Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes : l'adaptation à un monde en évolution

Le codage automatique des professions

Arne Bethmann, Malte Schierholz, Knut Wenzig et Markus Zielonka[1]

Résumé

En Allemagne, le codage des professions est effectué principalement en utilisant des dictionnaires suivies d'une révision manuelle des cas qui n'ont pas pu être codés. Puisque le codage manuel est coûteux, il est souhaitable de coder le plus de cas possible automatiquement. Parallèlement, le codage automatique doit atteindre au moins le même niveau de qualité que le codage manuel. À titre de solution possible, nous employons divers algorithmes d'apprentissage automatique pour effectuer la tâche en utilisant une quantité importante de professions codées manuellement dans le cadre d'études récentes comme données d'apprentissage. Nous déterminons la faisabilité de ces méthodes en évaluant la performance et la qualité des algorithmes.

Mots-clés : Codage des professions; apprentissage automatique; bayésien naïf; catégorique bayésien

1. Introduction

Ces dernières années, plusieurs études par panel à grande échelle réalisées en Allemagne ont révélé qu'il existait une demande de codage des réponses aux questions ouvertes sur les professions (p. ex., l'étude par panel nationale sur l'éducation (NEPS), le panel socioéconomique allemand (SOEP) et l'enquête par panel sur le marché du travail et la sécurité sociale (PASS)). Jusqu'à présent, en Allemagne, le codage des professions a été effectué principalement de façon semi-automatique, en employant des approches par dictionnaire suivies d'un codage manuel des cas qui ne peuvent pas être codés automatiquement.

Comme le codage manuel des professions entraîne des coûts considérablement plus élevés que le codage automatique, il est fort souhaitable d'accroître la proportion du codage qui peut être effectuée automatiquement. Parallèlement, la qualité du codage est de la plus haute importance et requiert un examen minutieux. La qualité du codage automatique doit atteindre au moins le niveau de la qualité du codage manuel afin de préserver un bon ratio coût d'enquête versus erreur d'enquête. Considérant l'erreur d'enquête globale, cela libérerait des ressources consacrées auparavant à la réduction de l'erreur de traitement et permettrait d'employer ces ressources pour réduire les autres sources d'erreur.

Contrairement aux approches par dictionnaire, qui sont utilisées principalement pour le codage automatique des professions dans les enquêtes allemandes, nous employons deux algorithmes d'apprentissage automatique (c.-à-d. l'algorithme *bayésien naïf* et l'algorithme *bayésien multinomial*) pour effectuer la tâche. Puisque nous disposons d'une quantité importante de professions codées manuellement lors d'études récentes, nous les utilisons comme données d'apprentissage pour la classification automatique. Cela nous permet d'évaluer la performance ainsi que la qualité — et donc la faisabilité — des algorithmes d'apprentissage automatique pour accomplir la tâche du codage automatique des réponses aux questions d'enquête ouvertes sur les professions.

[1] Arne Bethmann, Institute for Employment Research, Regensburger Straße 104, Nuremberg, Allemagne, 90478 (arne.bethmann@iab.de); Malte Schierholz, Mannheim Centre for European Social Research, University of Mannheim, Mannheim, Allemagne, 68131, (malte.schierholz@mzes.uni-mannheim.de); Knut Wenzig, Allemagne Institute for Economic Research, Berlin, Allemagne, 10108 (kwenzig@diw.de); Markus Zielonka, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, Bamberg, Allemagne, 96047 (markus.zielonka@lifbi.de)

2. Algorithmes

Jusqu'à présent, les algorithmes d'apprentissage automatique n'ont été appliqués au codage des professions que par quelques institutions (voir, p. ex., Thompson, Kornbau et Vesely 2012). L'approche que nous utilisons ici a été mise en œuvre par Schierholz (2014) en utilisant deux ensembles de données d'enquête recueillis par l'Institut allemand de recherche sur l'emploi (IAB).

La tâche de codage des réponses aux questions d'enquête ouvertes sur les professions en utilisant une approche d'apprentissage automatique peut être décrite en gros par la procédure en trois étapes suivante :

1. trouver des attributions de catégorie antérieures dans les données d'apprentissage pour chaque cas figurant dans les données à coder en se basant sur la chaîne de texte provenant de la question d'enquête ouverte q_i et éventuellement des variables auxiliaires x_i;

2. estimer les probabilités d'exactitude $\hat{P}_{cor}$ pour chaque catégorie de profession c_j sachant l'information présente dans les données d'apprentissage :
$$\hat{P}_{cor}(c_j|q_i, x_i);$$

3. attribuer une (ou plusieurs) catégories à chaque cas en se basant sur les probabilités d'exactitude.

