OptiClust4Rec: Optimisation du clustering des données de patients suivant une thérapie médicale pour l'aide à l'amélioration de la qualité de vie

Yanis Bouallouche*, Juba Agoun**
Mohand-Said Hacid**

*Université de Montpellier
yanis.bouallouche@etu.umontpellier.fr
**Université Claude Bernard Lyon 1, CNRS, LIRIS, 69100, Villeurbanne, France
juba.agoun@univ-lyon1.fr
mohand-said.hacid@univ-lyon1.fr

Résumé. Lors de l'introduction de nouvelles thérapies médicales, un ensemble de données présentant des caractéristiques sémantiques variées est recueilli auprès d'une cohorte de participants. Le recours à l'apprentissage non supervisé est privilégié comme première étape pour explorer ces données, permettant ainsi d'extraire des informations précieuses avant d'entreprendre la tâche laborieuse de l'étiquetage. Le clustering émerge comme l'une des techniques offrant un aperçu complet de l'analyse exploratoire des données, facilitant l'identification de communautés de patients. Avec OptiClust4Rec [1], nous proposons une méthodologie visant à caractériser les groupes, fournissant ainsi des recommandations pour les patients suivant un traitement thérapeutique. Cette approche repose sur l'utilisation de deux ensembles de données distincts : le premier contenant les données cliniques des patients et le second regroupant les réponses des patients à un questionnaire portant sur leur qualité de vie. Notre objectif principal est d'optimiser le processus de clustering et la réduction de la dimensionnalité, en nous appuyant sur des métriques concises et une analyse de la topologie des données, afin d'étiqueter efficacement les différents clusters et de générer des règles d'association pertinentes.

1 Introduction

L'industrie de la santé génère une quantité importante de données chaque jour. De nouvelles thérapies sont testées sur des patients lors d'essais cliniques. Les données acquises sont sémantiquement différentes, couvrant l'état de santé, les effets secondaires, le potentiel de travail et le mode de vie. Elles sont recueillies pour révéler des inférences causales concernant l'efficacité des traitements. Ces données sont donc devenues cruciales pour prédire les futures conditions de santé, réduire les coûts des traitements et améliorer la qualité de vie en général. Par exemple, pour améliorer le traitement par immunothérapie des patients, l'analyse des

1. Article accepté à CoopIS 2023

données médicales en amont permet la production de recommandations et de lignes directrices pour les praticiens.

Dans l'aide à la prise de décision médicale, le regroupement classique des patients s'avère être une approche fiable. Cela implique d'identifier les caractéristiques dominantes au sein des groupes de patients et de suivre leur évolution de santé. Les techniques non supervisées permettent une analyse initiale des relations entre les données, sans nécessiter de connaissances spécialisées dans le domaine. Bien que les méthodes supervisées aient fait leurs preuves en termes d'efficacité, la création d'ensembles de données étiquetés reste une tâche chronophage et exigeante en main-d'œuvre. C'est pourquoi, avec OptiClust4Rec, nous proposons une méthodologie non supervisée axée sur les données, permettant la découverte de motifs cachés dans les données cliniques des patients ainsi que dans leurs réponses à des questionnaires.

OptiClust4R introduit une interface web conviviale visant à évaluer les capacités des algorithmes de clustering pour des ensembles de données spécifiques. Cette évaluation utilise des mesures statistiques basées sur des métriques internes, complétées par une exploration visuelle pour analyser les variations des mesures avec divers paramètres et techniques de réduction de dimensionnalité. Guidé par l'approche illustrée dans la figure 1 cet outil assiste les utilisateurs dans l'optimisation du processus de clustering, aboutissant à des groupes distincts, qui sont ensuite caractérisés en extrayant des caractéristiques essentielles. Pour générer des recommandations pour les patients, nous utilisons deux ensembles de données : l'un contenant des analyses et des informations sur les patients, et l'autre contenant leurs réponses à des questionnaires. OptiClust4Rec est conçu pour regrouper les patients en fonction des deux ensembles de données et attribuer des étiquettes aux groupes. Cela fournit ainsi aux utilisateurs finaux des règles d'association et facilite ensuite la génération de recommandations.

