UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

Measuring distances between medical entities. Step 1: DrugBank

Alberto Olivares-Alarcos — Iva Stankovic — Humberto González and Horacio Rodríguez

Department of Computing Science, Universitat Politècnica de Catalunya, UPC

Abstract

We face in this paper the problem of computing distance measures between medical entities. Specifically we deal with the most challenging type of medical entity: drugs. Three different similarity measures between drugs are presented, based each one on specific dimensions of drugs description, namely textual, taxonomic and molecular information. All the information has been extracted from the same resource, the DrugBank database

Evaluation

Clustering

Similarities have been used to cluster the drugs into groups. Then, we studied the **ATC Code distribution** of those clusters in order to check if our similarity measurements are good

Ground Truth

The similarity of a list of **100 pairs of drugs** was annotated by experts. We have taken it from [Franco et al., 2014] and modified and adapted to our convenience

[Franco et al., 2014] Franco, P., Porta, N., Holliday, J. D., and Willett, P. (2014). The use of 2d fingerprint methods to support the assessment of structural similarity in orphan drug legislation. Journal of Cheminformatics, 6(1):5.

Overview

Conclusion

Contributions

- **Novel** text based similarity over DrugBank
- **Novel** taxonomy based similarity over DrugBank
- Molecular Structure based similarity over DrugBank
- **Three similarity** measurements within the **same framework**
- Qualitative indirect evaluation based on clustering
- Quantitative direct evaluation using a small ground truth
- MIT License Code provided on a GitHub repository (**Python**)

Conclusions

Three different similarity measurement over drugs from DrugBank have been implemented:textual, taxonomic and molecular. To our knowledge theres is no other work which includes these three measures within the same framework. A evaluation of the implemented similarities has been performed by means of both indirect (Clustering) and direct (Ground Truth) evaluation.

The **Clustering evaluation** has provided lights and shadows, while in some cases we have been able to cluster properly the drugs based on their ATC Codes, we have not in several cases. This does not strongly implies our similarity measures are not good. Spectral Clustering, used in this work, and graph-based semi-supervised learning algorithms, in general, are well known to be sensitive to how graphs are constructed from data. In particular if the data has proximal and **unbalanced** clusters these algorithms can lead to poor performance.

On the other hand, some **promising results** have been found in the evaluation based on the ground truth, specially, for the similarity based on Molecular Structure. Nevertheless, the results are not definitive, a need of a **larger ground** truth is clear.