Le premier algorithme utilisé est basé sur l'approche bayésienne naïve (BN) bien connue :

$$\hat{P}_{cor}(c_j|q_i, x_i) \propto \hat{P}(c_j) \times \hat{P}(x_i|c_j) \times \hat{P}(q_i|c_j)$$
$$\propto \hat{P}(c_j) \times \hat{P}(x_i|c_j) \times \prod_{v=1}^{V} (0.95\hat{P}(T_v|c_j) + (1 - 0.95)\hat{P}(T_v))^{w_{iv}}$$

Toutes les estimations $\hat{P}$ dans les formules susmentionnées sont calculées à partir de fréquences relatives des données d'apprentissage. La première formule est dérivée du théorème de Bayes et l'hypothèse douteuse, « naïve », selon laquelle les covariables X et les chaînes de texte Q (avec leurs réalisations respectives x_i et q_i) sont stochastiquement indépendantes en fonction des catégories cibles c_j. Comme il est fréquent de ne pas trouver de concordances exactes avec les réponses textuelles dans les données d'apprentissage, les chaînes q_i sont subdivisées en mots multiples T_v. $\hat{P}(q_i|c_j)$ est alors calculée comme le produit des probabilités estimées qu'un terme T_v soit utilisé par un répondant si la catégorie c_j est correcte. L'autre algorithme est basé sur une analyse bayésienne conjuguée à la loi multinomiale (BMN). La probabilité a posteriori s'évalue alors par

$$\hat{P}_{cor}(c_j|q_i) = (1 - \omega)\hat{P}(c_j) + \omega\hat{P}(q_i|c_j).$$

Ici, $\hat{P}_{cor}$ est une moyenne pondérée de la fréquence relative pour chaque catégorie et la fréquence relative de la catégorie sachant qu'exactement la même réponse a été fournie dans les données d'apprentissage. Les poids ω sont plus grands quand la réponse q_i figure plus souvent dans les données d'apprentissage :

$$\omega = \frac{\#\{q_i\}}{\#\{q_i\} + 0.5}.$$

Une difficulté importante du codage automatique tient au fait que la plupart des emplois sont rares dans la population et que de nombreuses formulations textuelles sont possibles pour spécifier le même code d'emploi. Par conséquent, l'estimateur MV $\hat{P}_{cor}(c_j|q_i) = \frac{\#\{c_j, q_i\}}{\#\{c_j\}}$ sera égal à un si la réponse q_i ne figure qu'une seule fois dans l'ensemble de données d'apprentissage. Il s'agit manifestement d'une situation non souhaitable, parce que si cela était vrai, nous coderions chaque réponse q_i dans la catégorie c_j. L'algorithme BMN résout ce problème en réduisant la pondération de l'estimateur MV jusqu'à une valeur plus raisonnable.

Une moyenne pondérée est également incluse dans la formule BN, mais les poids ne sont pas une fonction des fréquences de terme T_v et la réduction de la pondération souhaitée n'est pas bien effectuée. Un avantage de l'algorithme BN est plutôt que les réponses q_i sont subdivisées en mots $\#\{T_v\}$, ce qui permet de prédire les catégories lorsqu'aucune réponse donnant une concordance exacte n'est trouvée. Des descriptions plus détaillées de ces algorithmes figurent dans Schierholz (2014).

3. Données

L'ingrédient essentiel de bons résultats d'apprentissage automatique consiste en une grande quantité de données de haute qualité. Pour les analyses effectuées ici, nous utilisons des données provenant de l'étude par panel nationale allemande sur l'éducation (NEPS; Blossfeld, Roßbach et Maurice 2011). Les données comprennent les réponses codées à diverses questions ouvertes sur les professions et les aspirations professionnelles pour toutes les cohortes de départ de la NEPS. On s'est servi pour le codage de la classification allemande des professions mise à jour récemment « Klassifikation der Berufe 2010 » (KldB2010; Bundesagentur für Arbeit 2011). Jusqu'à présent, le processus de codage comprend l'utilisation à grande échelle de dictionnaires et de systèmes de suggestions automatisés développés au Centre de données de recherche de la NEPS.

