Deloitte.

Risk Advisory
Avril 2019

Point de vue

Classification non supervisée : utilisations innovantes en banque

Le machine learning a connu ces dernières années un essor important en raison de l'existence de nouvelles sources de données, de systèmes informatiques plus puissants, mais aussi du perfectionnement de nouveaux algorithmes.

Les techniques de clustering permettent aujourd'hui de répondre à de nombreuses problématiques bancaires peu ou pas adressées par les méthodes actuelles.

La connaissance des ses clients, la gestion de la concentration ou l'identification précoce des fraudes sont autant de thématiques pour lesquelles le clustering peut apporter une solution innovante.

Pour tirer profit de ces avancées, encore faut-il maitriser ces algorithmes, ainsi que leur mise en œuvre. Les enjeux de ces méthodes sont en effet très différents de ceux des méthodes actuelles, se focalisant sur une variable d'intérêt préalablement déterminée.

« Les méthodes de classification non supervisées sont disruptives pour de nombreuses thématiques rencontrées par les banques. »

Hervé Phaure
Associé Risk Advisory

Panorama des techniques de clustering

Parmi les techniques statistiques les plus intéressantes, figure le clustering, méthode d'optimisation visant à regrouper les données en se basant sur leur proximité dans l'espace de représentation. Dans cette optique, les données les plus similaires sont rassemblées dans des groupes appelés « clusters » qui doivent être homogènes et bien distincts les uns des autres.

Le sujet du clustering a été abordé dans plusieurs disciplines et dans des contextes bien différents. En effet, le terme cluster est apparu pour la première fois en Avril 1954 dans l'article « *Use of Cluster Analysis with Anthropological Data* » de Forrest E. Clements, publié dans le « *American Anthropologist* ». Depuis, les méthodes ont évolué et l'éventail de leurs usages s'est également élargi : aéronautique, médecine, biologie, banque, etc.

Le clustering fait référence aux algorithmes de classification non supervisés. Il s'agit d'un ensemble de méthodes remplissant l'objectif d'identifier une structure propre à partir de caractéristiques mesurées sur chacune d'elles sans à priori. Au contraire, pour la classification supervisée, l'appartenance des données aux différents groupes est supposée connue. L'objectif est alors de construire une règle de classement pour prédire le groupe d'appartenance des nouvelles observations.

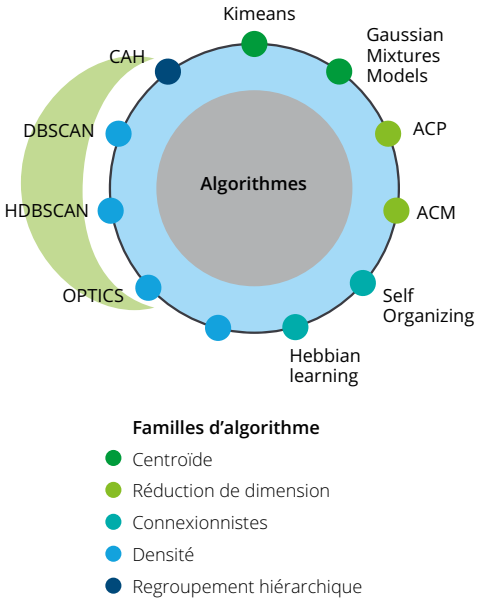
Plusieurs méthodes de clustering prévalent :

- les méthodes basées sur un centroïde ;
- les méthodes basées sur réduction de dimension ;
- les méthodes de regroupement hiérarchique ;
- les méthodes basées sur la densité ;
- les méthodes connexionnistes.

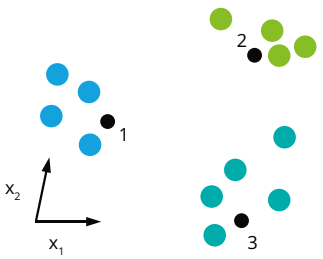
Les trois algorithmes les plus fréquemment utilisés sont décrits ci-dessous :

