Représentativité, généricité et singularité : augmentation de données pour l'exploration de dossiers médicaux

Joris Falip, Amine Aït Younes, Frédéric Blanchard, Michel Herbin

Représentativité, généricité et singularité : augmentation de données pour l'exploration de dossiers médicaux

Joris Falip*, Amine Aït Younes*, Frédéric Blanchard* Michel Herbin*

*CReSTIC, Université de Reims Champagne Ardenne
joris.falip@univ-reims.fr,
http://crestic.univ-reims.fr

1 Introduction

Les données médicales, issues entre autres de dossiers médicaux électroniques, sont de plus en plus abondantes et proposent des défis qui leur sont propres. Pour exploiter ces données, les experts ont besoin d'outils d'exploration et de visualisation favorisant l'émergence de nouvelles hypothèses cliniques. Il est ainsi important, dans le domaine de la santé, de tenir compte des cas atypiques et inhabituels tout en évitant la généralisation (**?**). Chaque cas est en effet un cas particulier, et tenter de générer des individus typiques issus de moyennes statistiques nous éloigne de la réalité en proposant aux professionnels de santé des modèles abstraits ne correspondant finalement vraiment à aucun des cas concrets qu'ils cherchent à explorer. Le projet *CoSyRES* vise à proposer des algorithmes adaptés à ces contraintes (**?**) pour la visualisation et l'exploration de ces données. Nous avons choisi un paradigme basé-instance, très proche de celui utilisé par les professionnels de santé se basant sur leur expérience. Dans cet article, nous proposons une méthode exploratoire permettant de faire émerger des associations de patients similaires et d'observer des regroupements de patients autour des cas jugés comme typiques. L'approche mise en avant ici est adaptée aux difficultés inhérentes aux bases de données médicales : informations absentes ou erronées, données hétérogènes et haute dimensionnalité (**?**).

2 Représentativité et émergence d'associations

La méthode exposée ici a pour but de faire émerger des associations entre les individus étudiés, permettant ainsi d'effectuer des regroupements autour d'individus emblématiques. Pour cela, chaque individu va voter en attribuant, selon chaque dimension, un score aux individus qui lui sont semblables. En prenant l'exemple d'un ensemble de données composé de **n** individus décrits chacun par **f** variables, l'émergence de cette structure a lieu en cinq étapes successives :

1. Etablissement, sur chacune des **f** dimensions, d'une matrice de dissimilarité entre les individus.
2. Chaque individu, pour chaque dimension, classe ses voisins par proximité. On obtient donc **f** matrices de rangs, comprenant chacune **n** classements.

3. Transformation des rangs en scores, chaque individu se voyant attribuer **f** scores par chacun des **n** autres individus. Ces scores peuvent être vus comme l'expression de "préférences" individuelles : un score élevé est attribué en cas de proximité forte, un score faible ou nul se voit attribué si les deux individus sont éloignés.
4. Emergence des préférences via l'agrégation des scores obtenus à l'étape précédente. Le score final d'un individu est obtenu en agrégeant les scores qui lui ont été attribués. La valeur obtenue est proportionnelle à la *représentativité* de l'individu, sa capacité à représenter un sous-groupe de la population étudiée.
5. Choix des représentants. En choisissant un facteur **k** modélisant le niveau de granularité souhaité, chaque individu va choisir parmi ses **k** plus proches voisins celui qui a le plus haut score comme étant l'instance la plus représentative de ses caractéristiques.

Des expériences menées sur des jeux de données synthétiques et contrôlés confirment la cohérence des résultats fournis par l'algorithme présenté.

3 Conclusion

L'algorithme proposé permet la visualisation et l'exploration de données de santé à l'aide d'un graphe orienté et valué. Cette approche orientée cas permet de conserver le côté personnel et individualisé de la donnée : un facteur très important dans les applications médicales. La structure de graphe et les notions de représentativité, généricité et singularité permettent d'augmenter les données. La visualisation est ainsi facilitée et l'exploration guidée. L'objectif est de faire émerger de nouvelles hypothèses cliniques, à partir des données patients.

Références

Blanchard, F., A. Aït Younes, et M. Herbin (2015). Linking data according to their degree of representativeness (dor). *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems 15*(4).

Blanchard, F., P. Vautrot, H. Akdag, et M. Herbin (2010). Data representativeness based on fuzzy set theory. *Journal of Uncertain Systems 4*(3), 216–228.

Nourizadeh, A., F. Blanchard, A. Aït Younes, B. Delemer, et M. Herbin (2013). Exploratory data analysis of insulin therapy in the elderly type 2 diabetic patients. *Studia Informatica Universalis 11*(3), 32–49.

Summary

In order to classify individuals according to exemplars that represent them accurately, data visualisation applied to the medical field need to avoid overgeneralization : each case must be treated as a particular instance. This paper presents an algorithm allowing each individual in a dataset to rank other individuals and vote for those that match their important features. Aggregating all these votes give us a way to visualize data according to typical individuals representing subsets of closely-related patients.