UNIVERSITE LUMIERE LYON LUMIERE II

Institut de la communication

MASTER 1

Informatique

Clustering et analyse multidimensionnelle

Rapport d'étude sous la direction de

Sabine LOUDCHER

Karidjatou DIABY

Mame Libasse MBOUP

Saliou NDAO

Deffa NDIAYE

Année universitaire 2020-2021

Table des matières

Introduction

L’analyse de données multidimensionnelle regroupe un ensemble de méthodes statistiques récentes utilisées pour analyser des données afin d’en tirer des informations pertinentes pour un besoin donné. Ces méthodes sont fondées soit sur des lois mathématiques, c’est le cas de l’analyse factorielle, soit sur l’informatique cas de la classification automatique. Le terme analyse factorielle désigne une sous-famille de méthodes de l'analyse des données, aux côtés des méthodes de classification automatique. Le clustering quant à lui est une approche qui vise à segmenter une population étudiée en plusieurs classes ou groupes homogènes. L’objectif est d’identifier et de caractériser les individus de chaque classe selon les besoins de l’analyse. Ces deux méthodes s’inscrivent dans une solide relation de complémentarité. Cependant, elles peuvent être utilisées indépendamment l’une de l’autre ou ensemble selon les besoins et la nature des données.

Dans ce dossier, nous tenterons de comprendre les tenants et aboutissants de chacune de ces méthodes énoncées ci-dessus. Pour cela, nous avons fait le choix d’appliquer ces dernières sur un jeu de données de modèles de véhicules afin d'étudier leurs principes de fonctionnement séparément.

Nous commençons par l’analyse multidimensionnelle, en réalisant une ACP et une AFC, puis nous explorons par la suite quelques méthodes de classification parmi les plus utilisées telles que k-means, k-medoids, CAH… et voir laquelle s’adapte le mieux par rapport à nos données.

I. Présentation du projet

1- Contexte

Dans le cadre de notre Master 1 en informatique, à l'Université de Lyon Lumière II, nous avons été appelés à réaliser un projet de clustering et analyse multidimensionnelle, nous permettant ainsi de mettre en pratique les connaissances acquises dans ce domaine. En effet, notre étude porte sur les analyses factorielles et la classification qui sont deux méthodes complémentaires. Elles permettent d'explorer et de comprendre les données de manière approfondie afin de pouvoir répondre à une problématique bien précise. Ainsi nous avons décidé d'orienter notre étude sur un jeu de données portant sur des voitures de différentes marques.

2- Problématique

L'analyse est le principal outil permettant d'obtenir de l'information à partir des données. Cependant, on est souvent confronté par plusieurs problèmes au niveau des données avant d'arriver à des résultats concluants. Quelques-uns des problèmes les plus courants :

- problèmes de réduction de dimension
- Nettoyages et normalisation des données
- Problèmes liés aux points abérrants
- Données manquantes
- Sensibilités par rapport au phénomène non-linéaire
- Variables non exprimées sur les mêmes unités
- etc.

Compte tenu de notre jeu de données et de nos besoins, nous avons eu à effectuer quelques-uns des traitements énumérés ci-dessus afin de répondre aux questions que nous nous sommes posées.

3- Objectif

Selon les la nature des données, nous nous sommes fixés comme objectifs :

- La réalisation d'une ACP normée en vue de cerner les ressemblances ou dissemblances entre les voitures en se basant sur leur caractéristique. Le choix d'une ACP est dicté par le fait que nos données ne sont pas exprimées sur les mêmes unités.
- La réalisation d'une AFC en vue de sortir les liens pouvant exister entre 2 variables catégorielles.
- L'application d'une méthode classification telles que K-means, k-Medoids, CAH... en vue de répartir les différents modèles de voitures en des groupes homogènes.

II. Compréhension des données

1 - Description des données

Notre jeu de données « cars.csv » provient du site SAS Help Data Sets. Il représente un ensemble de voitures en provenance de trois continents : Asie, Europe et Amérique.

Nos individus statistiques sont les modèles de voitures soit la variable Lib_Model. Les variables sont les caractéristiques des différentes voitures. Elles sont au nombre de 14 variables dont 9 quantitatives et 4 qualitatives. Les variables quantitatives sont :

- MSRP qui correspond au prix de vente conseillé
- Invoice : le prix de vente du véhicule
- Enginesize : la taille du moteur
- Horsepower : la puissance en chevaux
- Mpg_city : la consommation de carburant en ville,
- Mpg_highway : la consommation de carburant sur l'autoroute
- Weight : le poids du véhicule
- Wheelbase : l'empattement (écart entre les roues avant et arrière)
- Length : la longueur du véhicule.

