Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup

A. Yeh, L. Hirschman, A. Morgan, *BioInformatics*, **19**, 2003

Marie Guégan - Fiche de lecture, *Fouille de données et apprentissage*, M2R Info, mars 2006.

Résumé : La tâche 1 de la compétition Knowledge Discovery and Data Mining (KDD) Challenge Cup 2002 concerne l'extraction de données biologiques dans des textes pour leur insertion dans une base de données. Cet article expose l'élaboration, les résultats et les leçons à tirer de cette campagne.

1 Introduction

Les données biologiques nécessitent un transfert manuel par des experts depuis la littérature surabondante vers les bases de données afin d'être exploitables automatiquement. Les systèmes de fouille de textes élaborés pour faciliter la tâche n'ont jamais encore été comparés. Les auteurs ont donc lancé une campagne d'évaluation pour fournir des mesures d'évaluation de ces systèmes.

2 Elaboration du concours

Une tâche d'aide automatique a été définie après consultation d'experts, qui ont fourni les données d'apprentissage et de test (FlyBase) : en bref, identifier les articles "méritant" une exploitation manuelle. Des difficultés telles que le recueil d'articles complets, leur formatage (légendes, symboles) et la diversité des annotations par FlyBase ont conduit à simplifier fortement la tâche, qui reste ambitieuse. La mesure d'évaluation des systèmes candidats utilise la courbe ROC et la F-mesure.

3 Résultats

L'équipe gagnante sur les 18 participants s'est basée sur des règles construites manuellement pour rechercher des motifs linguistiques. D'autres y ont combiné un classifieur bayésien ou des SVMs. Tous ont écarté l'approche "sac de mots" qui casse les relations (interaction entre gènes...). Beaucoup ont eu recours à des experts biologistes. En moyenne, les participants ont été très prudents, préférant la précision au rappel.

4 Discussion : les leçons

Divers problèmes sont apparus à la constitution des données : les versions complètes des articles sont nécessaires mais rarement en libre accès ; les systèmes ne pouvant traiter le PDF, les articles ont été choisis en HTML ; le formatage supprime les figures et parfois aussi les légendes très porteuses d'information...

Les auteurs voulaient avant tout évaluer la campagne, qu'ils jugent réussie. Leurs critères sont : une répétition possible de l'évaluation, un nombre suffisant de participants, une tâche réaliste mais ambitieuse.

5 Conclusion

Cette campagne est un succès. Les résultats sont prometteurs mais méritent une analyse plus approfondie. Les auteurs envisagent d'autres campagnes. Ils souhaitent plus de participation des biologistes.

Avis personnel

Cet article est bien écrit, malgré quelques parties redondantes. Les auteurs se sont centrés sur l'élaboration de la campagne : la description en est très complète. On peut regretter que l'analyse des approches des participants ne soit pas plus poussée, comme dans [1]. Si l'on suit la catégorisation des articles du type "leçons à tirer" de [Lavrac et al., 2004][2], cet article serait plutôt classé dans "les études de cas bien détaillées".

[1] Charles Elkan (2001). Magical thinking in data mining : Lessons from CoIL Challenge 2000. *In Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 426-431

[2] Nada Lavrac, Hiroshi Motoda, Tom Fawcett (2004). Editorial : Data Mining Lessons Learned, *Machine Learning*, 57, 5-11