Organisation et optimisation des données pour l'apprentissage de structure d'un réseau bayésien multi-entités

H. Bouhamed [1], A. Rebai [2], T. Lecroq [1], M. Jaoua [3]

[1] Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes (LITIS), Université de Rouen, France
Heni.bouhamed@etu.univ-rouen.fr

[2] Centre of Biotechnology Sfax (CBS), Tunisia

[3] Multimedia Information Systems and Advenced Computing laboratory (MIRACL), Université de Sfax, Tunisie

Résumé

L'objectif de notre travail est de développer une nouvelle démarche d'extraction de connaissances à partir d'une grande masse de données, le résultat de l'application de cette démarche sera un système expert qui servira comme outils de diagnostic d'un phénomène liés à un énorme système d'information. Nous rappelons tout d'abord la problématique générale de l'apprentissage de structure d'un réseau bayésien à partir de données et nous proposons une solution d'optimisation de la complexité en utilisant des méthodes d'organisations et d'optimisations des données avant l'apprentissage et un formalisme spécifique des réseaux bayésiens qui est les MEBN (Multi entities bayesian networks ou réseau bayésien multi-entités). Enfin nous concluons sur les limites atteintes par ces travaux.

Mots clefs

Réseaux bayésiens multi entités, apprentissage de structure à partir de données, Fusion de scores, classification, analyse simple variable, sélection de classes.

1 Introduction

Dans les études qui visent à identifier les classes de variables responsables d'un phénomène donné on utilise souvent un grand nombre de variables chacun donnant un signal (mesuré par un test statistique) en faveur de l'association ou non avec un phénomène, les valeurs de tests constituent l'alphabet d'un texte et on recherche la présence de classes où se succèdent les très faibles valeurs de tests. Ensuite une fois ces classes sont identifiées, il faut les classer par ordre d'importance (les plus riches en valeurs faibles étant les plus importants) et évaluer le nombre de régions à retenir pour leur modélisation sous forme d'un modèle graphique probabiliste. Un défi important est alors de modéliser un grand nombre de variables sachant la complexité algorithmique d'apprentissage d'un modèle graphique probabiliste qui est exponentielle d'exponentielles selon l'augmentation du nombre de variables.

2 Problématique de l'apprentissage de structure d'un réseau bayésien à partir de données

Le nombre de graphes acycliques dirigés (DAG) pour représenter toutes les structures possibles pour les réseaux bayésiens est de taille super-exponentielle.

En effet, Robinson (1977) a prouvé qu'il était possible de donner ce nombre grâce à la formule récursive suivante :

$$r(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-1)} r(n-i) = n^{2^{O(n)}}$$

Ce qui donne :

r(1)=1

r(2)=3

r(3)=25

r(5)=29281

r(10)=$4{,}2 * 10^{18}$

Comme l'équation est super exponentielle, il est impossible d'effectuer un parcours exhaustif en un temps raisonnable dès que le nombre de nœuds dépasse 7.

La plupart des méthodes d'apprentissage de structure utilisent des heuristiques de recherches dans l'espace des graphes acycliques dirigés.

3 Analyse simple variable

Le test de khi-deux noté χ^2 permet de valider des hypothèses concernant une propriété concrète contenue dans une base de cas. Il existe différentes méthodes basées sur le test de χ^2, principalement les tests d'homogénéité et les tests d'adéquation. Dans notre étude nous l'utiliserons principalement pour le test d'indépendance basée sur les tableaux de contingences, avec une base d'exemples sur deux attributs, nous pouvons construire un tableau d'occurrences conjointes des différentes valeurs pour ces variables.

4 Classification des variables

La classification est le fait de créer des groupes de variables de manière à rassembler les variables qui portent les mêmes informations (redondantes, corrélées), et dissocier les variables qui expriment des informations complémentaires.
On se pose dans le contexte qu'une stratégie d'analyse de type simple variable soit limitée pour élucider l'ensemble des mécanismes impliqués pour le surgit d'un phénomène, la solution qu'on a proposée est la classification des variables et le calcul de score pour chaque classe.
La classification est le fait de créer des groupes de variables de manière à rassembler les variables qui portent les mêmes informations ou qui sont fortement corrélés.
On utilise la classification pour comprendre les structures sous-jacentes qui organisent les données (oppositions, complémentarité, concomitance)
Dans notre contexte on va utiliser une méthode non hiérarchique qui produit directement une partition des individus en un nombre fixé de groupes.

