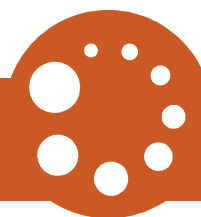




Plausi++: Contrôles automatiques de plausibilité pour vérifier la qualité et la fiabilité des données administratives et des données collectées par voie d'enquête

Résumé

EXPERIMENTAL STATISTICS



Neuchâtel, 2019

Éditeur:	Office fédéral de la statistique (OFS)	Concept de mise en page:	Section DIAM
Rédaction:	Dr. Christian Ruiz	Téléchargement:	www.statistique.ch
Contenu:	Dr. Christine Ammann Tschopp, Elisabeth Kuhn, Laurent Inversin, Dr. Mehmet Aksözen, Stefan Rüber	Copyright:	OFS, Neuchâtel 2019 La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée
Domaine:	00 Bases statistiques des généralités		
Langue du texte original:	Allemand		
Traduction:	Services linguistiques de l'OFS		



Projet pilote Plausi++ : résumé

Le 30 juin 2019

Des contrôles de plausibilité (ci-après «plausis») sont réalisés pour vérifier la qualité et la fiabilité des données administratives et des données d'enquête. Les données qui semblent clairement incorrectes ou du moins peu plausibles sont retournées aux fournisseurs, qui sont priés de les renvoyer corrigées ou avec un commentaire explicatif. Jusqu'ici, ces contrôles de plausibilité étaient effectués à l'aide de règles automatiques et de seuils empiriques, ou manuellement. Dans le projet pilote «Plausi++», nous avons créé un nouveau type de contrôles de plausibilité basé sur des algorithmes d'apprentissage automatique (ML pour « Machine Learning ») qui détectent automatiquement les combinaisons de données possiblement erronées. En l'associant à un mécanisme de feed-back automatisé, il nous est possible d'envoyer au fournisseur de données une explication compréhensible pour l'utilisateur. Cela nous a permis de compléter les contrôles de plausibilité existants, de réduire le temps requis et d'améliorer la qualité des données.

Résultats

Objectif 1 de l'algorithme ML («partie prédictive») : l'utilisation de l'algorithme ML a permis d'améliorer sensiblement la qualité des données. La comparaison avec la situation actuelle (« baseline ») a montré que Plausi++ avait détecté **29% d'erreurs de plus** que les règles de plausibilisation classiques. L'algorithme ML fonctionne avec une **précision de prédiction de plus de 94%**.

Les erreurs détectées par Plausi++ **portent uniquement sur le contenu**. Plausi++ **vient ainsi compléter les règles classiques**, dont seules 21% concernent le contenu et 79% portent sur des aspects formels (par ex. respect des codes, absence d'informations non renseignées). **Plausi++ ne remplace donc pas les plausis classiques, mais s'y associe pour couvrir aussi bien le contenu que la forme.**

Autre indice de l'amélioration de la qualité des données : la **détection d'erreurs systématiques à un niveau agrégé**. Dans l'ensemble, **Plausi++ détecte plus d'erreurs systématiques**. Par exemple, **deux fois plus d'erreurs** de ce type ont été trouvées lors d'une utilisation test qu'avec les plausis classiques.

L'apprentissage automatique supervisé («supervised ML») comprend deux phases. Si dans la première s'entraîne longuement l'acquisition des connaissances, dans la deuxième phase, la phase d'utilisation, **Plausi++ n'a besoin que de quelques minutes** pour être exécuté. Cela devrait permettre de réduire les échanges avec les fournisseurs de données et donc de gagner du temps, ce que l'on ne pourra vérifier qu'une fois l'implémentation dans l'environnement de production réalisée.

Objectif 2 Mécanisme de feed-back («partie explicative») : un mécanisme de feed-back automatisé explique le résultat de l'algorithme de ML au fournisseur de données d'une manière compréhensible pour un être humain. Une procédure a été mise en place qui permet d'implémenter entièrement une telle réponse et de mettre en évidence les variables pouvant poser problème (par ex. l'âge ou la catégorie de personnel). Il n'a pas encore été possible de tester d'autres approches.

Objectif 3 Documentation : une documentation détaillée a été rédigée de manière à ce que cette méthode puisse être adaptée et reproduite selon les besoins d'autres domaines et sections.



Objectif 4 Intégration de la production : le projet pilote a évalué la faisabilité d'une intégration dans le système de production. Naturellement, beaucoup d'inconnues subsistent encore telles que le choix de l'infrastructure informatique finale ainsi que d'autres éléments qui ne font pas partie du projet pilote. Selon l'état de nos réflexions, nous pensons que Plausi++ devrait pouvoir bien s'intégrer dans le processus de production des plausis classiques.

Conclusion

Les quatre objectifs du projet pilote sont atteints, plusieurs tests ont été effectués suite auxquels Plausi++ peut désormais être intégré à la production. S'il a été relativement facile d'intégrer un algorithme de ML au sens strict, les mécanismes de feed-back ont engendrés plus de difficultés. Une documentation détaillée a été établie qui pourra être adaptée aux besoins spécifiques.

Méthodes d'analyse complémentaires utilisées :

- Nous avons utilisé pour la partie prédictive des algorithmes tels que «Gradient Boosting Machines» et «Random Forests». Le premier a permis d'obtenir une plus grande précision de prédiction.
- Pour la partie explicative, nous avons utilisé les algorithmes «Ceteris Paribus» et «Local Interpretable Model-Agnostic Explanations» (LIME). Le mécanisme de feed-back fonctionne pour l'heure avec *Ceteris Paribus*