Actes des 27èmes Rencontres de la
Société Francophone de Classification

14-16 septembre 2022

Université Lumière Lyon 2, Lyon

SFC'2022

Abstracts of the 27th Conference of the
French Speaking Society of Classification

Avant-Propos

Le présent recueil contient les résumés des communications présentées lors des 27èmes rencontres de la Société Francophone de Classification (SFC) organisées à l'Université Lumière Lyon 2 du 14 au 16 septembre 2022.

La SFC est une société savante qui privilégie les échanges d'expériences entre chercheurs et utilisateurs de la classification des milieux académiques et industriels.

Cette manifestation scientifique internationale a pour objectif de présenter des résultats récents et des applications originales en classification et en analyse de données sous toutes ses formes, mathématique, informatique et statistique, de favoriser les échanges scientifiques entre ces trois communautés autour de la thématique commune de la classification et de faire connaitre à divers partenaires extérieurs les travaux de ses membres.

Nous sommes d'autant plus heureux que ces rencontres à taille humaine et enrichissantes par ses échanges scientifiques autour de l'analyse des données et de la classification, aient pu reprendre enfin après 2 années d'interruption.

Pour cette 27ème édition, 18 communications ont été retenues, 5 en sessions plénières, 1 session dédiée au prix Simon Régnier et 12 en sessions libres.

Nous tenons à remercier tous les participants, les membres du comité de programme pour leur disponibilité ainsi que tous les membres du comité d'organisation pour l'important travail accompli, sans oublier le soutien du Laboratoire COACTIS et de la Direction de la Recherche (DRED) de l'Université Lumière Lyon 2.

Pascal Préa et Rafik Abdesselam
Présidents des Comités de Programme et d'organisation

Comité de Programme

Président

- Pascal Préa, *École Centrale Marseille*

Membres

- Séverine Affeldt, Université de Paris
- Alexandre Bazin, LIRMM, Montpellier
- Khalid Benabdeslem, LIRIS, Université Claude Bernard Lyon 1
- Karell Bertet, Université de La Rochelle
- Patrice Bertrand, Université Paris Dauphine
- Paula Brito, Université de Porto, Portugal
- François Brucker, École Centrale Marseille
- Véronique Cariou, ONIRIS Nantes
- Victor Chepoi, Aix-Marseille Université
- Guillaume Cleuziou, Université d'Orléans
- Jean Diatta, Université de La Réunion
- Nadia Ghazalli, Université du Québec à Trois-Rivières
- Yann Guermeur, Université de Lorraine, LORIA
- Mehdi Kaytoue, Université de Lyon, INSA, LIRIS
- Pascale Kuntz, Université de Nantes
- Lazhar Labiod, Université Paris Descartes
- Mustapha Lebbah, Université Paris 13
- Vincent Lemaire, Orange Lab
- Ahmed Moussa, ENSA Tanger, Maroc
- Mohamed Nadif, Université Paris Descartes
- Amedeo Napoli, LORIA, Nancy
- Ndèye Niang, CNAM Paris
- Marc Plantevit, EPITA, Lyon
- Allou Samé, Université Gustave Eiffel, Paris
- Arnaud Soulet, Université François Rabelais Tours

Comité d'Organisation

Président

- Rafik Abdesselam, Université Lumière Lyon 2

Membres

- Christine Sybord, Université Lumière Lyon 2
- Nathalie Rivier, Université Lumière Lyon 2
- Charlotte HIPPY, Université Lumière Lyon 2
- Manon Lambert, Université Lumière Lyon 2

Conférences plénières

Introduction à la "s-concordance" et à la "s-discordance" d'une classe avec une collection de classes.

Edwin Diday
CEREMADE, Université Paris Dauphine - PLS, France

Afin d'obtenir des modèles locaux on utilise alternativement des classes et leur représentation (centre de gravité, régression, lois de probabilités, facteurs d'ACP ou d'analyse canonique etc.) de façon à améliorer itérativement leur adéquation. De même la "s-concordance" et la "s-discordance" utilisent ces notions de classes associées à des représentations. Afin de mesurer la concordance ou la discordance d'une classe c avec une collection de classes P, pour une représentation x, nous introduisons d'abord deux fonctions de base. La première, exprime l'adéquation de la représentation x avec c. La, seconde, mesure la proportion de classes c'de P ayant même représentation x avec une adéquation à c'proche de celle de x à c (au sens d'une mesure de voisinage donnée). Nous donnons ensuite les définitions axiomatiques de la s-concordance et de la s-discordance, puis des exemples de familles de s-concordance et de s-discordance. Quelques propriétés de la s-concordance et s-discordance sont indiquées. Nous faisons apparaître des liens qui permettent de considérer les copules, la vraisemblance et la décomposition de mélange dans un cadre plus général pouvant conduire à des algorithmes plus performants. Tout cela ouvre la voie à de nombreuses pistes de recherche, développement et applications.