FIG. 1 – Illustration résumant notre approche

2 Aperçu du système

Dans notre approche, nous abordons deux défis. Tout d'abord, nous proposons un ensemble de métriques et d'outils de visualisation qui permettront aux utilisateurs d'optimiser le regroupement des données. Ensuite, en exploitant les résultats du clustering, nous nous concentrerons

sur chaque cluster pour extraire les variables qui le caractérisent, créant ainsi essentiellement une forme d'étiquetage automatique basée sur ce que l'on appelle les *salient features*.

2.1 Clustering

Il existe un éventail de techniques de clustering bien établies parmi lesquelles choisir, incluant celles basées sur les centroïdes, la connectivité et la densité. Le processus de sélection de la méthode appropriée ainsi que des paramètres pertinents pour un ensemble de données donné est généralement conduit de manière déterministe. En pratique, le choix entre différentes méthodes de clustering, ainsi que la détermination du nombre optimal de clusters, nécessite souvent de tester divers algorithmes. Certains chercheurs préfèrent utiliser une méthode par défaut (comme *k-means*) avec un nombre de clusters déjà déterminé en fonction de leurs connaissances du domaine, tandis que d'autres optent de manière subjective pour la méthode la plus récente disponible Parker et Barnard (2019). Par la suite, nous aborderons les deux optimisations que nous ciblons avec notre outil.

2.1.1 Trouver un algorithme de clustering adapté

Sélectionner le modèle approprié parmi les algorithmes de clustering existants représente un défi. Notre approche repose sur l'utilisation d'une technique appelée *Persistent Homology*, une technique fondamentale dans le domaine de l'Analyse Topologiques des Données (TDA) Wasserman (2018). Cette méthode examine de manière systématique les relations entre les points de données à différentes échelles, offrant des informations cruciales sur la présence, la configuration géométrique et la densité des groupes potentiels. L'application de Persistent Homology génère un diagramme de persistance.

La figure 2 montre un exemple d'un diagramme de persistance. Dans ce diagramme, chaque couleur distincte représente un groupe d'homologie unique. En particulier, H0 représente les composantes connectées, fournissant des indications pour estimer un nombre potentiel de clusters. Les points H1 signalent les vides unidimensionnels, tandis que le groupe H2 décrit les vides bidimensionnels, souvent appelés *cavités*. Le diagramme trace de manière complexe les changements homologiques à mesure que nous explorons différentes échelles de distance ou de proximité au sein de l'ensemble de données. Chaque composante connectée et vide se matérialise sous la forme d'un point sur le graphique, avec son début (apparition) et sa fin (disparition) tracés le long des axes horizontal et vertical. Un écart significatif d'un point par rapport à la diagonale indique la présence persistante de la composante correspondante, suggérant potentiellement des structures robustes et durables au sein des données. Après un examen approfondi du diagramme de persistance pour un ensemble de données donné, l'écart notable de certains points H1 et H2 par rapport à la diagonale suggère fortement la présence de structures non linéaires. Celles-ci peuvent impliquer des formes sphériques ou des clusters qui se chevauchent. Par conséquent, cette observation soutient fortement la recommandation d'utiliser un algorithme basé sur la densité tel que DBSCAN, reconnu pour sa capacité à fonctionner exceptionnellement bien dans des scénarios complexes de ce type.

