4] & Jean-Claude Boulet[3]

1 INRA, LBE, avenue des Etangs, 11100 Narbonne
2 UFR pharmacie, 5, BOULEVARD HENRI IV CS 1904434967 - MONTPELLIER CEDEX 2
3 INRA, UMR1083 SPO, 2 place Viala, 34060 Montpellier cedex02
4 IRSTEA, UMR ITAP, 361 rue Jean -François Breton, 34196 Montpellier

Résumé.

Pour permettre à un plus grand nombre de pratiquer la chimiométrie, ChemProject (chemproject.org) a développé (1) un MOOC, CheMoocs, pour la diffusion des connaissances théoriques en chimiométrie et (2) un logiciel gratuit et ergonomique, nommé ChemFlow.

CheMoocs est un MOOC, Massive Open Online Courses, sur la chimiométrie qui s'est déroulée du 16 septembre au 25 novembre 2016 avec 1570 inscrits via la plateforme FUN, France Université Numérique. C'est le résultat d'un projet éponyme de deux ans financé par Agropolis Fondation. Ce projet a mobilisé une trentaine de personnes, dont un grand nombre de chimiométriciens francophones. Ce MOOC diffuse la connaissance à l'aide de 21 modules (ACP, prétraitement, régression, PLS, discrimination, robustesse, multibloc, etc) pendant que ChemFlow, le logiciel en permet sa pratique donc d'acquérir des compétences en chimiométrie. Cette application web se base sur une plateforme bioinformatique Galaxy (galaxyproject.org) gratuite où les outils bioinformatiques ont été remplacés par des outils de chimiométrie (ACP, PLSR, PLS-DA, MCR-ALS, ICA, ACOM, EPO, etc). Sur 1570 inscrits au MOOC, 650 comptes chemflow ont été créés. Durant ce mooc, le serveur public de l'inra de toulouse (https://vm-chemflow.toulouse.inra.fr) a répondu à 47000 requêtes via les outils de chimiométrie.

Le caractère très collaboratif de ce projet fait que le congrès JDS est une excellente occasion de faire un point devant toute la communauté des statisticiens Francais.

L'accent sera mis sur les perspectives 2017 avec l'enrichissement des cours CheMoocs, l'évolution du logiciel ChemFlow et le développement de la base de données ChemData ainsi que sur le projet global.

Mots-clés. Chimiométrie, Enseignement de la statistique, Logiciels, Stabilité mathématique, Qualité, fiabilité, Analyse des données, classification.

1 Structure du texte long

Introduction

La chimiométrique vise à extraire des informations à partir des spectres. Ces spectres peuvent être produits à partir de spectromètres infrarouge qui sont largement utilisés dans la recherche universitaire et de l'industrie comme outil de mesure simple, rapide, pas cher et sûr.

Depuis les années 70, la France perd en compétences en chimiométrie car on manque de formation dans le domaine de la chimiométrie, manque d'un outil informatique gratuit et facile à utiliser, et on manque d'une banque de données commune.

Afin de répondre à ces problématiques, le projet ChemProject a vu le jour. Ce projet vise à développer la connaissance et l'utilisation de la chimiométrie. Il est destiné à permettre au plus grand nombre de pratiquer la chimiométrie. Dans ce cadre deux piliers ont été développés : (1) un MOOC, CheMOOCs, pour la diffusion des connaissances théoriques en chimiométrie; (2) un logiciel ergonomique, nommé ChemFlow gratuit et sans connaissance en programmation pour la pratique de la chimiométrie;

CheMOOCs

Suite au constat d'un réel manque de formation dans le domaine de la chimiométrie, l'idée a été de monter une formation en ligne gratuite et ouverte à tous : CheMoocs qui est un MOOC, Massive Open Online Courses, cours en ligne dédiés à la chimiométrie. Après un an de préparation (définir le scénario pédagogique, tourner les films, mettre en ligne, développer le logiciel, etc), la première session s'est déroulé en automne 2016 du 16 septembre au 25 novembre sur une durée de 8 semaines. Ce projet a fédéré un grand nombre d'acteurs dont 24 collaborateurs venant de toute la France de l'INRA, l'IRSTEA, SupAgro l'université de Lille, Brest, Montpellier mais aussi des acteurs du privé comme Ondalys.