Mots-clés : Analyse des données symboliques, s-concordance, copulas, vraisemblance, décomposition de mélange de lois de probabilités.

Références

F. Afonso, E. Diday, C. Toque (2018) "Data Science par Analyse des Données Symboliques". Book (448 pages). TECHNIP editor.

Diday, E. (2020) "Explanatory tools for Machine Learning in the Symbolic Data analysis Framework". In: Diday, E.,Guan, R., Saporta, G., Wang, H. (eds.) Advances in Data Science, pp. . Eds. ISTE-Wiley.

E. Diday (2016) "Thinking by classes in data science: the symbolic data analysis paradigm". WIREs Computational Statistics Symbolic Data Analysis Volume 8, September/October 2016. Wiley Periodicals, Inc. 191. First published: 19 August 2016 https://doi.org/10.1002/wics.1384.

Nelsen, R.B. (1998) An introduction to copulas. Lecture notes in Statistics. Springer.

Classification de variables autour de variables latentes (CLV) et extensions

Evelyne Vigneau
Oniris, StatSC, Nantes, France

La méthode de Classification de variables autour de Variables Latentes, CLV, est présentée. Basée sur des critères explicites et un algorithme d'optimisation itératif simple, des extensions de l'approche de base ont été développées pour prendre en compte des structures de données et/ou des problèmes spécifiques. La présentation a pour objet d'en dressera un panorama assez large, illustré par des exemples de mise en œuvre.

Références

Cariou, V., Alexandre-Gouabau, M.-C., Wilderjans, T. F. (2020). Three-way clustering around latent variables (CLV3W) approach with constraints on the configurations to facilitate interpretation. Journal of Chemometrics, 35, e3269.

Llobell, F., Vigneau, E., Cariou, V., Qannari, E.M. (2019). ClustBlock: Clustering of Datasets. R package, version 2.1.1.

Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2020). Analysis and clustering of multiblock datasets by means of the STATIS and CLUSTATIS methods. Application to sensometrics. Food Quality and Preference, 79: 103520.

Vigneau, E. , Qannari, E. M. (2003). Clustering of variables around latent components. Communications in Statistics - Simulation and Computation, 32(4): 1131-1150.

Vigneau, E., Charles, M., Chen, M. (2014). External preference segmentation with additional information on consumers: a case study on apples. Food Quality and Preference, 32: 83-92.

Vigneau, E., Chen, M., Qannari, E. M. (2015). ClustVarLV: An R Package for the Clustering of Variables Around Latent Variables. RJournal, 7, 134-148.

Vigneau E., Chen M. (2016). Dimensionality reduction by clustering of variables while setting aside atypical variables. Electronic Journal of Applied Statistical Analysis, 9(1): 134-153.

Wilderjans, T. F., Cariou, V. (2016). CLV3W: A clustering around latent variables approach to detect panel disagreement in three-way conventional sensory profiling data. Food Quality and Preference, 47, 45-53.

Codage optimal : nouveaux regards sur un ancien problème

Gilbert Saporta
CEDRIC, CNAM

Le traitement de variables qualitatives avec un très grand nombre de catégories en apprentissage automatique est l'occasion de revisiter la théorie du codage optimal et ses applications.

Coder une variable qualitative consiste à attribuer des valeurs numériques à ses modalités, donc à la transformer en une variable numérique discrète. Un codage revient alors à définir les coefficients d'une combinaison linéaire des indicatrices des modalités sous certaines contraintes comme la monotonicité en cas de variables à modalités ordonnées.

La transformation de variables qualitatives en variables quantitatives a une longue histoire remontant à K. Pearson, R.A. Fisher, L. Guttman, C.Hayashi, etc. Elle fut à l'origine de l'analyse des correspondances (Lebart et Saporta, 2014). Les années 70 et le début des années 80 furent celles de la recherche de codages (appelés scores ou scaling) optimaux dans des contextes supervisés ou non supervisés où s'illustrèrent des chercheurs comme J. de Leeuw, S. Nishisato, Y.Takane, M.Tenenhaus, F.Young. On se reportera à Young (1981). Ces recherches furent popularisées par des logiciels : procédures Prinqual et Transreg de SAS, SPSS Categories.

