EPIC10 : Mesure, analyse et modélisation des données biologiques

Résumés des séminaires recherche

Le jeudi 14 octobre 2010 à Polytech'Lille (Université Lille 1)

Extraction et mesure de forme en imagerie médicale : les sinus de Valsalva

T. Sliwa (Le2i - UMR 5158, Univ. Bourgogne)

Sur la base de travaux menés par Cédric Blanchard (doctorant), Tadeusz SLIWA (encadrant), Yvon Voisin (directeur de thèse), Alain Lalande (CHU) et Olivier Bouchot (CHU).

L'imagerie médicale conduit à la production d'importantes quantités d'images souvent incomplètement exploitées. Se pose alors la question de l'automatisation d'une aide à l'inspection visuelle, par l'extraction automatique de zones d'intérêt susceptibles d'attention de la part du praticien, ou d'une aide à la décision, par des outils d'analyses plus objectifs que l'utilisation de protocoles manuels ambigus et de données de comparaison non fiables. Cela se traduit souvent par la mise au point d'outils d'extraction et d'analyses de formes adaptés à un besoin précis. Les difficultés sont liées, d'une part au fait que des problèmes fondamentaux comme celui de la segmentation d'image demeurent des problèmes ouverts et difficiles, et d'autre part au fait que le monde médical nécessite une fiabilisation maximale des informations à extraire.

A titre d'exemple, le cas de l'évaluation des sinus de Valsava par IRM est abordé ici. Bien que l'échographie reste la technique de référence, l'IRM est une méthode alternative ou complémentaire pour l'étude des valves aortiques. Actuellement, il n'y a pas de protocole de référence concernant la mesure de ces éléments et la classification des cas pathologiques est parfois ambigüe. Sans compter que les données statistiques disponibles pour l'aide à la décision sont très dépendantes de protocoles dont la fiabilité est actuellement discutée. Bien que l'imagerie faisant partie prenante de la décision dans la cadre d'un prise en charge chirurgicale, l'élaboration éventuelle d'une prothèse sur laquelle doivent s'insérer des valves se fait souvent de manière artisanale et non adaptée à la morphologie exacte du patient, aboutissant en général au choix d'un gabarit de prothèse composée généralement d'éléments relativement simples (ex : cylindre), et pouvant donner lieu à terme à divers problèmes secondaires. L'utilisation de séquence ciné-IRM en apnée entraîne des contraintes d'acquisition limitant la résolution spatiale de l'image. Comparé à la taille de l'organe, on peut considérer que la résolution spatiale est relativement faible.

Un logiciel est mis au point permettant d'extraire et d'analyser automatiquement les sinus de Valsava sur des séquences en ciné-IRM, que cela soit pour la constitution de données statistiques, la remise en cause éventuelle des classifications actuelles, la définition de protocoles objectifs et répétables ou bien même intégrant peut-être à plus long terme la fabrication automatisée de prothèses patient-adaptés. Tout ceci de préférence en temps-réel (ou du moins très rapidement), de manière à pouvoir fournir les résultats dès la fin de l'examen et donc réagir rapidement dans les cas d'urgence, ce qui n'est pas évident étant donné la lourdeur des traitements d'images parfois nécessaires. Comme pour de nombreuses applications de traitement d'images, de nombreux outils mathématiques (méthodes numériques, géométriques, statistiques), informatiques (algorithmique, adéquation algorithme-architecture, interfaçage) voire électroniques (temps-réel) interviennent. Par ailleurs, comme c'est souvent le cas aussi, certains outils développés visent à améliorer la résolution de problèmes plus généraux.

Tous ces éléments seront présentés ici, comportant notamment une introduction au domaine vaste et ouvert de l'extraction et de l'analyse des formes, sous-domaine de la vision artificielle (lui-même sous-domaine de l'intelligence artificielle), domaine à la fois déjà très avancée et très embryonnaire et au niveau duquel de nombreuses "écoles" sont sans cesse en compétition.

Mots-clés : imagerie médicale, extraction de primitives, morphologie, classification, mesure, architecture.

Datamining sur données génomiques : approches par méthodes d'optimisation

C. Dhaenens (LIFL / INRIA & Polytech'Lille, Univ. Lille 1)

La recherche génomique est un véritable enjeu pour notre société et de nombreux laboratoires de spécialités différentes (recherche en biologie, en médecine ou en technologie de l'information) se regroupent pour participer à des recherches sur des thèmes précis. En ce qui nous concerne, nous étudions, par exemple, les facteurs génétiques de prédisposition à certaines maladies multifactorielles telles que le diabète, l'obésité et les maladies cardiovasculaires.

