Journées de méthodologie statistique de l'Insee 2022

PERSPECTIVES FUTURES EN COUPLAGE D'ENREGISTREMENTS

Gautier Gissler (*) et Abel Dasylva (*)

(*) Statistique Canada, Direction de la méthodologie

gautier.gissler@statcan.gc.ca
abel.dasylva@statcan.gc.ca

Mots-clés : classification non supervisée, couplage probabiliste, Fellegi-Sunter, k-moyennes, estimation de seuils, estimation des erreurs.

Domaine concerné : appariement probabiliste, classification

Résumé

Plusieurs recherches sont en cours pour améliorer le couplage probabiliste à l'aide de techniques d'apprentissage automatique et développer des modèles pour les erreurs associées.

Pour le couplage probabiliste, les techniques d'apprentissage automatique sont d'intérêt pour la phase d'indexation ou celle de classification de paires. Cependant le plus grand défi consiste sûrement en le manque ou l'absence de données d'entraînement, indispensables pour l'utilisation d'algorithmes d'apprentissage supervisé. Pour surmonter ce problème, Christen [1] a, entre autres, proposé une méthode de classification en deux étapes qui a été adaptée dans le logiciel généralisé de couplage probabiliste développé à Statistique Canada, G-Coup [2], en tant qu'estimateur automatique de seuil.

Dans la méthodologie d'appariement probabiliste sous-jacente à G-Coup, fondée sur la théorie de Fellegi-Sunter [3], la classification des paires potentielles en trois ensembles (rejetées, possibles et définitives) dépend d'une part de la bonne estimation des poids des paires (basé sur un ratio de probabilités : probabilité d'être un *match* sur la probabilité d'être un *unmatch*) et d'autre part de la bonne estimations de deux seuils (inférieur T_λ et supérieur T_μ) qui délimitent ces trois zones en fonction des poids des paires. Ainsi, une paire de poids w sera rejetée si $w < T_\lambda$, possible si $T_\lambda \leq w < T_\mu$, et définitive si $w \geq T_\mu$. En pratique, on observe qu'une bonne classification n'est pas obtenue dès le premier coup et qu'il faut itérer le processus de calcul des poids (dont le dénominateur, la probabilité d'être un *match*, est estimée à partir des paires non-rejetées, ensemble qui est affecté par la précédente classification), de choix des seuils et de reclassification à partir de ces derniers pour raffiner le couplage.

L'étape du choix des seuils est donc essentielle puisque d'elle découle la classification des paires et le calcul de nouveaux poids à l'itération suivante. Cependant, la détermination des seuils reste un défi pour le couplage probabiliste et plusieurs méthodes peuvent être utilisées pour réaliser cette étape. L'une d'entre elles, utilisée en pratique par les méthodologistes de l'Environnement des Couplages des Données Sociales à Statistique Canada pour sa fiabilité expérimentée, s'appuie sur l'examen des paires et profils de concordances et nécessite donc l'intervention d'un expert en couplage.

Les estimateurs automatiques de seuils ont, eux, pour objectif de déterminer les seuils inférieurs et supérieurs sans intervention humaine pour un couplage similaire à ceux effectués en pratique (ex : dans le cadre de l'ECDS), en ayant donc à disposition que les données du couplage (non-supervisées), c'est-à-dire les paires potentielles, leur précédent statut de classification (si disponible) et leur vecteur de poids. Différentes méthodes d'apprentissage automatique peuvent être utilisées, en particulier dans le cadre de l'estimateur en deux étapes (partie non-supervisée qui permet de créer un ensemble d'entraînement, généralement procédée avec k-moyennes, puis partie supervisée). Le choix des paramètres, ainsi que la question de la représentation des données (utilisation de tout l'ensemble des paires potentielles ? quelles composantes du vecteur de poids garder ? notion de distance entre les paires ?) entrent alors en jeu. De plus, une propriété désirable de ces estimateurs est la convergence des poids (et la classification des paires potentielles induite) lorsque leurs seuils obtenus sont utilisés au fur et à mesure des itérations. Également, le résultat final du couplage utilisant ses estimateurs automatiques doit satisfaire des critères de qualité (pouvant être exprimé à partir des taux de faux positifs $\hat{\mu}$ et taux de faux négatifs $\hat{\lambda}$). Finalement, le développement d'un estimateur automatique de seuil peut éventuellement amener à une méthode alternative de classification des paires qui pourrait être aussi considérée et évaluée.

Les recherches portent aussi sur des modèles pour estimer les taux d'erreurs sans vérification manuelles et sans données d'apprentissage. En effet, la mesure de ces taux est indispensable lorsque le couplage d'enregistrements sert à produire des statistiques officielles. Récemment, un modèle a été proposé à partir du nombre d'enregistrements adjacents à un enregistrement donné [4][5], soit n_i pour l'enregistrement i d'une première source donnée qui est appariée à une seconde source contenant la première. Sous des conditions générales, lorsque les sources sont grandes et sans doublons et qu'elles sont liées de sorte que la décision de lier deux enregistrements ne fasse pas intervenir d'autres enregistrements, il est possible de modéliser n_i par un mélange de distributions de la forme suivante.

$$n_i \qquad \sum_{g=1}^{G} \alpha_g \, Bernoulli\left(p_g\right) * Poisson\left(\lambda_g\right),$$

où ¿ dénote l'opération de convolution et les paramètres sont estimés par la maximisation de la vraisemblance composite des n_i. Ce modèle généralise le modèle de Blakely et Salmond [6] et possède de nombreux avantages car il prend en compte toutes les interactions entre les variables de façon implicite. En outre il s'applique aux couplages probabilistes, déterministes et hiérarchiques.

Bibliographie

[1] Christen, P. (2007). A two-step classification approach to unsupervised record linkage In Proc. Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Australia. CRPIT, 70. Christen, P., Kennedy, P.J., Li, J., Kolyshkina, I. and Williams, G., J., Eds. ACS. 111-119.

[2] Guide de l'utilisateur de G-Coup 3.5 (2021)

[3] Fellegi, I., and Sunter, A. (1969). « A theory of record linkage », Journal of the American Statistical Association, vol. 64, pp. 1183-1210.

[4] Dasylva, A. and Goussanou, A. (2020). Estimating linkage errors under regularity conditions, Proceedings of the Survey Methods Section, American Statistical Association.

[5] Dasylva, A., et Goussanou, A. (2021). "Estimer les faux négatifs dus aux pochettes dans le couplage d'enregistrements", Techniques d'enquêtes, vol. 47.

[6] Blakely, T., and Salmond, C. (2002). "Probabilistic record linkage and a method to calculate the positive predicted value", Journal of Epidemiology, 31, 1246-1252.