Pendant près de 30 ans, le sujet ne suscita plus guère de recherches ; les applications où on attribue des notes aux catégories des prédicteurs devinrent routinières comme les scores de risque en banque et en assurance.

Avec la disponibilité de données massives, les chercheurs et praticiens de l'apprentissage se sont trouvés confrontés à des données catégorielles, mal adaptées aux réseaux de neurones et possédant des dizaines ou des centaines de catégories (comme des codes postaux par exemple). Voir Hancock & Khoshgoftaar (2020).

Ignorant généralement les travaux des statisticiens, on a vu fleurir différentes méthodes d'encoding essentiellement pour des problèmes supervisés. Di Ciaccio (2022) indique que Scikit-learn propose 17 méthodes différentes qu'il sépare en trois groupes : les méthodes où le codage d'une variable ne dépend pas des autres variables (en particulier de la réponse) comme le Hash encoding, celles où le codage ne dépend que de la réponse (moyenne conditionnelle), et le One-Hot Encoding qui n'est autre que la mise sous forme disjonctive avec autant d'indicatrices que de modalités.

La grande dimension de certaines données catégorielles soulève alors des problèmes de stabilité et de surajustement que l'on négligeait dans les applications statistiques classiques où le nombre de modalités est faible et où la démarche apprentissage-test était peu fréquente.

La confrontation de ces deux mondes permet d'envisager un renouveau des méthodes de codage (voir Meulman et al., 2019).

Références

Di Ciaccio A. (2022). Optimal Coding of categorical data in machine learning. To be published in "Studies in Classification, Data Analysis and Knowledge Organization".

Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. Journal of Big Data, 7(1), 1-41.

Lebart, L., & Saporta, G. (2014). Historical Elements of Correspondence Analysis and Multiple Correspondence Analysis. In Blasius J. & Greenacre, M., editors : Visualization and Verbalization of Data, 73-86. Chapman and Hall/CRC.

Meulman, J. J., van der Kooij, A. J., & Duisters, K. L. (2019). ROS regression: Integrating regularization with optimal scaling regression. Statistical science, 34(3), 361-390.

Young, F. W. (1981). Quantitative analysis of qualitative data. Psychometrika, 46(4), 357-388

Méthodes de régression par clusters pour données massives agrégées sous forme de distributions

Rosanna Verde
Département de Mathematique et Physique
Università della Campania “Luigi Vanvitelli”, Caserta, Italie

L'exposé se concentre sur les développements récents de la régression par clusters lorsque des données massives sont agrégées sous la forme de distributions.

L'ensemble E des objets ei à clustériser est décrit par des variables distributionnelles $\{Y, X1, \ldots, Xp\}$, avec Y la variable de réponse et Xj les prédicteurs. Chaque objet est représenté par p + 1 fonctions de densité de probabilité, ou équivalents fonctions empiriques.

Les Méthodes de Régression par Clusters (MRC) proposées sont basées sur un algorithme de clustering dynamique (CD), où les centroïdes des clusters sont des modèles de régression linéaire pour des données de distributions. Les objets sont affectés aux clusters dont la somme des erreurs au carré du modèle estimé de la classe est minimale. L'objectif est d'étudier une relation de dépendance entre variables distributionnelles lorsque les données sont structurées en classes.

Les méthodes MRC nécessitent de se référer à une méthode de régression pour variables distributionnelles.

En 2015 deux modèles de régression ([1], [2]) ont été proposés tous deux basés sur la distance de Wasserstein en norme carrée dans un espace linéaire, sous une contrainte de non-négativité des paramètres, puis sur un algorithme des moindres carrés non linéaires pour garantir que la variable de réponse soit toujours une variable distributionnelle, Le premier est basé sur une décomposition du modèle, et le seconde sur une transformation symétrique des distributions.