Ce MOOC s'adressait aux personnes pour lesquelles les mots « spectroscopie », « chimiométrie » ou « analyse de données multivariées » sont déjà connus et suscitent un intérêt. Qu'elles soient :

- étudiants, pour une remise à niveau avant leur entrée en master ou en thèse impliquant la chimiométrie
- étudiants en mathématiques intéressés par des applications pratiques à l'algèbre matricielle
- stagiaires, niveau master/ingénieur, ou étudiants en thèse, ayant besoin d'utiliser ponctuellement des outils de chimiométrie
- techniciens utilisateurs de spectromètres, souhaitant mieux comprendre les traitements de leurs données
- ingénieurs ou chercheurs développant des méthodes rapides d'analyse, au laboratoire comme sur le terrain
- professeurs pouvant s'appuyer sur ces ressources pédagogiques

Ce MOOC était composé d':

- une première partie introductive pour la première semaine
- une deuxième qui a duré 6 semaines, appelée « tronc commun » qui reprenait les bases de la chimiométrie avec les statistiques « simples », l'ACP (Analyse en Composantes Principales), les prétraitements, la classification non supervisée, le régression non linéaire, les bonnes pratiques de la modélisation et la discrimination ; et enfin
- la huitième et dernière semaine où des parcours « pour en savoir plus » étaient proposés permettant aux apprenants d'approfondir au choix une méthode : optimisation (la sélection de variable, la robustesse), les méthodes de décomposition spectrale (ICA et MCR) et/ou l'analyse multibloc.

Cette formation s'est déroulée via la plateforme FUN (https://www.fun-mooc.fr), France Université Numérique. Cette plateforme est un Groupement d'Intérêt Public,GIP, FUN-MOOC. Celle ci présente actuellement plus d'1 million d'inscriptions et 150 cours sont disponibles. Différents supports pédagogiques ont été proposés pour aider les élèves : des cours en vidéo, des documents pdf, des exercices sous forme de quizz, des logiciels, un forum. Les étudiants et internautes peuvent les suivre de manière interactive et collaborative et à leur rythme.

Chaque semaine du tronc commun, deux méthodes de chimiométrie étaient présentées avec une vidéo explicative d'une quinzaine de minutes en moyenne, repris dans un document détaillé en pdf, des quizz pour tester leurs connaissances et des exercices pour les mettre en application sur le logiciel ChemFlow (le deuxième pilier de ChemProject) afin de pratiquer et de s'entraîner.

Un forum et des rendez vous live hebdomadaire, Webinaire, ont été mis en place afin que les apprenants puissent échanger entre eux et avec les experts. Des webinaires était diffusé hebdomadairement et rediffusé sur YouTube. A cette occasion les experts répondaient en direct aux questions du forum sur le cours diffusé dans la semaine. Effectivement nous avons eu par exemple beaucoup de questions des étudiants du master OPEX à Rennes. On a pu voir également des groupes de travail émergeaient comme on a pu le voir à l'INRA. Un contrôle continu a été effectué au fil des sessions pour permettre aux apprenants de valider les connaissances acquises. Les apprenants souhaitant obtenir l'attestation ont dû passer un examen en répondant à un challenge chimiométrique.