5 Fusion des scores de chaque classe

La question qui se pose maintenant est comment on va déduire un score qui représente toutes les variables d'une classe connaissant les scores statistiques de chaque variable.
Après une étude des méthodes utilisées en matière de fusion de score statistique dans la littérature en informatique et plus précisément en fouille et extraction de connaissances à partir de données on a pu trouver seulement l'utilisation de méthodes qui fusionnent les scores de variables indépendantes comme par exemple : Moyenne où Max des scores (H.N. Prakash and D.S. Guru 2010) et (Damian M. lyons and D. Frank Hsu 2009), Somme, min où produit des scores (Karthik Nandakurman and Anil K. Jain 2005).

Vu l'insuffisance des méthodes cités dans la littérature spécialisée en informatique on a utilisé une méthode statistique qui s'appelle Truncated Product Method (Zaykin, D. V. et al 2002) dont l'algorithme est le suivant :

Algorithme Truncated Product Method

1: Construction d'une matrice de corrélation pour les variables de chaque classe

2: Calculer le facteur de cholesky noter C pour chaque Matrice

3: Choisir la valeur maximale π des p-valeurs qui seront choisies

4: Calculer $W_0 = \prod_i^L p_i^{I(p_i \le \pi)}$

5: A=0

6: Générer aléatoirement L valeur indépendante $u_1^*,.....,u_L^* \in [0,1]$ et qui forment le vecteur R^*

7: Transformer le vecteur R^* en un autre vecteur R ayant les valeurs $u_1,...,u_L$ avec l'équation suivante : $R = 1 - \Phi\left\{C\Phi^{-1}(1-R^*)\right\}$

//notons $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$ et $\Phi^{-1}(x) = \int_{x}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$

8: Calculer $W = \prod_i^L u_i^{I(u_i \le \pi)}$

9: Si $W \le W_0$ alors incrémenter A par 1

10: Répéter les étapes 6-9 B fois

11: La p-valeur combinée = A/B

6 Classement

On peut appliquer des transformations aux scores de manière à ce qu'une forte valeur de score implique un haut degré d'association.

On peut appliquer la transformation Logarithmique suivante :

$-\text{Log}_{10}$(p-valeur)

Après la transformation des scores, un tri décroissant sera effectué sur les classes selon leur score.

7 Sélection des classes très liées aux phénomènes

Le but de cette étape est de proposer un ensemble de régions.

On a constaté la quasi absence, dans la littérature en informatique, de véritables méthodes où algorithmes pour proposer un ensemble de classes alors on a proposé l'utilisation de la méthode de Karlin et Altschul (1993)

Principe

On considère que parmi les k scores identifiés, on va retenir les r premiers.

Les scores $H^1, H^2, ..., H^K$ sont utilisés dans une nouvelle somme de statistiques de la manière suivante:

$$T^i = \sum_{j=1}^{i} H^j$$

$$T^1 = H^1$$

$$T^2 = H^1 + H^2$$

$$T^i = H^1 + H^2 + ... + H^i$$

Des p-valeurs $(P_T^1, ..., P_T^k)$ seront estimées pour obtenir la distribution empirique de chaque T^i

La p-valeur est alors estimée par la proportion de valeurs simulées qui dépasse la valeur observée T^{obs}

$$Pt = \frac{Cardinalité\left\{T^B \geq T_{obs}\right\}}{B}$$

Karlin et Altshul (1993) considèrent que l'ajout d'une région est intéressant s'il ne diminue pas la significativité de la sélection ($P_T^{(i+1)} \leq P_T^{(i)}$)

r correspond aux premières régions où il n'y a pas d'augmentation de la p-valeur.

r = argmin ($P_T^{(i+1)} \geq P_T^{(i)}$).

8 Apprentissage des classes sélectionnées sous forme d'un réseau bayésien

Dans les premières parties de notre travail on a déjà définie des méthodes pour le filtrage du nombre classes selon leurs implications sur un phénomène donné.

Dans cette partie on va présenter une nouvelle méthode d'apprentissage de structure d'un réseau bayésien en s'inspirant du formalisme de (Laskey 2007) appelé « Réseau bayésien multi-entités ».

Laskey (2007) propose un formalisme qui unifie la logique du premier ordre et la théorie des probabilités. Ce formalisme s'appelle les réseaux bayésiens multi-entités (MEBN pour Multi-entity Bayesian network). Les MEBN sont formés de fragments (appelés MFrags) qui représentent alors la distribution jointe d'un sous-ensemble de variables.

Application

Un apprentissage de structure sera fait entre les variables de chaque classe et la variable du phénomène qui sera le premier nœud du graphe (Figure 1).