Compte tenu des développements les plus récents en analyse de données distributionnelles (DD), un nouveau modèle de régression introduit une transformation des distributions par les logarithmiques des fonctions quantiles dérivées [3], afin de représenter les fonctions de densité dans un espace de Hilbert et de résoudre certaines issues dans le modèle de régression de DD. La présentation traitera des avantages de cette nouvelle approche de régression en MCR. Dans l'algorithme MRC proposé, l'étape de partitionnement est effectuée selon un algorithme de regroupement dynamique basé sur la minimisation d'une fonction de perte, basée sur la somme des erreurs au carrés dans chaque classe, en utilisant la distance de Wasserstein en norme carrée. L'affectation des éléments aux classes est cohérente avec les modèles de régression et la transformation des fonctions de distributions, de sorte que la convergence est prouvée à une valeur stable.

Certaines applications sur des jeux des données artificielles et réels (par exemple, sur les données COVID-19) permettent de montrer l'efficacité des méthodes MRC proposées.

Références

[1] Irpino, A., Verde, R.: Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance, Advances in Data Analysis and Classification 9 (1) 81-106 (2015)

[2] Dias, S., Brito, P.: Linear regression model with histogram-valued variables, Statistical Analysis and Data Mining: The ASA Data Science Journal 8 (2) 75-113 (2015)

[3] Zhao, Q., Wang, H., Lu, S.: M-LDQ feature embedding and regression modeling for distribution-valued data, Information Sciences, 609, 121-152 (2022)

[4] Petersen, A., Muller, H.: Functional data analysis for density functions by transformation to a Hilbert space, Annals of Statistics 44 (1) 183-218 (2016)

[5] Dias, S., Brito, P.: Analysis of Distributional Data, Chapman and Hall/CRC (2022)

Classification non supervisée de données textuelles

Mohamed Nadif
Centre Borelli UMR 9010, Université Paris Cité, France

La classification non supervisée (ou clustering) est devenue incontournable dans le domaine de l'intelligence artificielle, notamment pour les données textuelles [1]. Plusieurs approches et algorithmes de type apprentissage statistique ou apprentissage profond sont employés pour classifier des documents. Nous passons d'abord en revue différents types de représentation des documents et des termes pour le traitement, puis nous présentons des approches de clustering s'appuyant sur la factorisation matricielle [2, 3, 4, 5], la modularité [6], les modèles probabilistes [7, 8] ou encore la méthode ensemble [9, 10, 11].

Mots-clés : Classification non supervisée, données textuelles

Références

[1] M. Nadif, F. Role. Unsupervised and self-supervised deep learning approaches for biomedical text mining, briefing in Bioinformatics, 22(2): 1592-1603, 2021.

[2] Ch Fettal, L. Labiod, M.Nadif. Efficient Graph Convolution for Joint Node Representation Learning and Clustering. ACM WSDM, 289-297, 2022.

[3] M. Febrissy, A. Salah, M. Ailem, M. Nadif. Improving NMF clustering by leveraging contextual relationships among words. Neurocomputing 495: 105-117, 2022.

[4] S. Affeldt, L. Labiod, M. Nadif, Regularized Bi-Directional Co-Clustering, Statistics and Computing, 31(3):32, 2021.

[5] A. Salah, M. Ailem, M. Nadif. Word Co-occurrence Regularized Non-Negative Matrix Tri-Factorization for Text Data Co-clustering". AAAI: 3992-3999, 2018.

[6] R. Boutalbi, M. Ait Saada, A. Iurshina, S. Staab, M. Nadif. Tensor-based Graph Modularity for Text Data Clustering, ACM SIGIR, 2227-2231, 2022.

[7] P. Riverain, S. Fossier, M. Nadif. Semi-supervised Latent Block Model with pairwise constraints. Mach. Learn. 111(5): 1739-1764, 2022.

[8] A. Salah, M. Nadif. Directional Co-clustering. Advances in Data Analysis and Classification 13(3): 591-6201-30, 2019.

[9] R. Boutalbi, L. Labiod, M. Nadif. Implicit consensus clustering from multiple graphs. Data Mining and knowledge Discovery, 35(6):2313-2349, 2021.

[10] M. Ait Saada, F. Role, M. Nadif. How to Leverage a Multi-layered Transformer Language Model for Text Clustering: an Ensemble Approach. ACM CIKM 2021: 2837-2841, 2021.

[11] S. Affeldt, L. Labiod, M. Nadif. Ensemble block co-clustering: a unified framework for text data. ACM CIKM, 5-14, 2020.