Au bilan nous avons eu 1570 inscrits de toutes origines : principalement des gens résidant en France 60% et 15% qui venaient du Maghreb. La parité a presque était atteinte avec 40% de femmes pour 60% d'hommes. L'âge des inscrits s'étendaient de 18 à 85 ans avec une majorité de trentenaires ayant un cursus de formation supérieur. Environ 300 personnes ont suivi le mooc jusqu'au bout, donc environ 20%. Bien que accessible à tous, ce mode d'enseignement ne convient pas à tout le monde. Livrés à eux-mêmes sans encadrement pédagogique, beaucoup d'élèves abandonnent en cours de route. Sur Coursera, seulement 10% des étudiants inscrits suivent les cours jusqu'à leur terme. Dans ce type de formation il y a aussi des personnes qui viennent picorer des savoirs, chercher juste ce qu'elles veulent sans suivre la totalité de la formation. 127 attestations de suivi ont été délivrés avec succès en répondant au quizz (30% de la note) et à un challenge sur 129 réponses au challenge donc un taux de réussite de 98%.

Fort de ce bilan plutôt positif nous pensons déjà à la suite avec la volonté de proposer de nouvelles sessions de CheMoocs dès l'automne 2017 à minima en reprenant les mêmes modules. Nous sommes toutefois ouverts à d'autres domaines spectraux tel que le moyen infrarouge, l'ultra-violet, le visible, la fluorescence, le Raman, etc. Nous aimerions également améliorer ce MOOC comme la traduction en anglais, un ajout de modules sur des prérequis (qui on pu manquer à certains apprenants) ou encore en développant de nouveaux parcours « pour en savoir plus ».

ChemFlow

Dans le cadre du MOOC, Chemoocs, nous avions besoin d'un logiciel gratuit et facile d'utilisation, c'est à dire sans ligne de commande, et regroupant les méthodes de chimiométrie abordées dans le MOOC. Or ce MOOC abordait un large panel de méthodes de chimiométrie allant des classiques telles que l'ACP, la PLS (PLS-R et PLS-DA) mais aussi des méthodes de décomposition spectrale (ICA, MCR-ALS, etc), de transfert d'étalonnage (EPO, etc) ou encore de l'analyse multibloc (ACOM).

Il a été donc conçu avec deux objectifs :

- Un support pour l'enseignement, comme on vient de le voir au-dessus.
- Un support pour la recherche développement. Chemflow contient des nouvelles méthodes de statistiques, validées, compatibles avec différents formats de données (csv, jdx, dx), ce qui permet de les mettre en œuvre très rapidement. Nous espérons que Chemflow devienne ainsi un benchmark des nouvelles méthodes chimiométriques.

Les spécifications requises avant de débuter le projet ChemFlow étaient les suivantes :

- Un outil gratuit;

- Un outil qui recycle le code de Matlab, Scilab, R, Python et C;

- Un outil accessible via Internet et responsive design accessible depuis n'importe quel écran tel un smartphone.

Voilà pourquoi nous avons choisi comme base de développement Galaxy qui est un outil initialement dédié à la bioinformatique. Cet outil informatique est nommé ChemFlow puisque toutes les fonctions dédiées à la bioinformatique ont été remplacé par nos propres fonctions de chimiométrie. Ces fonctions peuvent être classées en 3 catégories :

- manipulation de nos données;

- statistiques de base;

- chimiométrie qui exécute des méthodes telles que les étalonnages et les classifications.

Ce logiciel est donc bien gratuit et libre, sous la forme d'une application web, accessible à travers un navigateur internet. La chimiométrie est donc praticable à travers un ordinateur, une tablette ou même un smartphone. Il permet d'appliquer toutes les méthodes à travers des outils interfacés qui ne requiert aucune connaissance en programmation informatique.

Ce logiciel requiert une authentification et son interface se présente de la manière suivante :

- le panneau à gauche est un menu avec de liens vers des outils/fonctions notamment chimiométrique, rangés par catégories
- celui de droite, un historique contenant les données importées ou générées par les outils
- et au centre, un affichage de l'interface de l'outil sélectionné ou des données de l'historique
- différents formats de données peuvent être importés : csv, mat, rdata, via une URL.