	Description	Plus	Difficulté
K-Means	S'appuie sur une mesure de distance afin de créer k clusters en minimisant la somme des carrés des distances entre un point et la moyenne des points de son cluster.	• Grande simplicité • Rapidité	• Choix du nombre de clusters
CAH	Permet de partitionner un jeu de données de manière hiérarchique en agrégeant à chaque étape les deux clusters les plus proches. La distance entre un individu et un groupe ainsi que celle entre deux groupes est ainsi calculée à chaque étape.	• Stabilité	• Choix du nombre de classes
DBSCAN	S'appuie sur la densité estimée des clusters pour effectuer le partitionnement. Il utilise 2 paramètres : la distance ε et le nombre minimum de points MinPts devant se trouver dans un rayon ε pour que ces points soient considérés comme un cluster.	• Pas de choix du nb de clusters • Robuste au données aberrante (isolées) • Performant sur les formes arbitraires	• Mis en défaut en cas de densités locales hétérogènes • Choix des paramètres

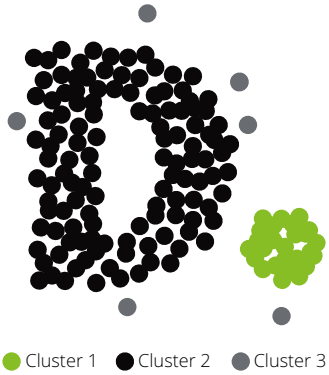
Familles d'algorithme



Résultats de l'algorithme K-Means dans un plan en 2d



Résultats de l'algorithme DBSCAN dans un plan en 2d



Enjeux et contraintes d'utilisation

Le principal attrait des techniques de clustering, et des algorithmes non supervisés plus généralement, est de ne pas faire d'hypothèses sur les données. Il est donc possible de faire émerger des schémas inattendus, à la fois d'un point de vue métier et en comparaison des techniques supervisées les plus courantes.

Cependant, le principe fondateur de ces techniques est aussi une très forte contrainte. Il induit des complexités spécifiques à chaque étape de la construction d'un modèle :

- sélection et Retraitement des variables ;
- optimisation des algorithmes et quantification de la performance ;
- interprétation et appropriation des résultats.

Sélection et retraitement des variables

Contrairement aux modèles supervisés, il n'est pas possible de mettre en place des indicateurs d'intensité avec une variable dépendante. La sélection des variables ne repose donc pas sur les mêmes tenants et n'est dans tous les cas pas soumis aux mêmes contraintes (ex : corrélation entre les variables). Les techniques de clustering permettent aisément d'intégrer un nombre important de variables, sans que cela ne nuise à l'interprétation des résultats.

Le retraitement des données est cependant un préalable à l'application des différents algorithmes, fonctionnant pour la plupart à partir d'un calcul de distance entre les points et nécessitant donc des données quantitatives. A cet effet, plusieurs stratégies sont envisageables. Le choix dépend du type de variable à intégrer : ordinales, nominales ou continues. Il pourra par exemple se faire suivant la démarche suivante :

1. Ajustements préliminaires : à défaut de pouvoir s'appuyer sur une variable cible (ex : défaut, fraude...), l'homogénéité des classes devra être privilégiée. Une démarche systématique peut être mise en place afin de réaliser le traitement préalable des données (traitement des valeurs manquantes, regroupement des modalités des variables catégorielles, discrétisations des variables continues).
2. Codage des variables au format numérique devant être réalisé en choisissant avec précaution la nature du codage (ex : codage en dummies simple), car cela à une incidence dans le calcul des distances : Il s'agira en particulier de coder les variables en veillant à ne pas surpondérer certaines modalités. C'est par exemple le cas si une variable peut prendre de nombreuses valeurs et que le codage : 1, 2, ..., N est retenu. En effet, le calcul des distances privilégiera les variables les plus éloignées du nuage des points.

Une démarche systématique comprenant la sélection et le retraitement des variables peut être définie afin de permettre une mise en place rapide des algorithmes de clustering, qui pourra ensuite être reconduite sans ajustement particulier.