Toutes les réponses aux questions ouvertes ont été codées manuellement. Mais les personnes chargées du codage se sont appuyées sur des suggestions générées par ordinateur dérivées de matériel codé antérieurement issu des vagues précédentes de l'étude, à savoir les termes de classification et mots-clés officiels fournis par le Bureau fédéral de l'emploi (BA). Dans les cas de chaînes quasi identiques, une seule suggestion a été présentée. Dans une deuxième boucle, un superviseur chevronné a vérifié les résultats. Les étapes manuelles sont appuyées par les capacités du moteur de recherche du Bureau fédéral allemand de l'emploi. Malgré le système de suggestions effectué par ordinateur, le processus tout entier dépend encore du codage et de la vérification manuelle pour s'assurer d'une bonne qualité des données (Munz, Wenzig et Bela, à paraître).

Ces données de la NEPS comprenant plus de 300 000 réponses déjà codées représentent une source idéale de données d'apprentissage (où les résultats codés sont utilisés comme données d'entrée aux algorithmes) et de données tests (où les codes existants sont utilisés pour vérifier les résultats des algorithmes de codage).

4. Résultats préliminaires

Afin de comparer la performance de la classification, nous avons exécuté les deux algorithmes sur des sous-ensembles des données d'apprentissage en utilisant diverses tailles d'échantillon. La figure 1 donne les résultats pour les deux algorithmes en utilisant 300 000 cas dans les données d'apprentissage et un échantillon de 7 500 cas comme données de test devant être codées[2].

L'axe des x représente le *taux de production*, c'est-à-dire le pourcentage de cas de test qui ont été assignés automatiquement à une catégorie. Pour ces taux de production, les *taux de concordance cumulés* correspondants sont représentés sur l'axe des y. Les taux de concordance cumulés représentent le pourcentage de cas codés correctement, ce qui signifie que la classification automatique effectuée par l'algorithme et la classification manuelle d'après les données d'apprentissage concordent. En principe, le taux de concordance devrait diminuer lorsque le taux de production augmente. Cela tient au fait que l'algorithme code les cas faciles pour commencer[3]. L'augmentation du taux de production force alors l'algorithme à classer également les cas présentant plus d'incertitude, ce qui accroît la quantité d'erreurs de classification.

En ce qui concerne la classification de haute qualité, l'approche BMN produit des résultats considérablement meilleurs. Pour un taux de production de 50 %, le taux de concordance est d'environ 94 % pour l'approche BMN et

[2] À l'heure actuelle, il s'agit des résultats pour un échantillon unique. Des vérifications transversales sont prévues et les résultats seront publiés dans l'avenir.

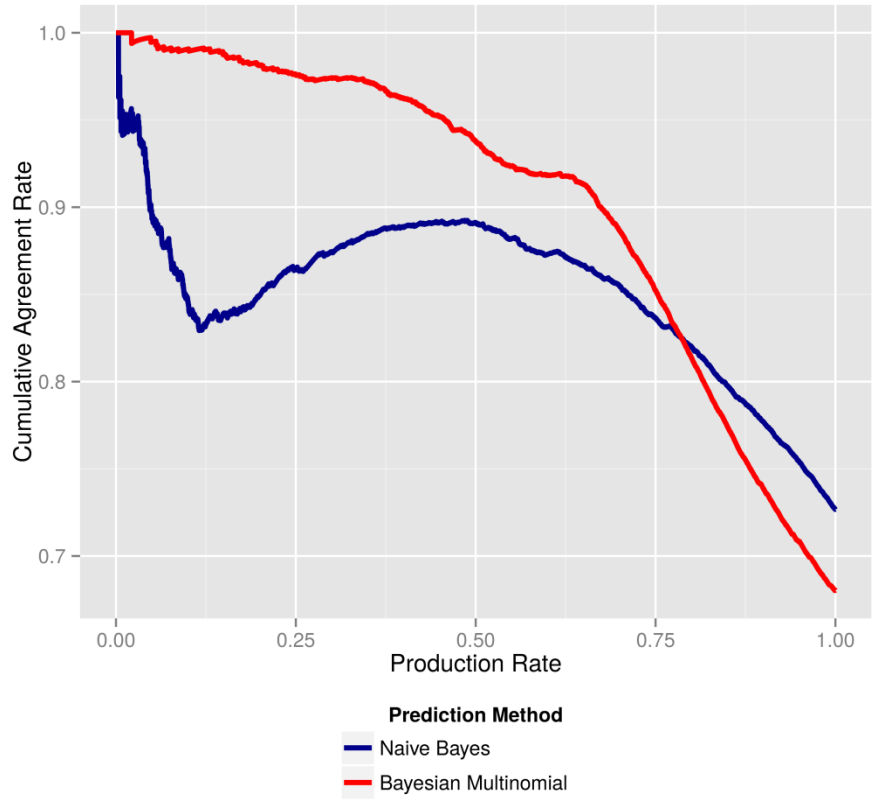
[3] C'est-à-dire les cas ayant une forte probabilité d'être corrects quelle que soit la classe.