Les variables qualitatives sont :

- Continent : le continent d'origine du véhicule
- Type : le type du véhicule (SUV, Sport…)
- Cylinders : le nombre de cylindres du véhicule
- Lib_Make : la marque du véhicule

Les variables quantitatives représentent nos variables actives et les variables qualitatives représentent nos variables supplémentaires

2 – Prétraitement

Nous avons eu à faire quelques modifications sur nos données afin de pouvoir les exploiter le plus efficacement possible et pour avoir plus de précisions dans nos résultats.

Tout d’abord nous avons vérifié si le fichier ne contenait pas de doublons ou de valeurs manquantes. Ensuite nous avons procédé à une standardisation des nos variables quantitatives afin d’éviter que celles à forte variance pèsent indûment sur nos résultats car elles sont d’unités et d’ordre de grandeurs différentes.

III. Analyse factorielle

1 - L'Analyse en Composantes Principales (ACP)

L'analyse en composantes principales est une méthode de réduction de dimension qui permet de réduire le nombre de variables et de rendre l'information moins redondante.

Ainsi, nous avons utilisé l'ACP pour répondre aux questions suivantes :

- Est-ce que toutes les voitures se ressemblent ?
- Si tel n'est pas le cas, quelles voitures se ressemblent, quelles voitures diffèrent ?
- Est-ce que certaines caractéristiques sont liées ?
- Quels critères peuvent expliquer la taille de la voiture (EngineSize) ?

Nous allons essayer de répondre à toutes ces questions en appréhendant l'information de deux manières différentes :

- Étudier la proximité entre les individus (voitures) d'une part.
- Analyser les liaisons entre les variables d'autre part.

Choix du nombre d'axes *(figure 1 de l'annexe)*

Nous avons retenu les deux premiers axes pour représenter les associations entre les individus et les variables. Cependant on va se concentrer que sur le premier axe car il comporte plus d'inerties (58.6%). Nous allons donner une sémantique à notre axe en analysant les résultats des individus et des variables.

Analyse du résultat des individus *(figure 2,3,5 de l'annexe)*

Le côté positif de l'axe 1 est caractérisé par les modèles de voiture Excursion 6.8 XLT, Escalade EXT, Yukon XL 2500 SLT, H2 etc. Ces modèles de voiture ont à peu près les mêmes caractéristiques, par opposition aux voitures : Land Cruiser et MR2 Spyder convertible. Les voitures 2dr Sierra HD 2500, Highlander V6 et Forester X caractérisent le côté négatif de l'axe 1.

En ajoutant la variable supplémentaire Cylinders à l'ACP (figure 9 de l'annexe), on constate que les voitures sont ordonnées sur l'axe 1 selon le nombre de cylindres. Les voitures avec 3 ,12, 10 cylindres sont moins nombreux et différents des voitures avec des cylindres compris entre 4 et 8 qui se ressemblent plus.

On a procédé de la même manière pour les autres variables supplémentaires. *(cf annexe)*

Analyse du résultat des variables *(figure 13 de l'annexe)*

Le côté positif de l'axe 1 est caractérisé par les variables EngineSize, Weight, Horsepower. Donc du côté positif de l'axe 1, sont projetées les grandes voitures, à poids lourd avec une grande puissance en chevaux. Et du côté négatif de l'axe 1, nous retrouvons des voitures qui sont mal notées en EngineSize, Weight, Horsepower mais bien notées en MPG_City et MPG_Highway. On observe également une forte corrélation entre les variables Weight, EngineSize, Length et Wheelbase d'une part et les variables MGP_City et MPG_Highway d'autre part. Donc une voiture bien notée en Weight et EngineSize sera également bien notée en Length et Wheelbase. Cependant elle est mal notée en MSPR, Invoice.

2 - L'Analyse factorielle des correspondances ou l'AFC

L'analyse factorielle des correspondances est une méthode qui permet d'étudier les liens entre deux variables qualitatives. Elle permet également d'analyser les associations ou correspondances entre les modalités en colonnes et en lignes.

L'objet de notre étude est d'identifier deux variables intéressantes (ici Type et Continent) que nous allons croiser ensemble. Nous avons tout d'abord construit un tableau de contingence avec le croisement de ces deux variables avant de proposer une représentation graphique. Nous étudierons ensuite les ressemblances et les différences entre les profils lignes et les profils colonnes pour étudier le lien entre ces deux variables catégorielles. Nous allons donc répondre aux questions suivantes :

• Quelle est la distribution des types de voitures selon les continents (profils ligne) ?

• Cette distribution diffère-t-elle d'un continent à un autre (distance entre les profils) ?

La question statistique est de savoir si le type de voiture dépend-il du continent, par exemple le type Sports est-il plus présent en Europe ? Nous allons vérifier cela à l'aide du test du Khi-2.