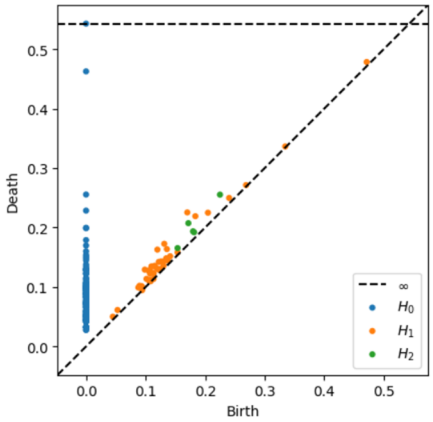

FIG. 2 – Exemple d'un diagramme de persistence

2.1.2 Trouver le nombre optimal de clusters *k*

Les méthodes de clustering qui ne dépendent pas de la densité requièrent de connaître à l'avance le nombre de clusters à former. Par conséquent, déterminer le nombre exact ou optimal représente un défi, étant donné l'absence d'une méthode universelle pour identifier le nombre idéal selon les caractéristiques spécifiques d'un ensemble de données donné. Deux approches couramment utilisées pour aborder cette problématique sont la méthode du coude (Elbow Method) et le score de silhouette (Silhouette Score).

La *elbow method* évalue la compacité d'un cluster en calculant la somme des carrés des distances intra-cluster (Within-Cluster Sum of Squared, WCSS) pour différents nombres de clusters. Elle observe une diminution de la WCSS à mesure que le nombre de clusters augmente, indiquant ainsi une meilleure compacité. Cependant, il arrive un point dans lequel l'ajout de plus de clusters n'améliore plus la qualité. Le point de cette décroissance marquée, représenté visuellement comme un coude dans le graphique de l'évolution de la WCSS par rapport au nombre de clusters, indique le nombre de clusters optimal. Il est à noter que cette méthode ne prend pas en compte la séparation entre les clusters, un aspect important du clustering idéal.

En revanche, le score de silhouette fournit une évaluation plus complète de la qualité des clusters en tenant compte à la fois de la cohésion (distance moyenne à l'intérieur d'un groupe) et de la séparation (distance moyenne au groupe voisin le plus proche) pour chaque point de données. Cependant, il peut rencontrer des difficultés à identifier des clusters complexes de formes diverses. Il est également important de noter que dans certains cas, les clusters identifiés par le score de silhouette peuvent présenter des distributions inégales, ce qui peut entraîner des groupes avec seulement quelques observations tandis que le reste est dispersé à travers d'autres groupes.

Pour déterminer la valeur optimale de k, nous faisons appel à deux mesures internes : la connectivité et la variabilité. Notre objectif est d'analyser comment la variabilité évolue par rapport à la connectivité. Cette analyse nous permet de repérer le point caractéristique où la variabilité diminue significativement tandis que la connectivité augmente légèrement. Ce point est identifiable comme un genou (knee) sur le graphique illustrant l'évolution de la

variabilité en fonction de la connectivité, indiquant ainsi le nombre optimal de clusters. Grâce à l'algorithme *kneed* Satopaa et al. (2011), nous sommes en mesure de déterminer ce point de manière précise, plutôt que de se fier uniquement à une évaluation visuelle. Notre approche est appliquée à sept algorithmes différents et se conclut par un processus de vote. Par exemple, si quatre des sept algorithmes recommandent quatre clusters comme nombre idéal, alors quatre sera retenu comme le nombre optimal.

Étant donné que nous travaillerons avec des données comportant un grand nombre de variables, leur traitement pose un problème bien connu appelé le **fléau de la dimension**, soulignant ainsi l'importance de la réduction de la dimensionnalité. Il existe plusieurs techniques et algorithmes dans la littérature pour effectuer cette réduction. L'Analyse en Composantes Principales (PCA) est la plus connue. Elle opère en transformant les variables initiales en un nouvel ensemble de variables, appelées composantes principales, qui sont des combinaisons linéaires des variables originales. Cela permet de mettre en évidence les relations linéaires et les tendances dans les données. Toutefois, dans notre approche, nous avons privilégié l'utilisation d'UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) McInnes et al. (2018), en raison de son efficacité à préserver la structure non linéaire et la topologie des données. Nous explorons diverses réductions de dimensionnalité, en considérant un maximum de $\sqrt{N}$ dimensions pour N observations. L'objectif est de déterminer la dimension la plus stable, où la variance du nombre optimal de clusters pour chaque algorithme est la plus faible.

