Classification sur données médicales à l'aide de méthodes d'optimisation et de datamining, appliquée au pré-screening dans les essais cliniques

Julie Jacques

Université Lille 1
École Doctorale Sciences Pour l'Ingénieur Lille 1 Nord-de-France
Laboratoire d'Informatique Fondamentale de Lille (UMR CNRS 8022)
Centre de Recherche Inria Lille Nord Europe

THÈSE

pour obtenir le grade de
Docteur de l'Université de Lille I
Discipline : Informatique
Numéro d'ordre : 41262 | Année : 2013

Classification sur données médicales à l'aide de méthodes d'optimisation et de datamining, appliquée au pré-screening dans les essais cliniques

par

Julie JACQUES

Date de soutenance : 2 décembre 2013

Jury

Directeurs :	Clarisse DHAENENS	Professeur des Universités, Université Lille I
	Lætitia JOURDAN	Professeur des Universités, Université Lille I
Rapporteurs :	Jean-Charles BILLAUT	Professeur des Universités, Université de Tours
	Nadia BRAUNER	Professeur des Universités, Université Grenoble I
Examinateurs :	Stéphane BONNEVAY	Maître de conférences, HDR, Université Lyon I
	Denis BOUYSSOU	Directeur de recherche, Université Paris Dauphine
	Sophie TISON	Professeur des Universités, Université Lille I
Invité :	David DELERUE	Gérant, Société Alicante

Abstract Medical data suffer from uncertainty and a lack of standardization, making them hard to use in medical software, especially for patient screening in clinical trials. In this PhD work, we propose to deal with these problems using supervised classification methods. We will focus on 3 properties of these data : imbalance, uncertainty and volumetry. We propose the MOCA-I algorithm to cope with this partial classification combinatorial problem, that uses a multi-objective local search algorithm. After having confirmed the benefits of multiobjectivization in this context, we calibrate MOCA-I and compare it to the best algorithms of the literature, on both real data sets and imbalanced data sets from literature. MOCA-I generates rule sets that are statistically better than models obtained by the best algorithms of the literature. Moreover, the models generated by MOCA-I are between 2 to 6 times shorter. Regarding balanced data, we propose the MOCA algorithm, statistically equivalent to best algorithms of literature. Then, we analyze both theoretically and experimentally the behaviors of MOCA and MOCA-I depending on imbalance. In order to help the decision maker to choose a solution and reduce over-fitting, we propose and evaluate different methods to handle all the Pareto solutions generated by MOCA-I. Finally, we show how this work can be integrated into a software application.

Keywords Imbalanced data, partial classification, combinatorial optimization, multi-objective optimization, machine learning, datamining.

Résumé Les données médicales souffrent de problèmes d'uniformisation ou d'incertitude, ce qui les rend difficilement utilisables directement par des logiciels médicaux, en particulier dans le cas du recrutement pour les essais cliniques. Dans cette thèse, nous proposons une approche permettant de pallier la mauvaise qualité de ces données à l'aide de méthodes de classification supervisée. Nous nous intéresserons en particulier à 3 caractéristiques de ces données : asymétrie, incertitude et volumétrie. Nous proposons l'algorithme MOCA-I qui aborde ce problème combinatoire de classification partielle sur données asymétriques sous la forme d'un problème de recherche locale multi-objectif. Après avoir confirmé les apports de la modélisation multi-objectif dans ce contexte, nous calibrons MOCA-I et le comparons aux meilleurs algorithmes de classification de la littérature, sur des jeux de données réels et asymétriques de la littérature. Les ensembles de règles obtenus par MOCA-I sont statistiquement plus performants que ceux de la littérature, et 2 à 6 fois plus compacts. Pour les données ne présentant pas d'asymétrie, nous proposons l'algorithme MOCA, statistiquement équivalent à ceux de la littérature. Nous analysons ensuite l'impact de l'asymétrie sur le comportement de MOCA et MOCA-I, de manière théorique et expérimentale. Puis, nous proposons et évaluons différentes méthodes pour traiter les nombreuses solutions Pareto générées par MOCA-I, afin d'assister l'utilisateur dans le choix de la solution finale et réduire le phénomène de sur-apprentissage. Enfin, nous montrons comment le travail réalisé peut s'intégrer dans une solution logicielle.