On constate sur la table que l'effectif théorique de la modalité Hybrid est faible (< à 5). Pour éviter que l'AFC ne soit biaisée nous allons l'exclure du tableau de contingence. Deux premières visualisations permettent de voir que les Sedan sont bien présents en Europe, en Asie et aux USA.

Analyse des profils lignes

Les profils lignes obtenus traduisent la distribution du type pour chaque continent. On constate qu'aux Etats-Unis plus de 60% des voitures sont de type Sedan, 17 % de type SUV, 11% de type Truck, 6% de type Sports, et 4.8% de Wagon.

Analyse des profils colonnes

Ils correspondent à la distribution du continent pour chaque Type. On constate que pour le type Sports 34.7% des voitures proviennent d'Asie, 46.9% d'Europe et 18.4% des USA.

Intéressons-nous maintenant à la dépendance entre ces deux variables !

Vu la faible valeur de la p-value (<0.05) on rejette l'hypothèse d'indépendance ; ainsi les deux variables sont liées. En conclusion, les individus n'ont pas la même appétence vis-à-vis des types de voitures d'un continent à un autre. Réalisons une AFC pour expliquer le lien entre ces deux variables.

Choix du nombre d'axes *(figure 14 de l'annexe)*

On remarque que l'axe 1 représente plus de 96% de l'inertie du nuage, ce qui nous conduit à concentrer l'essentiel de nos interprétations sur cet axe.

Interprétation sémantique de l'axe 1 avec les profils lignes *(figure 15 et 16 de l'annexe)*

Regardons les profils qui ont une forte contribution et un cosinus2 élevé sur l'axe 1. Les profils des Etats-Unis et de l'Europe représentent respectivement le côté positif et le côté négatif de l'axe 1. Donc la distribution des types de voitures aux USA n'est pas la même que celle d'Europe.

La représentation graphique sur le plan factoriel consolide notre idée selon laquelle les profils des modèles européens s'opposent aux profils des modèles américains. Que peut-on dire sur les modèles asiatiques ?

Les profils des modèles asiatiques sur l'axe 1 sont projetés près du centre de gravité ou de l'origine des axes qui représente le profil moyen. Cela veut dire que les voitures qui proviennent d'Asie ont une distribution du type de voiture qui ressemble à la distribution du type dans tout le jeu de données.

Interprétation sémantique de l'axe 1 avec les profils colonnes *(figure 17 et 18 de l'annexe)*

Les profils Truck caractérisent le côté positif de l'axe 1 contrairement aux profils Wagon et Sports qui sont eux projetés du côté négatif de l'axe 1. Cela veut dire que la distribution des modèles de voitures pour les types Sport par continent n'est pas la même que pour le type Truck. On peut en déduire que les voitures de type Truck et de type Sport proviennent de continents différents.

Au regard de ces graphiques, nous affirmons les oppositions que nous venons de voir entre les profils de type Sports et ceux de type Truck.

3 - La classification

3.1- CAH

La méthode de **classification hiérarchique** est une approche alternative au clustering partitionnel pour regrouper des objets en fonction de leur similitude. Contrairement aux méthodes k-means et k-medoids, la classification hiérarchique ne nécessite pas de pré-spécifier le nombre de clusters à produire. Le clustering hiérarchique peut être subdivisé en deux types : classification ascendante hiérarchique (CAH) et classification descendante hiérarchique.

Dans le cadre de notre étude, nous avons fait le choix d'appliquer la CAH sur notre jeu de données afin de regrouper les véhicules dans des clusters homogènes. Nous avons utilisé la librairie **factominer** avec laquelle la CAH est réalisée sur les résultats de l'analyse factorielle. Autrement dit, réaliser à priori une ACP et utiliser par la suite des objets résultats de l'analyse factorielle pour réaliser la CAH. En effet, l'ACP peut être faite seulement sur la base des premiers composants conservant la totalité de l'information et non sur l'ensemble des données car les derniers composants peuvent représenter du bruit, et donc les enlever permettra de stabiliser les résultats de la classification.

Après réalisation de l'ACP normée et de la CAH sur les variables synthétiques, nous obtenons un arbre hiérarchique avec les gains d'inertie lors du passage d'une classe à une autre. Ces derniers nous permettent de voir combien de classes on peut conserver pour faire un découpage. En plaçant la croix et en cliquant sur le niveau de coupure proposé, on obtient une projection des points sur le plan factoriel avec une coloration des individus selon leur classe d'appartenance. Le niveau de coupure du dendrogramme nous renvoie ainsi trois classes (voir Markdown).