2.2 Caractérisation des clusters

Dans la perspective de notre objectif principal, qui est de travailler avec des données sémantiquement variées en vue de générer des règles d'association, il devient crucial que nos données soient étiquetées de façon appropriée. Après l'application du processus de clustering, notre objectif est de distinguer les traits saillants de chaque cluster en extrayant les éléments pertinents. Nous adoptons à cette fin l'approche décrite dans Khoie et al. (2017). À ce stade, les observations sont réparties en deux groupes : *in-pattern* et *out-pattern*. En analysant ces groupes au sein de chaque cluster, nous pouvons identifier les caractéristiques marquantes et déterminer si les variables associées présentent des valeurs élevées ou basses. Le schéma représenté dans la figure 3 est divisé en six étapes, que nous allons détailler ci-dessous :

1. La première étape consiste à calculer le centroïde X_k de chaque groupe. Ici, P_i représente le point i, et N le nombre total d'observations dans un cluster.
2. La deuxième étape consiste à diviser les points en deux groupes : les *in-pattern* et les *out-pattern*. Pour chaque point i, nous calculons d_i, qui représente la distance euclidienne entre le point i et son centroïde. Si d_i se situe dans l'intervalle décrit dans la figure, alors le point i est un *in-pattern*, sinon il est un *out-pattern*. μ_k et σ_k représentent respectivement la moyenne et l'écart type du groupe k, et z est une constante
3. La troisième étape implique le calcul de la moyenne des *in-pattern* et *out-pattern* pour chaque cluster k et chaque variable v. Ici, φ_{in} représente l'ensemble des *in-pattern* et φ_{out} représente l'ensemble des *out-pattern*.
4. La quatrième étape consiste à calculer un facteur de différentielle pour chaque variable et chaque cluster.

5. La cinquième étape comprend le calcul de la moyenne et de l'écart type pour toutes les variables et pour chaque cluster.
6. Enfin, pour la sixième et dernière étape, une variable v pour un cluster donné k est considérée comme saillante si le facteur de différentielle répond aux deux conditions dans la figure. De plus, si le facteur de différentielle est positif, cela signifie que les valeurs de la variable v pour le cluster k sont majoritairement élevées, tandis que s'il est négatif, cela indique des valeurs majoritairement basses.

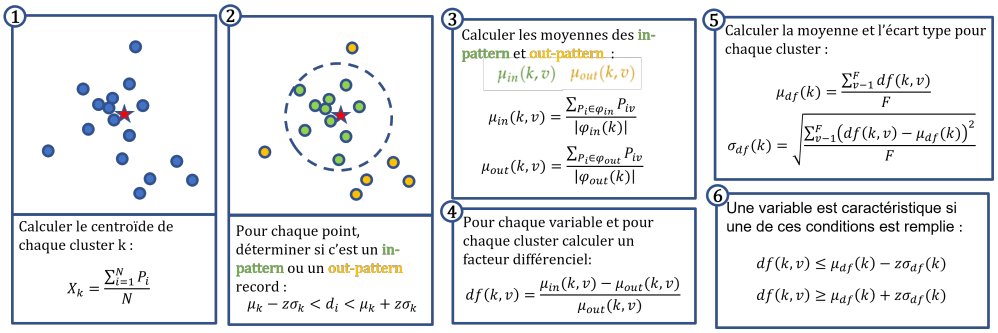

FIG. 3 – Les étapes décrivant l'approche utilisée pour l'extraction des variables caractéristiques

3 OptiClust4Rec

OptiClust4Rec [2] est une application web conçue pour l'analyse et la visualisation de jeux de données, qu'ils soient médicaux ou non. Elle utilise principalement des techniques non supervisées, principalement le clustering, et propose des conseils grâce à l'utilisation de la réduction de la dimensionnalité, de l'analyse de données topologiques et de méthodes d'étiquetage automatique.

Comme le montrent les captures d'écran de la figure 4, OptiClust4Rec dispose de plusieurs interfaces utilisateur, principalement trois, que nous détaillons dans les sections suivantes :

A. Cette zone présente les résultats des différentes opérations de clustering dans différentes dimensions. Après avoir chargé les ensembles de données dans une autre interface, l'utilisateur obtient des résultats comparés aux métriques internes les plus connues de la littérature.