Optimisation des algorithmes et quantification de la performance

La mesure de la performance d'un modèle de type clustering suppose de recourir à des indicateurs spécifiques car indépendants de la prédictivité au regard d'une variable à modéliser. Deux dimensions en particuliers doivent être considérées :

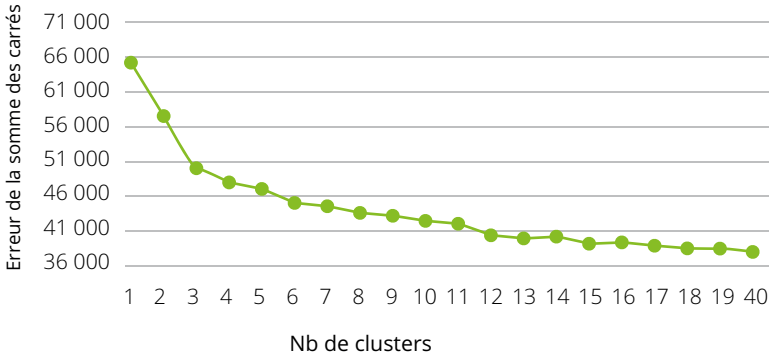
1. Homogénéité et spécificité des clusters : indicateurs directement calculables, permettant de catégoriser chaque cluster selon l'homogénéité, ainsi que la spécificité de chacun d'eux par rapport à la population d'ensemble. Seuls les indices utilisant des informations internes sont retenus.

Indicateurs	Description
Indice Silhouette moyen	Utilise la notion de distance moyenne intra-cluster et le minimum de la distance moyenne entre un point et tous les autres appartenant à des clusters différents.
Indice de Dunn	Rapport entre la distance maximum intra-cluster et la distance minimale inter-cluster.
Somme des erreurs carrés	Somme des distances au carré des points intra-classe.
Calinski-Harabasz	Basé sur les moyennes inter-cluster et la matrice de covariance intra-cluster.
Davies-Boudin	Prend en compte la distance moyenne intra-cluser et la somme des distances inter-clusters.

2. Stabilité dans le temps : indicateurs permettant de pouvoir comparer les clusters déjà définis, ainsi que leur qualité, en suivant la stabilité des découpages et l'analyse des évolutions substantielles.

L'une des principales difficultés des méthodes de clustering est la définition du nombre de clusters optimal. En augmentant le nombre de classes, l'homogénéité totale des clusters augmente mécaniquement mais complexifie la compréhension d'ensemble. Une stratégie simple consiste à incrémenter les paramètres des algorithmes (K pour le K-Means, epsilon pour le Dbscan) et à retenir la dernière valeur qui induit un gain informationnel significatif. Pour cela, chacun des indicateurs définis précédemment peuvent être utilisés, et notamment l'erreur de la somme des carrés (voir figure à droite).

Variance de l'ensemble fdes points par rapport aux centres des clusters en fonction du nombre de clusters



Interprétation et appropriation des résultats

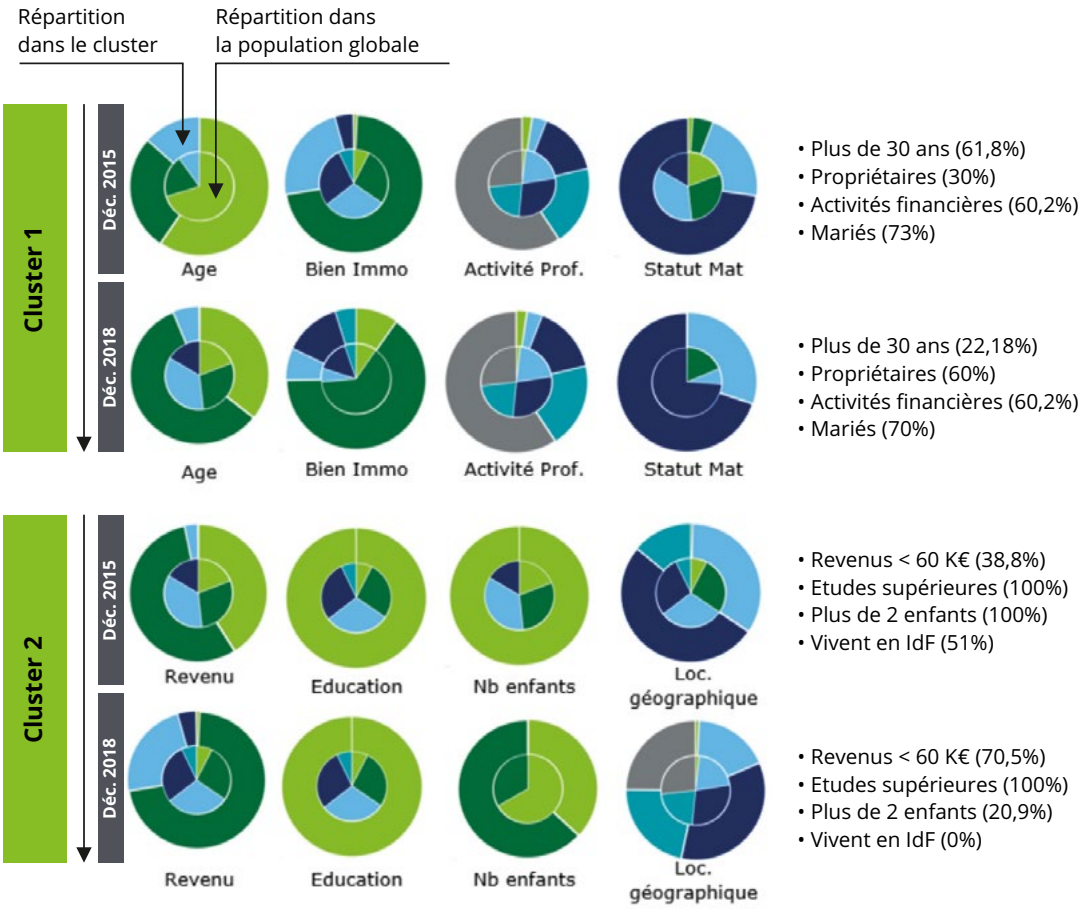
L'interprétabilité des résultats des algorithmes de clustering est un point essentiel pour pouvoir les utiliser de manière pratique.

