BIG DATA ET INTELLIGENCE ARTIFICIELLE : EXEMPLES D'APPLICATIONS POUR LA FILIERE VIGNE ET VIN

Pascal Neveu[1]; Coraline Damasio[1]
[1] MISTEA, INRAE, l'Institut Agro, Université de Montpellier, Montpellier 34060, France
Contact : Pascal.Neveu@inrae.fr

Le Big Data et l'Intelligence Artificielle (IA) désignent deux tendances technologiques en plein essor. Le potentiel du Big Data et de l'IA repose principalement sur leur convergence pouvant offrir de nombreuses possibilités pour la filière vigne et vin. En effet, dans l'ensemble de la chaîne de production viticole, les données se multiplient avec l'usage de smartphones, capteurs, équipements connectés, satellites, réseaux sociaux ou encore avec l'usage de services informatisés (traçabilité, systèmes d'aide à la décision technique ou économique, informatisation des déclarations…). Que faire de toutes les données produites ? Comment traiter les données pour exploiter au mieux ces informations au profit des acteurs de la filière et des consommateurs ? Une réponse se trouve dans l'association du Big Data aux méthodes d'IA qui vont permettre de produire des connaissances nouvelles, fournir une aide pour prendre la bonne décision ou simplement effectuer un diagnostic. Mais pour être en capacité de développer pleinement ce potentiel, il est crucial de structurer les données et d'opérer des liens entre les ensembles de données.

Les données du Big Data de la filière viti-vinicole

Les sources de données du Big data viticole sont multiples et prolifèrent. Dans le domaine des sciences de données, le Big Data se caractérise par des mots commençant par "V" et le Big Data viticole n'échappe pas à cette règle mais avec ses spécificités.

- Le « Volume » toujours croissant des données, est lié au développement des technologies, à l'usage du Web et à la réduction des coûts de stockage de l'information qui permet potentiellement de les réutiliser.

- La « Variété » est la caractéristique forte du Big Data des domaines de la vigne et du vin. Au-delà des différents types d'information (texte, mesure, image, spectre, simulation...), ces domaines demandent de prendre en compte des sources de données nombreuses et extrêmement diverses, provenant :
    - de différentes échelles allant du territoire à la génomique en passant par différents niveaux tels que l'exploitation, la parcelle ou la plante.
    - des informations spatiales et temporelles pouvant décrire des zones, des trajectoires, des transformations, des opérations culturales ou œnologiques sur des sites et des installations hétérogènes.
    - des interactions et échanges qui se produisent à toutes ces échelles avec l'environnement (sol, eau, climat, bioagresseurs, fermentations,…) durant les différentes étapes.
    - des évolutions sur les outils, l'instrumentation, les équipements, les modes de conduite, les produits, les règles, etc.
    - des transformations qui vont de la parcelle à la bouteille et la caractérisation de produits à différents stades qui va jusqu'à la dégustation
    - des impacts sociaux, économiques, environnementaux, sanitaires...

Au-delà des données, c'est aussi la variété des acteurs qui les produisent et qui les manipulent, qui est en soi un challenge. Ces acteurs (viticulteurs, œnologues, agronomes, généticiens, biologistes, industriels, distributeurs, experts, consommateurs ...) ont par essence des vocabulaires et des modes de production de données très hétérogènes. Cette forte variété rend difficile l'intégration de ces données. Cela demande donc de nouvelles approches pour structurer, lier et exploiter toutes ces données potentiellement disponibles.

- La «Vélocité » correspond à la nécessité de pouvoir gérer et analyser de grandes masses de données en temps réel. En agriculture, cet aspect va prendre de l'importance avec la multiplication des sources de données (capteurs, crowd sourcing) comme c'est déjà le cas dans les plateformes de phénotypage haut-débit.

- D'autres «V» peuvent être mentionnés : «Véracité» et «Validité» qui caractérisent la qualité de sources de données ; la «Visualisation» qui est une problématique pour de grands ensembles de données complexes ; la «Visibilité» de faits ou d'éléments pertinents qu'il faut extraire de gros volumes de données ; enfin l'analyse de toutes ces données peuvent poser des problèmes d'atteinte de la «Vie privée» des acteurs de la filière.

Enfin, le véritable enjeu du Big Data c'est la Valeur que l'on peut en obtenir. Cet enjeu dépend de notre capacité à structurer le Big Data. Ceci permettra de réutiliser et d'agréger des données de différentes façons mais aussi de pouvoir les lier avec d'autres données et de nourrir les algorithmes de l'IA. Autrement dit, l'innovation autour du Big Data est indispensable pour créer de la valeur et cela passe par la structuration des données ainsi que par le développement de nouvelles approches relevant de l'analyse des données et de l'intelligence artificielle.