légèrement inférieur à 90 % pour l'approche BN. Jusqu'à un taux de production d'environ 80 %, l'algorithme BMN donne de meilleurs résultats que l'algorithme BN. À ce stade, les deux algorithmes donnent un taux de concordance d'environ 83 %. Aux taux de production plus élevés, l'algorithme BN donne des résultats un peu meilleurs, le taux de concordance étant d'environ 73 % comparativement à 68 % seulement pour l'algorithme BMN. Les taux de concordance sont généralement plus faibles pour les plus petits ensembles de données d'apprentissage, mais la performance relative des deux algorithmes est très semblable pour les diverses tailles d'échantillon.

Souvent, ce n'est pas le niveau de concordance à un certain taux de production qui présente un intérêt pour l'évaluation de l'algorithme de codage, mais plutôt le taux de production qui pourrait être atteint à un taux de concordance fixe. Pour le codage automatique des données de l'American Community Survey (ACS), seul un taux de concordance d'au moins 95 % est considéré acceptable (Thompson, Kornbau et Vesely 2012). Dans le tableau 1, nous comparons la performance des deux algorithmes au niveau de concordance de 95 % et au niveau — un peu moins exigeant — de 90 %.

Aux deux niveaux, la performance de l'algorithme BN est moins bonne que celle de l'algorithme BMN. Donnant systématiquement des taux de production inférieurs à 5 %, l'algorithme BN semble inutilisable à toutes fins pratiques. En outre, si l'on examine la figure 1, il semble que des niveaux légèrement plus faibles de taux de concordance fixes produiraient vraisemblablement des taux de production acceptables.

Figure 1
Taux de production



Données de la NEPS; $N_{Test} = 7\,500$***;*** $N_{Apprentissage} = 300\,000$

Tableau 1
Taux de production

	Taux de concordance fixes			
	90 %		95 %	
$N_{Apprentissage}$	BN	BMN	BN	BMN
25 000	0,5	51,5	0,3	36,4
50 000	0,9	55,6	0,1	38,1
100 000	1,1	60,2	0,3	41,9
300 000	4,9	67,4	3,1	45,8

Données de la NEPS; $N_{Test} = 7\,500$; BN = bayésien naïf; BMN = bayésien multinomial

La performance en fonction de la taille d'échantillon évolue comme prévu : les grands ensembles de données produisent de meilleurs résultats de classification. Clairement, pour les applications pratiques, les tailles d'échantillon de données d'apprentissage devraient se situer dans la tranche des grands nombres à six chiffres, du moins pour le scénario analysé ici[4].

Tableau 2
Corrélations : Mesures du statut et du prestige

	ISEI-08		SIOPS-08		% valides	
$N_{Apprentissage}$	BN	BMN	BN	BMN	BN	BMN
25 000	0,904	0,968	0,922	0,973	78	54
50 000	0,907	0,966	0,928	0,973	80	58
100 000	0,917	0,967	0,937	0,973	82	61
300 000	0,929	0,964	0,945	0,971	85	68

Données de la NEPS; $N_{Test} = 7\,500$; BN = bayésien naïf; BMN = bayésien multinomial

Pour de nombreuses questions de recherche appliquée, le code de profession correct n'est pas une priorité absolue. Il suffirait souvent d'obtenir un compte rendu raisonnablement précis du statut socioéconomique ou du prestige de la profession d'une personne. Les échelles utilisées à ces fins sont souvent dérivées du code de profession (p. ex., l'ISEI ou la SIOPS dérivée de la CITP). Même si les professions peuvent différer à de nombreux égards, des professions similaires fournissent également souvent des niveaux semblables de prestige ou de statut socioéconomique. Par conséquent, même les cas pour lesquels le codage automatique a fourni une classification incorrecte pourraient produire des mesures de statut et de prestige correctes à condition que le code de profession qui a été attribué mène au même score d'ISEI ou SIOPS.

Nous avons évalué la qualité des mesures selon l'ISEI et la SIOPS pour les codages automatiques effectués par les deux algorithmes en établissant la corrélation entre les scores obtenus à ceux dérivés d'après le codage manuel[5]. Les algorithmes BN ainsi que BMN ont donné d'assez bons résultats, les corrélations variant de 0,904 pour l'algorithme BN pour une taille d'échantillon de 25 000 à 0,971 pour l'algorithme BMN pour 300 000 cas dans les données d'apprentissage (voir le tableau 2).