Les éléments définissant chacun des clusters doivent être immédiatement mis en évidence.

Cela nécessite la mise en place d'un indicateur classant les variables les plus importantes dans la définition de chacun des clusters.

Une représentation graphique élaborée doit permettre d'interpréter rapidement les clusters et présenter les résultats à différents types d'interlocuteurs.

Comparaison de clusters identifiés entre deux dates sur la base des quatre variables d'intérêt principales



- Plus de 30 ans (61,8%)
- Propriétaires (30%)
- Activités financières (60,2%)
- Mariés (73%)

- Plus de 30 ans (22,18%)
- Propriétaires (60%)
- Activités financières (60,2%)
- Mariés (70%)

- Revenus < 60 K€ (38,8%)
- Etudes supérieures (100%)
- Plus de 2 enfants (100%)
- Vivent en IdF (51%)

- Revenus < 60 K€ (70,5%)
- Etudes supérieures (100%)
- Plus de 2 enfants (20,9%)
- Vivent en IdF (0%)

Cas d'usages métiers

Pour répondre aux différents enjeux d'une application pratique du clustering, nos équipes ont développé une démarche complète pouvant répondre à la fois aux défis du retraitements des données, du paramétrage des algorithmes et du reporting des résultats.
Nous pensons que l'apport du clustering dans le cadre d'un processus systématique peut créer une rupture dans la manière d'appréhender de nombreux sujets. Trois cas d'usage ont été identifiées pour illustrer ces opportunités :

- Qui sont mes clients ?
- Risque de concentration
- Détection précoce des fraudes

Premier cas d'usage – Qui sont mes clients ?

Problématique

Pour le pilotage des activités de crédit, la connaissance de la base client est un réel enjeu, permettant la mise en place d'actions ciblées déterminantes pour atteindre les objectifs de développement commercial et de rentabilité.

Dans le cadre de campagnes marketing par exemple, le suivi de l'efficacité et de l'impact sur la population est primordial. La campagne doit pouvoir être réajustée rapidement en fonction des retours, et notamment dans un environnement où l'information circule rapidement.

Pour la mise en place de nouvelles offres, produits ou activités, la connaissance et le suivi de la population du portefeuille clientèle sont indispensables pour identifier les risques de dérapages et les effets d'aubaine péjorant.

Dans ce cadre, il est donc primordial de pouvoir identifier systématiquement et très rapidement la composition du portefeuille clientèle, mais aussi de mettre en évidence toute tendance d'évolution ou d'émergence de sous-population particulières.

Limites des méthodes actuelles

Les techniques actuelles, notamment l'Analyse en Composantes Principales (ACP) qui est la plus populaire, permettent d'explorer des jeux de données multidimensionnels. La visualisation des observations dans un espace réduit (deux ou trois dimensions) est la première étape permettant ensuite d'identifier des groupes homogènes ou des observations atypiques.