Les applications de l'Intelligence Artificielle

L'intelligence artificielle (IA) est la théorie et le développement de systèmes informatiques capables d'effectuer des tâches qui normalement nécessitent de l'intelligence. Plus largement, l'IA est le nom donné aux algorithmes (et aux programmes et systèmes réalisant et utilisant ces algorithmes) qui produisent ou imitent un comportement "intelligent".

Les méthodes les plus connues du grand public sont celles relevant de l'apprentissage automatique (machine learning) avec notamment le « Deep Learning » (apprentissage profond) qui est développé avec succès depuis une dizaine d'années. La majorité de ces méthodes nécessitent un grand nombre d'exemples pour réussir à paramétrer un système d'aide à la décision afin de reproduire des décisions d'experts. Des résultats impressionnants ont été obtenus principalement en classification d'images, par exemple pour reconnaître des adventices à partir de photos, mais aussi pour contribuer à des applications plus complexes. La société DeepMind a ainsi développé une application capable de battre le champion du monde du jeu de Go en s'appuyant sur un très grand nombre de parties disputées avec des humains puis sur des millions de parties entre ordinateurs. Ces résultats spectaculaires laissent entrevoir une multitude d'applications pour la filière viti-vinicole qu'il faut cependant appréhender avec précaution. Les méthodes d'apprentissage automatique posent encore en effet beaucoup de question qui relèvent de la recherche fondamentale, notamment sur leur sensibilité au jeu de données utilisé dans la phase d'apprentissage. Une partie de ces méthodes nécessitent une très grande quantité de données (souvent des milliers voire des millions d'exemples) qui ne sont pas toujours disponibles pour des applications en viticulture ou en œnologie.

Par ailleurs, l'intelligence artificielle ne se limite pas aux techniques d'apprentissage. Elle englobe d'autres méthodes qui peuvent également contribuer à valoriser les données du Big Data Agricole. Sans donner une liste exhaustive, on peut citer les enjeux liés à la représentation de connaissances pouvant être de différentes natures (expertes ou encore imprécises). C'est une étape indispensable au développement d'outils d'aide à la décision s'appuyant sur des algorithmes efficaces pour mettre en œuvre des raisonnements complexes intégrant de nombreuses contraintes.

Structuration et liage de données

Constituer les ensembles de données nécessaires à l'entraînement des algorithmes de l'intelligence artificielle, demande de mobiliser de nombreuses sources de données et requiert que ces sources soient clairement décrites et interopérables. Pour cela l'organisation et la description des données issues de ces sources doit répondre à différentes exigences :

- les objets (parcelles, cuves, vins, exploitations, enquêtes, etc), les actions (opérations, ajouts, traitements) et les problèmes (gels, pannes, etc) doivent être formellement identifiés de façon unique, non ambiguë et standardisées (URI, DOI, RFID, etc). Les variables doivent faire l'objet de conventions de nommage.

- des métadonnées standardisées doivent clairement décrire les données pour les comprendre, c'est à dire indiquer les Quoi ? Ou ? Comment ? Quand ? ainsi que pour les gérer (informatiquement et administrativement) et enfin pouvoir les lier. Afin de garantir que des différentes données se complètent ou s'enrichissent, il faut s'assurer qu'on « parle » bien de la même chose. L'usage de ressources sémantiques informatisées (dictionnaires, thésaurus, taxonomies, ontologies) permet de formaliser les concepts et les termes associés aux données. Enfin, les relations entre concepts doivent être formalisées dans des ontologies et vont permettre d'effectuer des liens, par exemple entre une cuve et un vin, ou encore un enquêteur et une exploitation.

Les données liées (Linked data) visent à favoriser la publication et l'usage de données structurées, non pas sous la forme de silos de données isolés les uns des autres, mais en les reliant entre elles pour constituer un ensemble global d'informations (les données sont organisées dans un graphe) en s'appuyant sur les standards du Web. Cette organisation permet d'incorporer des connaissances dans les ensembles de données et des connaissances ce qui est essentiel pour la qualité de ces données et pour favoriser leur (re)utilisation. Ces graphes de données contenant une mine d'informations importantes sont prêts à être exploitées et analysées pour découvrir de nouvelles connaissances.