Communications libres

Méthodes de classification symbolique appliquées aux estimations d'intervalles de quantiles de coûts de production

Dominique Desbois
AgroParisTech, INRAE - Université Paris-Saclay, France

Cette communication utilise la classification des données symboliques pour explorer les similitudes entre distributions d'estimations quantiles conditionnelles, en l'appliquant au problème de l'allocation des coûts spécifiques en agriculture. Après avoir rappelé le cadre conceptuel de l'estimation des coûts de production agricole, la première partie présente le modèle empirique, l'approche de régression quantile et la technique de classification des données d'intervalle utilisée. La seconde partie présente l'analyse comparative entre douze États membres européens des résultats issus de la classification hiérarchique divisive des intervalles d'estimation, appliquée à l'estimation du coût des fertilisants.

Modèles de mélanges finis pour une distribution de loi Beta sous-jacente avec une application à des données sur la COVID-19

Cédric Noel[12], Jang Schiltz[2]
[1]Université de Lorraine, IUT de Thionville-Yutz, France
[2]Université du Luxembourg

Nous introduisons une extension du modèle de mélanges finis de Nagin à des données qui suivent une distribution de loi Beta et nous présentons notre paquet R trajeR qui permet de calibrer ce genre de modèles.
Dans une deuxième partie de l'article, nous utiliser ce modèle pour analyser l'efficacité des mesures sanitaires prises dans les différents pays pendant les premiers seize mois de la pandémie de COVID-19.

Classification et décomposition de séries temporelles pour l'analyse de données thermiques de bâtiments

Louise Bonfils, Allou Samé, Latifa Oukhellou
COSYS-GRETTIA, Université Gustave Eiffel, Marne-la-Vallée, France

Cet article aborde la problématique de la classification non supervisée de séries temporelles dépendant de multiples facteurs explicatifs observés et non observés. Les motivations applicatives de ce travail concernent principalement la caractérisation du comportement dynamique d'occupants de bâtiments à partir de mesures thermiques (ex. température intérieure et extérieure, ensoleillement).
Dans cette optique, nous proposons un modèle de classification qui s'inscrit dans la lignée des modèles probabilistes à variables latentes tels que les mélanges de régressions. La spécificité du modèle proposé réside dans le fait qu'il permette, en plus du partitionnement de séries temporelles, de décomposer chaque classe en deux composantes, l'une traduisant l'effet de variables explicatives connues, et l'autre l'effet de facteurs dynamiques inconnus. L'estimation des paramètres du modèle repose sur une variante variationnelle de l'algorithme EM.

Classification par recouvrement d'X-arbres

François Brucker
LIS, ECM, Marseille, France

Nous montrons dans cette communication une méthode de classification utilisant un recouvrement d'X-arbres. Cette méthode permet de décrire parfaitement les données tout en évitant l'explosion combinatoire des classes.

Clustering multi-tranche pour les tenseurs d'ordre

Dina Faneva Andriantsiory[1], Joseph Ben Geloun[1], Mustapha Lebbah[2]
[1]LIPN-UMR 7030, Institut Galilée – Université Paris 13, France
[2]LIPN, Institut Galilée, Université Sorbonne Paris Cité (USPC), université Paris 13, France

Nous proposons une nouvelle méthode appelée "Multi-Slice Clustering" (MSC) ou Clustering multi-tranche pour les données tensorielles d'ordre 3. Nous proposons d'analyser pour chaque dimension, les composantes principales des colonnes de chaque matrice de tranche. Nous définissons ainsi une mesure de similarité entre les différentes tranches, permettant ainsi d'identifier un cluster comme un ensemble de tranches fortement similaires. L'intersection des trois clusters de chaque dimension représente le clustering du tenseur. L'efficacité de notre algorithme est évaluée sur des données synthétiques.

Classification Topologique d'Individus

Rafik Abdesselam
ERIC – COACTIS, Université Lumière Lyon 2, France

La classification d'objets-individus est l'une des approches les plus utilisées pour explorer des données multidimensionnelles. Les deux stratégies non supervisées courantes sont la classification ascendante hiérarchique (CAH) et les k-means, utilisées pour identifier des groupes d'objets similaires dans un ensemble de données afin de le diviser en groupes homogènes. L'objectif est d'établir une typologie, ou segmentation, c'est-à-dire une partition, ou répartition des individus en classes homogènes, ou catégories. La Classification Topologique d'Individus proposée, notée CTI, étudie un ensemble homogène d'individus-lignes d'un tableau de données, elle est basée sur la notion de graphes de voisinage. Les colonnes-variables sont plus ou moins corrélées ou liées selon que les variables sont quantitatives ou qualitatives ou encore mixtes. Elle est illustrée ici à l'aide d'un jeu de données réelles continues, cependant, elle peut être également appliquée avec des données binaires ou mixtes.