Dans cet environnement, on peut :

- éditer des fichiers (supprimer des individus ou extraire des variables)
- créer des graphiques interactifs grâce à la librairie GoogleVis du logiciel R
- appliquer des méthodes de chimiométrie classiques : ACP, PLSR, discrimination, etc
- appliquer des méthodes de chimiométrie plus avancées: transfert d'étalonnage, MCR-ALS, ICA, ACOM, EPO, etc
- créer, éditer et exécuter des workflows sous forme graphique
- partager ces worflows, des historiques de traitement entre utilisateurs de ChemFlow
- tracer des opérations donc faire de la science reproductible.

Le deuxième but de ChemFlow était de diffuser des méthodes développées par des chercheurs. Or, les personnes développant ces méthodes utilisent des langages différents. Par exemple, la méthode EPO est codée en Scilab alors que la méthode ACOM est écrite en Octave, clone de Matlab. Nous avons pu les incorporer directement dans Chemflow puisqu'il utilise le moteur Galaxy. Galaxy est une plateforme web, au départ dédiée à bioinformatique que nous avons dérivée pour la chimiométrie. Cette plateforme permet d'intégrer et d'interfacer des codes provenant de nombreux langages. On peut donc intégrer facilement et rapidement de nouvelles méthodes et nous aimerions que ChemFlow devienne un projet encore plus collaboratif où vous pourriez rendre vos méthodes facilement accessibles à tous.

Ce logiciel est une application web ce qui implique :

- aucune installation sur votre ordinateur, votre navigateur suffit !
- vous pouvez vous connecter à votre compte depuis n'importe où du moment que vous avez une connexion Internet.
- vous profitez de la puissance de calcul des serveurs.
- vous bénéficiez de la même version logiciel et des dernières mises à jour en temps réel.
- vos données et vos traitements chimiométriques sont sécurisés.

Il est ainsi installé sur 3 serveurs : (1) un de développement / test à l'EIC de Montpellier, (2) un public à l'INRA de Génotoul (https://vm-chemflow.toulouse.inra.fr) puis (3) un autre à France Grille (GIS) pour l'innovation méthodologique et la sauvegarde.

Sur 1570 inscrits sur la plateforme Fun, 630 comptes ChemFlow ont été créés durant CheMOOCs. Le serveur a répondu à 47000 requêtes qui sont des calculs demandés par l'utilisateur via les outils du logiciel. Ainsi 47000 calculs ont été résolu par le serveur durant le mooc avec une moyenne de 1000 requêtes ChemFlow par jour et une pointe à 4720 le jour du challenge et 20 000 sur le mois de novembre. 50% des apprenants CheMoocs, ayant répondu au sondage, ont utilisé ChemFlow et parmi eux pourtant + de 50% utilisaient déjà un autre logiciel de chimiométrie avant de commencer ce MOOC.

Les perspectives pour ce logiciel sont :

- l'implémentation de nouvelles méthodes de chimiométrie adaptées à de nouveaux types de données (OMICS, spectres de masse,…)
- la connexion à une base de données spectrales (ChemData)
- la création d'une workflowthèque
- l'amélioration continue de l'architecture du logiciel pour une utilisation de plus en plus intuitive.
- un logiciel plus facilement collaboratif et
- un accompagnement permanent durant l'utilisation du logiciel via chemflow@chemproject.org

Conclusion

L'ensemble du projet a été décrite tel que nous avons pu le réaliser jusqu'à aujourd'hui. Le troisième pilier est une base de données spectrales pour agréger et diffuser des données spectrale, nommée ChemData. Cette dernière est cependant encore en développement. Ce projet a été soutenu par un financement d'Agropolis fondation,que nous remercions. Pour l'avenir nous cherchons une autre source de financement sous forme par exemple d'un mécénat d'entreprise comme le fait SupAgro Fondation.