Cependant, la mise en place de cette méthode peut être jugée comme longue, coûteuse et n'offrant que de faibles garanties quant aux résultats espérés. Elle est donc rarement mise en œuvre ou dans un cadre trop statique, la problématique étant finalement peu adressée.

Application du clustering

Les algorithmes de clustering sont particulièrement efficaces pour le ciblage ou la segmentation de clients. L'identification des clusters est rapide et précise, permettant de dégager rapidement un consensus sur la répartition d'un portefeuille.

La comparaison de plusieurs photos à fréquence variable (voir figure de droite) permet de constater l'évolution de clusters identifiés ou l'émergence de concentrations particulières inattendues et d'en dégager des constats pouvant permettre d'ajuster par exemple une compagne commerciale mal ciblée.

Les évolutions les plus significatives peuvent ensuite faire l'objet d'une analyse ad-hoc selon la même méthode afin de mettre en évidence leurs fondements.

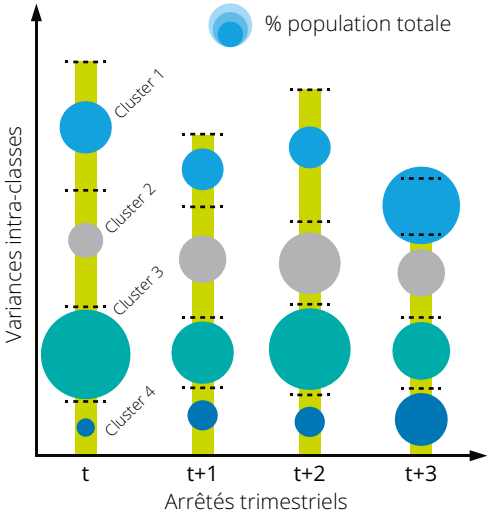
Deuxième cas d'usage – Concentration

Problématique

La concentration est un enjeu particulier pour la banque, pouvant notamment provoquer une augmentation forte de la sinistralité du portefeuille si elle n'est pas correctement appréhendée. En effet, en cas de corrélation importante des tiers composants le portefeuille, un défaut peut se propager à d'autres tiers par effet de contagion. En outre, un retournement conjoncturel affectant défavorablement le taux de défaut du portefeuille peut être démultiplié en cas de concentration forte. Un risque de concentration maitrisé rend donc la banque moins vulnérable aux difficultés économiques d'une industrie ou d'une région particulière.

Dans les faits, la gestion opérationnelle de la concentration n'est pas adressée, sauf dans le cadre de mesure de type Pilier 2. Une évolution significative

Evolution trimestrielle du % de la variance totale et du % du nombre de tiers



de la concentration/diversification du portefeuille devrait permettre d'enclencher une réflexion sur l'appétit au risque et la mise en place d'actions de gestion en réaction.

Limites des méthodes actuelles
Dans le cadre du pilier II et notamment de l'ICAAP, la banque doit être capable d'évaluer son risque intrinsèque à la concentration. Pour cela, les banques recourent habituellement au calcul d'indices de concentration tels que Hirschman-Herfindahl et de méthodes plus ou moins complexes pour calculer des impacts en RWA. Ce calcul est réalisé sur une base d'axes de segmentation définis le plus généralement par une approche « experte ». Des seuils fixés permettent de déterminer l'intensité de la concentration.

Ainsi, la sélection des axes d'analyse est donc limitative et le lien avec le risque de défaillance ou la rentabilité très peu mis en évidence. De plus, la gestion opérationnelle de la concentration n'est pas adressée convenablement n'intégrant par les opportunités potentielles d'une vision plus large de la concentration.

Application du clustering
La démarche de clustering développée permet d'estimer la concentration d'une manière totalement inédite, intégrant la notion de risque rattachée à la sinistralité du portefeuille, mais aussi aux opportunités potentielles métiers.

Pour cela, la comparaison du portefeuille à deux dates permet de déterminer l'impact en termes de taux de défaut, de rentabilité liée à une augmentation de la concentration d'un cluster sur une modalité, d'une variable identifiée. Il faudra pour cela décomposer les différents effets (conjoncture, nouveaux produits…).