L'originalité commune des différentes problématiques est de rechercher non pas un seul facteur explicatif, mais bien une ou plusieurs combinaisons de facteurs (pouvant être de différentes natures : facteurs génétiques, facteurs environnementaux...) parmi un ensemble très grand de facteurs potentiels (plusieurs milliers). Nous sommes donc face à un problème d'optimisation combinatoire.

Nous nous plaçons en particulier dans le cadre de la génomique et de la post-génomique, caractérisées par un volume de données brutes en augmentation très rapide (grâce aux nouvelles technologies de récolte de ces données). En effet, la difficulté réside aujourd'hui non plus seulement dans l'obtention de ces données, mais également dans leur analyse. Ainsi, un de nos objectifs consiste à développer des méthodes d'analyse permettant d'extraire un maximum d'informations à partir des données récoltées par les biologistes et généticiens. La première étape concerne donc la modélisation des problématiques exprimées par les biologistes. En ce qui nous concerne, nous les modélisons en des problèmes d'extraction de connaissances que nous transformons ensuite en des problèmes d'optimisation combinatoire. Ceci nécessite donc de définir à la fois l'ensemble des solutions admissibles, et donc déterminer quel type de réponse est attendue (classification des instances, relations entre les attributs,...) et le(s) critère(s) à optimiser. Il est ensuite possible d'appliquer à ce problème transformé des méthodes d'optimisation.

Au cours de cet exposé la méthodologie générale de l'approche sera présentée. Puis dans un deuxième temps nous nous focaliserons en particulier sur une problématique que nous avons modélisée comme une recherche de règles d'association (problématique classique de datamining). Une caractéristique ici est que l'on recherche des associations de plusieurs facteurs parmi un grand nombre de facteurs potentiels. Ainsi, la combinatoire (le nombre de solutions possibles) est très grande. De plus, l'analyse du problème et des données nous a montré que nous devions développer des méthodes ayant un fort pouvoir d'exploration. Enfin, l'étude des critères d'évaluation des règles d'association nous a montré qu'il n'existe pas de critère universel et nous a conduit à modéliser la recherche de règles d'association en un problème multi-critère. Pour toutes ces raisons nous avons alors choisi d'utiliser les algorithmes évolutionnaires comme base de résolution. Actuellement des études sont menées pour améliorer ces méthodes en les hybridant avec des méthodes exactes permettant de résoudre des sous-problèmes (problèmes de tailles beaucoup plus petites).

Nouveaux défis en classification de données biologiques

C. Biernacki (Labo. Paul Painlevé, Univ. Lille 1)

La classification, supervisée ou non, est une problématique récurrente en biologie à laquelle de nouveaux défis sont régulièrement lancés. L'objectif est ici d'en donner des exemples ainsi que de présenter des solutions issues de recherches récentes.

Nous nous intéressons dans un premier temps au contexte supervisé. Les données sans étiquettes sont souvent beaucoup plus nombreuses que celles étiquetées (du fait de leur faible coût) et un enjeu important est de pouvoir les utiliser. On parle alors plutôt de classification semi-supervisée. Dans ce cadre, les modèles probabilistes génératifs s'imposent par leur capacité à exploiter ce nouveau type de données. Nous présentons en particulier une méthode de choix de modèles qui évite la lourdeur des méthodes de ré-échantillonnage imposée par l'utilisation des données sans étiquettes. Nous discutons également de situations réalistes où les données avec et sans étiquettes ne proviennent pas strictement de la même population, nécessitant un réajustement préliminaire.

Puis nous nous plaçons dans le contexte non supervisé, cas où la question du nombre de groupes est cruciale. Nous nous intéressons en particulier aux faibles tailles d'échantillons, situation réaliste lorsque les données biologiques sont acquises difficilement, rendant ainsi d'autant plus délicat l'identification des groupes et de leur nombre. Cette situation rend les critères de choix de modèles asymptotiques classiques peu performants et nous proposons des versions non asymptotiques, détectant ainsi les structures sous-jacentes dès de faibles tailles d'échantillons. Alternativement, nous discutons de la possibilité d'associer des petits échantillons provenant de populations différentes pour améliorer la qualité globale de l'estimation. L'ensemble de l'étude se fait de nouveau avec des modèles génératifs pour les mêmes raisons que précédemment.