Mots clefs Classification sur données asymétriques, classification partielle, optimisation combinatoire, optimisation multi-objectif, apprentissage, fouille de données.

Remerciements

Je tiens tout d'abord à remercier Clarisse et Laetitia, mes directrices de thèse, avec qui j'ai eu la chance et le plaisir de travailler pendant ces 3 ans. Je les remercie toutes les deux pour leurs nombreux conseils, relectures, encouragements et leur disponibilité.

Je souhaite remercier les membres du jury pour l'intérêt qu'ils ont porté à mon travail et les remarques pertinentes qu'ils y ont apporté : Jean-Charles Billaut et Nadia Brauner, rapporteurs, Stéphane Bonnevay et Denis Bouyssou, examinateurs, et Sophie Tison, présidente du jury.

Je tiens à remercier les Professeurs Régis Bordet et Régis Beuscart d'avoir partagé leur expérience sur les essais cliniques pour la conception d'Opcyclin, ainsi que Virginie Deprez et Anne-Marie Bordet d'avoir partagé leur expérience d'attaché de recherche clinique.

Je remercie également mes collègues d'Alicante David, Julien, Muriel, Charles et Benjamin avec qui j'ai travaillé sur Opcyclin, ainsi que Cédric, Jacques, Fabienne, Samuel et Eve que j'ai cotoyés pendant ces 3 ans à Alicante.

Je tiens aussi à remercier les différents membres permanents de l'équipe Dolphin pour leur accueil : El-Ghazali Talbi, Nouredine Melab, Luce Brotcorne, Bilel Derbel, Dimo, François, Sébastien, Arnaud et Marie. Sans oublier les non-permanents : Thé Van, Sezin, Ines, Nadia, Moustapha, Bayrem, Tuan, Aline, Ekaterina, Sophie, Martin, Mathieu et en particulier Yacine et Julie avec qui j'ai pu partager la thèse au quotidien avec ses hauts et ses bas, et dont les nombreuses sessions "thé" ont boosté ma productivité.

Je remercie mon compagnon, André, de m'avoir soutenue et encouragée pendant ces 3 ans. Je remercie également la famille Demoli pour son soutien et ses encouragements. Je tiens à remercier Floriane, ma petite soeur, pour son soutien, et ses conseils en anglais. Je remercie également mes parents, en particulier mon père pour m'avoir donné le goût des sciences et de l'innovation, et toujours contribué à développer ma curiosité scientifique, même s'il n'est plus là aujourd'hui.

Je remercie également mes amis pour leur soutien, et pour m'avoir rechargé les batteries quand j'en avais besoin : les amis d'Outre-Quiévrain : Florence, Michael et Astrid mais aussi les français : Yoann, Sandra, Virginie et Geoffrey.

Enfin, ce travail n'aurait pu avoir lieu sans le financement de la société Alicante. J'ai beaucoup apprécié de contribuer à certains des projets innovants d'Alicante, pour lesquels j'ai bénéficié de beaucoup de liberté et de confiance quant à la manière de les mener à bien.

Confidentialité

Conformément à la demande de la société Alicante, qui a financé ces travaux de recherche, le présent manuscrit est confidentiel pour une durée de 3 ans. Certaines publications issues de ce travail sont néanmoins consultables :

- Julie Jacques, Julien Taillard, David Delerue, Laetitia Jourdan Clarisse Dhaenens
  MOCA-I : discovering rules and guiding decision maker in the context of partial classification in large and imbalanced datasets
  Learning and Intelligent OptimizatioN Conference, LION 7, Lecture Notes in Computer Science 2013, Pages 37-51
- Julie Jacques,Julien Taillard, David Delerue, Laetitia Jourdan Clarisse Dhaenens
  The Benefits of Using Multi-objectivization for Mining Pittsburgh Partial Classification Rules in Imbalanced and Discrete Data
  Proceeding of the fifteenth annual conference on Genetic and evolutionary computation conference, Pages 543-550.