Détection d'observations atypiques dans des données distributionnelles multivariées

Ana Martins, Paula Brito, Sonia Dias, Peter Filzmoser
Université de Porto, Portugal

Nous considérons des données numériques distributionnelles, i.e., des données pour lesquelles une distribution empirique sous forme d'histogramme est enregistrée pour chaque unité et chaque variable descriptive. Nous proposons une mesure pour la détection des valeurs aberrantes multivariées, qui repose sur des projections unidimensionnelles et utilise des distances appropriées pour de telles données. La méthodologie proposée est valuée à l'aide de données simulées dans différentes configurations.

Identification de sous-espaces caractéristiques de classes issues de K-Means parcimonieuse.

Mory Ouattara[1], Abdoul Wahab Diallo, Ndèye Niang[2]
[1]Université de San Pedro, Côte d'ivoire
[2]CEDRIC, CNAM, Paris, France

Dans ce papier nous abordons les méthodes de classification adaptées aux données de grande dimension, plus précisément lorsque les individus sont décrits par des sous-espaces de variables. Nous proposons une nouvelle approche de sparse subspace clustering appelée Sparse Subspace K-means (SSKM) qui est basée sur une modification de la fonction de coût d'une version Sparse de l'algorithme K-means. La méthode proposée est illustrée sur des données simulées et sur un jeu de données réelles. Dans sa comparaison avec les méthodes de la littérature, SSKM se montre aussi bonne ou meilleure tant au niveau des indices de qualité de partition que de la détection de variables pertinentes.

Approximation Robinsonienne sur les PQ-arbres

Pascal Préa, François Brucker
LIS, Universités d'Aix-Marseille Université et de Toulon, École Centrale Marseille, France

Une dissimilarité D sur un ensemble S est Robinson si il existe un ordre total sur S tel que ∀x, y, z ∈ S, x < y < z ⟹ D(x, z) ≥ max{D(x, y), D(y, z)}. Un tel ordre est dit Robinson ou compatible (avec D). Un PQ-arbre sur S est un arbre qui représente un ensemble de permutations de S. Étant donnée une dissimilarité de Robinson D, l'ensemble des ordres compatibles avec D peut être représenté par un PQ-arbre.
Dans ce papier, nous considérons le problème suivant : étant donnés une dissimilarité D et un PQ-arbre T sur un ensemble S, approximer D en une dissimilarité de Robinson R telle que tous les ordres représentés par T soient compatibles avec R.
Nous montrons que, dans la plupart ces cas, ce problème peut se ramener à une régression isotonique et nous donnons des algorithmes efficaces (les complexités varient entre O(n2) et O(n3 log3 n)) pour résoudre les différentes versions (approxier selon la norme L1, L∞, . . .). En plus, il est possible d'améliorer incrémentalement la solution obtenue. On obtient ainsi une dissimilarité de Robinson R, plus proche de D, mais tous les ordres représentés par T ne sont pas compatibles avec R.

Comparaison des approches de clustering dans le cadre de la méthode Partial Least Squares Path Modeling (PLS-PM)

Sophie Dominique[12], Mohamed Hanafi[3], Fabien Llobell[4], Jean-Marc Ferrandi[5], Véronique Cariou[1]
[1]StatSC, ONIRIS, INRA, Ecole Nationale Vétérinaire, Agroalimentaire et de l'Alimentation Nantes-Atlantique, Nantes, France, [2]Addinsoft, XLSTAT, [3]StatSC, Oniris, Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), Nantes, France, [4]Addinsoft, XLSTAT, Paris, France, [5]LEMNA, Oniris, France

Les modèles à équations structurelles sont utilisés pour analyser l'ensemble des relations entre des variables manifestes mesurables et des concepts latents non observés. L'approche PLS-PM (Partial Least Squares Path Modeling) est une des méthodes d'estimation qui a gagné en popularité ces dernières années. L'hétérogénéité présente dans les données est supposée négligeable lorsque cet algorithme est appliqué à l'ensemble des individus. Toutefois, cette hypothèse n'est pas réaliste notamment dans les sciences sociales car les individus sont susceptibles d'avoir des comportements différents. Généralement, il existe une hétérogénéité non observée qui peut compromettre la validité et l'interprétabilité des résultats. Pour détecter cette hétérogénéité, différentes méthodes de classification peuvent être utilisées pour définir des classes homogènes partageant les mêmes relations (FIMIX-PLS, REBUS-PLS, PLS-POS, PLS-GAS, PLS-IRRS, PLS-SEM KMEANS, ...). Ce travail vise à présenter et comparer les différentes approches d'un point de vue théorique et pratique à partir d'une application marketing.