B. Cette zone affiche le résultat de Persistent Homology, qui fournit des informations sur la présence de cavités. Plus nous trouvons de points H1 et H2 s'écartant de la diagonale, plus nous recommandons l'utilisation d'algorithmes basés sur la densité.

C. Cette zone est destinée à afficher les résultats de la caractérisation des groupes. Une fois que l'utilisateur sélectionne la méthode de clustering appropriée, chaque cluster est étiqueté avec les variables saillantes.

2. Lien vers l'outil et la vidéo : `https://tinyurl.com/Demo-OptiClust4Rec`

FIG. 4 – Capture d'écran de l'interface d'OptiClust4Rec.

En fonction du résultat de la caractérisation des groupes de chaque jeu de données, l'utilisateur peut établir des corrélations entre les groupes des différents ensembles de données sémantiques.

4 Conclusion et perspectives

En conclusion, notre approche, OptiClust4Rec, propose une nouvelle méthodologie pour aborder l'optimisation du clustering en intégrant une analyse topologique des données, associée à l'utilisation de mesures statistiques robustes. Cependant, il est important de souligner que notre travail ouvre la porte à plusieurs avenues prometteuses pour la recherche future. Tout d'abord, l'exploration de techniques de réduction de la dimensionnalité plus avancées pourrait offrir des perspectives plus riches pour la caractérisation des clusters. De plus, étendre notre méthodologie à d'autres domaines médicaux et évaluer sa faisabilité sur des ensembles de données plus vastes représente une étape importante.

Finalement, OptiClust4Rec constitue une base solide pour l'optimisation du clustering, offrant ainsi la possibilité de générer des recommandations en matière de qualité de vie pendant les thérapies médicales. Nous sommes impatients de voir comment cette méthodologie évoluera et apportera des avantages tangibles aux patients et aux praticiens dans les années à venir.

Remerciements Ce travail est soutenu par le programme de recherche et d'innovation Horizon 2020 de l'Union européenne dans le cadre de l'accord de subvention n° 875171, du projet QUALITOP (Surveillance des aspects multidimensionnels de la qualité de vie après l'immunothérapie contre le cancer - une plateforme numérique intelligente ouverte pour la prévention personnalisée et la gestion des patients).

Références

Khoie, M. R., T. Tabrizi, E. Khorasani, et N. Marhamati (2017). A hospital recommendation system based on patient satisfaction survey. *Applied Sciences 7*, 966.

McInnes, L., J. Healy, N. Saul, et L. Großberger (2018). Umap : Uniform manifold approximation and projection. *Journal of Open Source Software 3*(29), 861.

Parker, A. J. et A. S. Barnard (2019). Selecting appropriate clustering methods for materials science applications of machine learning. *Advanced Theory and Simulations 2*(12), 1900145.

Satopaa, V., J. Albrecht, D. Irwin, et B. Raghavan (2011). Finding a "kneedle" in a haystack : Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pp. 166–171.

Wasserman, L. (2018). Topological data analysis. *Annual Review of Statistics and Its Application 5*(1), 501–532.

Summary

Upon the introduction of novel medical therapies, an array of semantically different data is gathered from the participant's cohort. Unsupervised learning is always privileged as a preliminary step for data investigation, to extract valuable information before embarking on the tedious task of data labeling. Clustering emerges as one of the techniques providing a comprehensive overview of exploratory data analysis, facilitating the identification of patient communities. With OptiClust4Rec, we propose a methodology aimed at characterizing the groups, thus providing recommendations for patients undergoing therapeutic treatment. This approach relies on the use of two distinct datasets: the first containing patients' clinical data and the second grouping patients' responses to a questionnaire on their quality of life. Our main objective is to optimize the clustering process and dimensionality reduction, relying on concise metrics and an analysis of the data's topology, in order to effectively label the different clusters and generate relevant association rules.