Exemples d'applications de l'IA pour la vigne et le vin

BigDataGrapes est un projet européen qui vise à accompagner les entreprises européennes des secteurs du vin et des cosmétiques naturels à devenir plus compétitives sur les marchés internationaux. Il s'efforce en particulier d'aider les entreprises de la chaîne de valeur viticole à profiter des outils du Big Data, en fournissant une aide à la décision résultant de l'analyse en temps réel et transversale de sources de données importantes et diverses. Par exemple, dans le cadre de ce projet BigDataGrape, des travaux basés sur des techniques d'apprentissage profond ont été mis

en œuvre pour développer un outil capable à partir d'images, de compter automatiquement le nombre de feuille par cep et d'en déterminer la position. Cet outil pourra être valorisé dans plusieurs cadres applicatifs comme pour alimenter des modèles agronomiques notamment de représentation 3D ou encore pour des applications en viticulture de précision en particulier pour limiter l'usage d'intrants. Ces travaux ont été réalisés par le CNR Italie et les unités LEPSE et MISTEA.
Dans le cadre du projet BigDataGrapes, les équipes du CNR en Italie ont également mis au point des méthodes d'apprentissage automatique pour mener des analyses prédictives sur des vins. Ces données proviennent de réseaux sociaux destinés à des utilisateurs passionnés par l'œnologie (voir le livrable 4.3 du projet BigDataGrapes). Le jeu de données comportait une description générale du vin, des données générées par l'utilisateur (notes et commentaires) ainsi que des informations sur le profil de l'utilisateur. Une ensemble d'algorithmes a été déployé pour évaluer le potentiel de pénétration du marché d'un vin donné, dans un nouveau pays. Cette capacité de pénétration a été estimée en entraînant ces algorithmes, avec les données générées par les utilisateurs, pour pouvoir prédire les notes potentielles d'un vin dans un pays cible à partir des caractéristiques du vin (cépage, origine, arômes, etc).

Toujours dans le cadre de ce projet, l'université de la KULeuven collabore avec MISTEA pour développer des outils de visual analytics afin de permettre d'explorer de larges jeux de données et leur donner du sens. Un outil en cours de développement permet de visualiser des variables liées à la viticulture et à la vinification, pour différents vins sur plusieurs années. L'utilisateur pourra sélectionner des arômes ciblés ou des caractéristiques spécifiques (cépage, années d'intérêt, etc.) et comparer les vins produits à l'aide de variables sélectionnées de la vigne au vin. Pour atteindre cet objectif, le travail préliminaire a été de lier sémantiquement l'ensemble des données afin de les connecter de façon intelligente pour garantir un continuum vigne-raisin-vin, à partir des données expérimentales de Pech Rouge (unité INRAE). Cet outil permettra de répondre à plusieurs questions telles que l'influence des effets climatiques ou du processus de vinification sur la composition aromatique d'un vin.

Dans le cadre d'une collaboration entre MISTEA et NYSEOS, une étude a été conduite afin de mieux comprendre les effets environnementaux sur les profils aromatiques des vins. Des mesures de composés aromatiques sur un ensemble d'apprentissage de plusieurs dizaines de vins différents ont été utilisées. Les vins étudiés sont issus de différents sites et de différents millésimes et 3 cépages ont été utilisés. Pour un cépage donné, c'est toujours le même type de conduite et le même type de vinification qui ont été effectués quel que soit le site ou le millésime. Le nombre important de composés aromatiques du vin ainsi que la forte hétérogénéité des données rend difficile la caractérisation des profils aromatiques et leur interprétation par des méthodes classiques. Pour conduire cette étude, les données ont été structurées sous forme de graphe, et l'idée suivie a été de rechercher un sous-ensemble minimum (une clé du point de vue informatique) de composés capable de discriminer les différents vins. Ce qui peut être interprété comme l'empreinte aromatique d'un vin. Ce travail a permis de développer un outil capable d'apprendre à distinguer les vins et à essayer de mimer le comportement d'un expert… L'outil peut apprendre l'empreinte aromatique de centaines, voire de milliers de vins différents. Il permet ensuite d'établir des liens entre des sous-ensembles de composés aromatiques et les différents ensembles de données décrivant les conditions environnementales.

Conclusion et perspectives

De nombreux outils d'aide à la décision, basés sur des méthodes de l'Intelligence Artificielle et s'appuyant sur les outils du Big Data, sont actuellement en cours de développement. On peut par exemple citer un outil de détection précoce de maladie (flavescence dorée) par des techniques d'IA (société Chouette) ou un service de caviste virtuel (Société Matcha). L'enjeu majeur pour le développement de ce type de solutions demeure la donnée. Le milieu viticole est un milieu en pleine transition technologique et de nombreux acteurs de la filière ont besoin de bénéficier de nouvelles avancées pour faire face à des défis comme l'adaptation au changement climatique ou la réduction des intrants. Pour atteindre cet objectif, il s'avère crucial de pouvoir collecter les données et les rendre accessibles à des communautés plus larges sans que cela soit complexe ou redondant mais c'est une vraie problématique et une étape fondamentale pour le développement de tous les outils intelligents actuels et à venir.