Algorithme d'identification des variations de signatures génomiques au sein des séquences virales

Dylan Lebatteux[12], Abdoulaye Baniré Diallo[1], Hugo Soudeyns[2], Isabelle Boucoiran[2], Soren Gantt[2]
[1]Laboratoire de bioinformatique, Université du Québec à Montréal, Canada
[2]Centre de recherche du CHU Sainte-Justine, Montréal, Canada

Dans le cadre de la surveillance des agents pathogènes, l'identification et la caractérisation des signatures génomiques virales sont des tâches essentielles. Ces signatures sont des séquences nucléotidiques spécifiques à des groupes de virus qui sont omniprésentes dans leur génome. Elles peuvent contribuer aussi aux études taxonomiques et phylogénétiques, la reconnaissance de variants émergents ou encore l'identification des différences phénotypiques entre les espèces. L'identification de ces variations en fonction des groupes de virus étudiés nécessite, dans les principales études, des alignements multiples et des analyses manuelles. Cet article présente KANALYZER, un algorithme automatisant l'identification des variations de signature génomique dans les séquences nucléotidiques. KANALYZER a été évalué sur des ensembles de séquences virales, ou il a identifié avec succès plus de 96% des variations de signatures discriminants les classes de virus.

Un algorithme simple et optimal pour la sériation circulaire stricte

Mikhaël Carmona[12], Victor Chepoi[1], Guyslain Naves[1], Pascal Préa[12]
[1]LIS, Aix-Marseille Université, CNRS & Université de Toulon, Marseille, France
[2]École Centrale Marseille, Marseille, France

La sériation circulaire (stricte) (ou reconnaissance des espaces de Robinson circulaires (stricte)) consiste, étant donné un ensemble X et une dissimilarité d à déterminer si les points de X peuvent être disposés sur un cercle de manière compatible avec d. Il existe plusieurs versions de ce problème.
Nous présentons un algorithme en *O(n log n)* qui calcule un ordre (circulaire) compatible pour la sériation circulaire stricte.
En complétant cet algorithme par une étape de vérification en *O(n^2),* on obtient un algorithme optimal qui résout toutes les versions de la sériation circulaire stricte.
Nous montrons aussi que les dissimilarités de Robinson circulaires (définies par l'existence, pour chaque paire de points, d'un ordre compatible sur un de deux arcs joignant ces points) sont exactement les dissimilarités de Robinson pré-circulaires (définies par une condition sur quatre points).

Contributions à la découverte de clés de liage dans des jeux de données RDF

Nacira Abbas[1], Alexandre Bazin[2], Jérôme David[3], Amedeo Napoli[1]
[1]INRIA, LORIA, Université de Lorraine Nancy, France
[2]LIRMM, Université de Montpellier, France
[3]INRIA, Grenoble INP, LIG, Université Grenoble Alpes, France

Une clé de liage entre deux jeux de données RDF D_1 et D_2 est un ensemble de paires de propriétés qui permet d'identifier des individus deux à deux, par exemple x_1 dans D_1 et x_2 dans D_2, et qui peut se matérialiser par un lien d'identité x_1 owl:sameAs x_2. Il existe plusieurs façons de découvrir des clés de liage, mais jusqu'à présent, aucune des méthodes ne prend en compte le fait que owl:sameAs est une relation d'équivalence. Si cela est fait il est possible de découvrir des clés de liage "non redondantes". Ainsi, dans cet article, nous présentons une méthode de découvertes de clés de liage qui s'appuie sur les "pattern structures" (PS), un extension de l'analyse formelle de concepts ou FCA ("Formal Concept Analysis"). Les PS permettent de construire un treillis de concepts où chaque concept a un extent représentant un ensemble de paires d'individus et un intent représentant les clés de liages candidates associées. De plus, nous étudions précisément la relation équivalence induite par une clé de liage et nous introduisons la notion de clé de liage non redondante.

Index des auteurs