Le suivi des concentrations identifiées, et de leur traduction dans des indicateurs de diversification particulier peut permettre d'objectiver la relation éventuelle entre la diversification moyenne du portefeuille et son évolution, et l'impact sur la sinistralité ou sur tout autre variable liée.

Troisième cas d'usage – Détection précoce de fraudes en financement
Problématique
La détection de fraude précoce par les banques demeure un enjeu de taille au regard de la grande diversification du service bancaire et de sa forte digitalisation. En effet, les fraudeurs font évoluer très rapidement leurs techniques pour échapper aux contrôles existants obligeant les banques à améliorer dans le même temps leur dispositif de surveillance. Or, cette adaptation peut se faire avec un retard important, pouvant entrainer des pertes substantielles.

Limites des méthodes actuelles : analyses semi-supervisées
Les banques utilisent des méthodes dites de « classifications supervisées » ou à partir de règles métiers pour identifier des cas de fraudes parmi des schémas déjà connus. Or, lorsque le mécanisme de fraude change ou se perfectionne, la banque n'a pas encore de règles permettant de détecter ces nouveaux cas. De nombreuses fraudes similaires peuvent donc se produire avant la matérialisation de la première perte. Le modèle doit donc évoluer, mais si cette évolution intervient trop tard, les pertes vont se cumuler.

Application du clustering
Les algorithmes de clustering permettent de construire des groupes de clients homogènes. Il est donc possible de mettre en place une démarche en trois étapes :

1. Constitution de clusters sur la base de dimensions pré-identifiées et susceptibles de qualifier des segments porteurs de fraude (la date de naissance, la nationalité, le type de crédit demandé, le montant du crédit, le motif de la demande, le salaire annuel brut du demandeur, sa situation professionnelle, sa situation familiale)

2. Caractérisation des clusters en fonction de la probabilité d'occurrence des fraudes

3. Suivi de l'évolution de ces clusters dans le temps, pouvant donner lieu à des analyses ad-hoc lorsque l'émergence d'une population est constatée à l'intérieur d'un cluster de fraude préexistant, ou dans un nouveau cluster pour lequel la probabilité d'occurrence d'une fraude est inconnue

La classification non supervisée : une solution nécessaire à une multitude de thématiques

La performance des algorithmes de clustering met en lumière la nécessité d'intégrer de manière systématique ces nouvelles techniques aux problématiques historiques du secteur bancaire.

La démarche proposée par nos équipes répond à tous les enjeux de ces nouvelles techniques pour pouvoir adresser un maximum de sujets.

La démarche privilégiée pour répondre à ces enjeux s'organise autour des étapes suivantes :

1

Réaliser un premier clustering à une date de référence permettant de dégager de façon automatique les différents axes de segmentation structurels (clusters) du portefeuille

2

Calculer les indicateurs d'homogénéité adaptés et caractériser les clusters suivant leur comportement sur différentes dimensions (Défaut, PNB...)

3

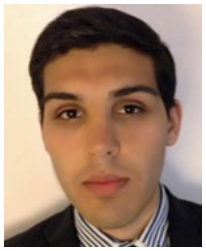
Réaliser une analyse spécifique dépendant du cas d'usage métier identifié

4

Mettre en place des actions opérationnelles ciblées pour répondre à la problématique identifiée et en suivre les effets dans le temps dans le cadre d'une application périodique et systématique de l'approche

Hervé Phaure
Associé Risk Advisory

Hervé est Associé responsable de l'activité Credit Risk Advisory. Il intervient depuis plus de 25 ans sur les problématiques de gestion de risques financiers, et accompagne les institutions financières dans leurs projets de gestion et de modélisation des risques de crédit et opérationnel, ainsi que dans l'optimisation de la gestion de leurs activités de crédit. Hervé est par ailleurs membre de l'ITG pour la transposition de la norme IFRS 9 et coordonne la Communauté Crédit EMEA de Deloitte.



Jérémy Sartre
Manager Risk Advisory

Jérémy est manager dans l'équipe Credit Risk du département Risk Advisory. Il intervient depuis 8 ans sur la gestion du risque de crédit et notamment sur la modélisation et la revue des modèles de risque de crédit. Il est impliqué dans ce cadre, sur le développement de solutions innovantes de type machine learning.

Un grand remerciement à Kéli Akakpo-Folly, Teico Kadadji et Djibril Diop pour leurs contributions.