Seuls les cas pour lesquels existait un code valide dans les données codées manuellement ou automatiquement ont été utilisés pour l'analyse. Cela s'est traduit par un nombre considérablement plus faible de cas pour l'algorithme

[4] Les résultats pour différentes tailles d'échantillon des données de test devant être codées ont été estimés, mais comme les variations étaient peu importantes, ils ne sont pas présentés ici.

[5] Pour ces analyses, nous avons commencé par recoder les codes de la KldB2010 attribués par le codage automatique ainsi que le codage manuel en codes de l'ISCO08. Cette conversion se fait raisonnablement bien étant donné que la KldB2010 a été élaborée en tenant compte de l'ISCO08. Ensuite, nous avons utilisé les codes de l'ISCO08 pour obtenir les scores ISEI et SIOPS. Un codage selon l'ISCO08 original aurait été préférable, mais il n'était pas disponible pour les données.

BMN que pour l'algorithme BN, et pourrait expliquer certains avantages du premier par rapport au second. Les cas non codés par l'algorithme BMN sont fort probablement ceux pour lesquels l'incertitude est grande. Ignorer ces cas non codés signifie moins d'erreurs de classification parmi les cas codés et donc de meilleures corrélations pour les mesures dérivées de statut et de prestige.

L'amélioration de la qualité lorsque la taille de l'échantillon de données d'apprentissage augmente est nettement moins prononcée qu'elle ne l'était pour les taux de concordance. Pour l'algorithme BMN, elle est même inexistante.

5. Conclusion et recherche future

Ces résultats préliminaires portent à conclure que le codage automatique des réponses aux questions d'enquête ouvertes sur la profession semble faisable, à condition de disposer d'un échantillon suffisamment grand de données d'apprentissage de haute qualité. La performance des deux algorithmes est acceptable, quoique l'algorithme BMN possède un avantage par rapport à l'algorithme BN quand on vise des taux de concordance fixes élevés (> = 90 %). Dans les cas où seules les échelles de prestige et de statut dérivées sont nécessaires, même de plus petits ensembles de données pourraient suffire pour fournir des estimations acceptables.

Nous avons qualifié nos conclusions de « préliminaires » pour plusieurs raisons. Premièrement, nous devons effectuer davantage d'analyses en nous servant des données et des algorithmes dont il est question ici afin de valider nos résultats et de confirmer leur robustesse. Par conséquent, l'une des prochaines tâches consistera à effectuer des vérifications transversales.

Deuxièmement, les algorithmes pourraient encore être améliorés. En plus de tester d'autres algorithmes d'apprentissage automatique (p. ex., forêts d'arbres décisionnels (*Random Forests*) ou machines à vecteurs de support (*Support Vector Machines*)), nous essaierons d'optimiser davantage les algorithmes BN et BMN. Une idée consiste à intégrer des mesures de distance (p. ex., distance de Levenshtein) dans les algorithmes pour mieux utiliser l'information contenue dans les données d'apprentissage. Les premiers essais ont donné des résultats prometteurs. En outre, nous essaierons d'élaborer une forme raisonnable de prétraitement des données afin d'obtenir des chaînes de texte plus épurées, ce qui à son tour devrait réduire la quantité de bruit dans les données et mener à des estimations plus précises des probabilités.

Enfin, l'un des principaux objectifs à long terme du projet est d'établir des pratiques exemplaires pour le codage automatique des professions qui pourraient alors être appliquées à plusieurs enquêtes à grande échelle en Allemagne.

Bibliographie

BLOSSFELD, H.-P., H.-G. ROßBACH et J. VON MAURICE, sous la dir. de (2011), « Education as a Lifelong Process. The German National Educational Panel Study (NEPS) », *Zeitschrift für Erziehungswissenschaft – Numéro spécial* 14. VS Verlag für Sozialwissenschaften.

BUNDESAGENTUR FÜR ARBEIT, sous la dir. de (2011), Klassifikation der Berufe 2010. Nürnberg.

MUNZ, MANUEL, WENGIZ K. et BELA D., (à paraître), « String coding in a generic framework ».

SCHIERHOLZ M. (2014), « Automating Survey Coding for Occupation ». FDZ-Methodenreport, 10/2014. Nürnberg.

THOMPSON M., KORNBAU M. E. et VESELY J. (2012), « Creating an Automated Industry and Occupation Coding Process for the American Community Survey », rapport inédit.