Maintenant que nous avons obtenu nos différentes classes, nous allons effectuer une description de nos clusters par nos différentes variables. On peut ainsi faire une :

- Description des clusters par les variables qualitatives
- Description des clusters par les variables quantitatives
- Description des clusters par les variables axes factoriels
- Description des clusters par les individus

Pour chaque point ci-dessus, une interprétation complète a été effectuée dans le markdown.

(*cf figure 19 à 24 de l'annexe)*

3.2- K-Means

Après la CAH et les clusters obtenus, nous allons maintenant travailler sur une stratégie de réallocation par agrégation autour de centres mobiles. k-means est une des méthodes de classification automatique les plus couramment utilisées. Elle permet de classer les individus en k groupes, avec k représentant le nombre de groupes prédéfinis. Cette classification est telle que les objets au sein du même groupe sont aussi semblables que possible (autrement dit une forte similitude intra-classe), tandis que les objets provenant de différents groupes sont aussi dissemblables que possible (c'est-à-dire une faible similarité interclasse). Ici, la méthode des k-means va nous permettre de consolider la classification hiérarchique déjà obtenue sur notre jeu de données. Ce qui va nous permettre de classifier les modèles de véhicules de manière plus homogène. La fonction k-means prendra en entrée les données centrées et réduites ainsi que le nombre de clusters permettra de lancer la classification.

(*cf figure 25 à 26 de l'annexe)*

3.3- K-medoids

L'approche des k-medoids est une alternative plus robuste au clustering k-means. Cela signifie que l'algorithme est moins sensible au bruit et aux valeurs aberrantes que les k-moyennes. En effet, il utilise les médoïdes comme centres de cluster au lieu des moyennes (utilisées dans les k-moyennes). L'idée est de déplacer des individus d'un groupe à l'autre afin d'obtenir une meilleure partition. Cela pourrait être mis en œuvre en s'appuyant sur des points représentatifs (médoïdes) dans les classes. L'algorithme Partitioning Around Medoids utilisera soit la distance euclidienne, soit la distance de manhattan. Comme la k-means, la méthode prend également en

entrée des données standardisées avant de chercher à calculer les dissimilarités. On peut avoir en composant les médoïdes des différentes classes. Nous pouvons aussi prolonger l'analyse en regardant le score ou degré d'appartenance à la classe par la représentation de la silhouette et en interprétant les valeurs de chaque groupe.

(*cf figure 27 à 28 de l'annexe)*

Conclusion

L'objectif de ce projet était de réaliser avec plusieurs méthodes multidimensionnelles et à l'aide de la classification une étude sur différents modèles de véhicule afin de faire ressortir une liaison entre elles. Pour cela nous avons sélectionné 3 variables que nous avons jugé assez pertinentes pour expliquer les relations entre nos individus : cylindres, type et continent. Ainsi nous avons grâce à ces différentes méthodes mis en pratique ce que nous avons acquis dans un cas concret. Les méthodes que nous avons utilisées sont les suivantes : l'analyse en composante principale, l'analyse factorielle des composantes et la classification (K-means, K-medoids et CAH).

Les conclusions que nous avons tiré de cette étude sont les suivantes :

Les types de véhicules ont une influence sur la manière dont seront représentées les données dans les axes. Au vu des résultats de l'ACP et de la représentation graphique nous avons tiré plusieurs conclusions notamment le fait que les véhicules de type Hybrid et SUV ont des caractéristiques très opposées les unes des autres de même que pour les véhicules de type sport et truck. Nos analyses nous ont également démontré que plusieurs variables étaient notées de la même manière, la taille du moteur et le poids de la voiture, ceci rentre en total concordance avec la réalité, une voiture qui aura un moteur volumineux pèsera lourd. D'autres variables concernent le prix des véhicules ou encore l'empattement et la longueur de la voiture sont notées de la même manière. Nous avons également pu constater que les types des voitures avaient un lien avec leur continent de provenance. En effet les voitures de type alors qu'au Etats-Unis les voitures les plus vendus sont celles de type Sedan, en Europe se sont les voitures de sport qui sont le plus prisées. On a pu conclure de cela que les individus n'étaient pas attirés par les mêmes types de voiture d'un continent à un autre. Au-delà de ces préférences, on voit une certaine opposition entre les continents dans les types de voitures, les types de voitures européens sont opposés aux types de voitures américains. Quant à la classification, la méthode hiérarchique CAH

Bien que nous ayons effectué une analyse complète à l'aide de plusieurs méthodes, notre jeu de données ne nous permettait pas d'aller au bout de ce que nous voulions entreprendre. En effet on aurait pu réaliser une ACM si nous avions un nombre conséquent de variables qualitatives. Mais également l'algorithme PAM peut être prolongé avec l'approche clara si nous avions eu des grandes bases.