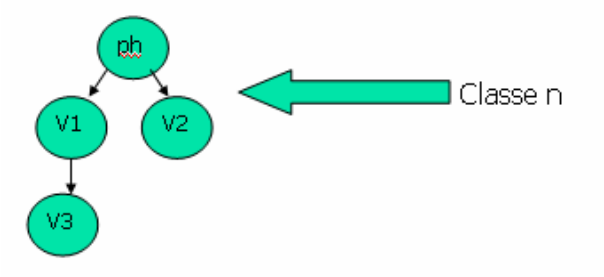


Figure 1 : Apprentissage de structure d'une classe n

Ensuite un apprentissage de structure sera fait entre les variables de toutes les classes qui ont une relation directe avec la variable phénomène et cette dernière. Enfin les noeuds restants des classes seront ajoutées au graphe final (Figure 2).

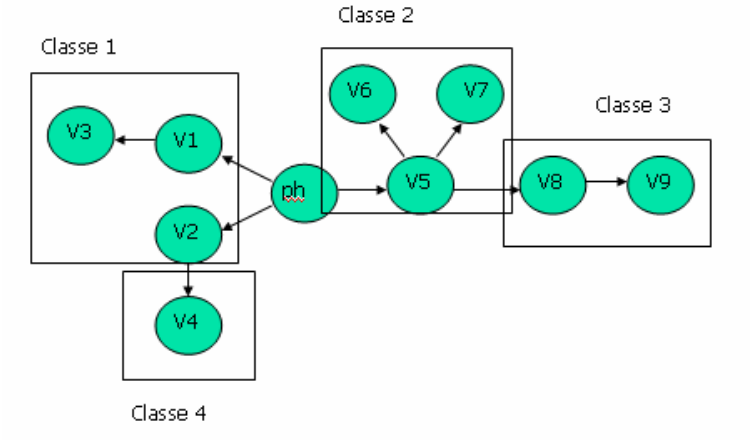


Figure 2 : Apprentissage de structure d'un réseau bayésien multi entités.

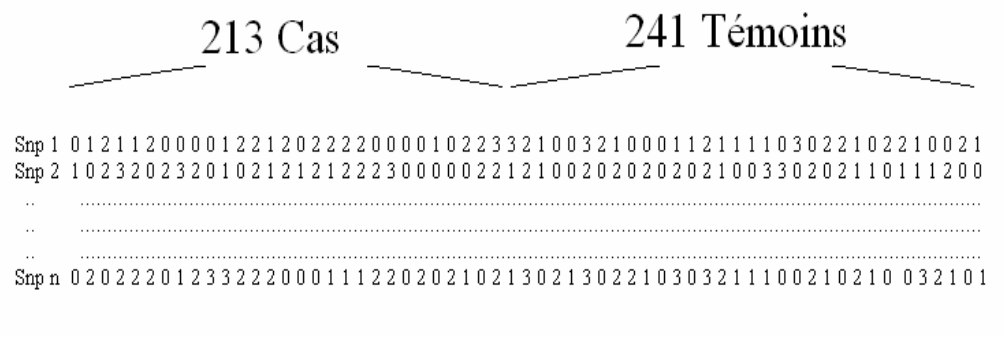
9 Données d'applications

On va s'intéresser aux données génomiques vu le nombre énorme des données et des informations que contient le génome humain et qui nécessite des méthodes évolués et optimisés pour en extraire des informations

Les ADN de 213 patients atteints de schizophrénie et de 241 témoins Canadien ont été génotypés pour ce jeu de marqueurs.

Les données d'études consistent en 164 SNPs sur le chromosome 13

La base de données consiste en un fichier texte remplie de la façon suivante :



0 : correspond au génotype aa

1 : correspond au génotype aA

2 : correspond au génotype AA

3 : correspond à une donnée manquante

Objectif du travail pour les données d'application

L'objectif dans un premier temps sera d'analyser des données génomiques pour la détection de régions hautement significative dans les études génétiques des maladies héréditaires complexes. Dans un deuxième temps l'objectif sera de modéliser les régions les plus intéressantes sous formes d'un réseau bayésien afin de servir comme outil de diagnostic.

Conclusion

Avec notre démarche on a pu identifier les classes les plus impliquées à un phénomène afin de les modéliser, mais pour le moment on n'a pas encore démontré la différence en termes de complexité entre l'apprentissage de structure classique et l'apprentissage de structure selon le formalisme « réseau bayésien multi entités ». Notre démarche d'organisation et de filtrage de données a fait ses preuves avec des données biologiques, reste à la vérifier sur une autre base de données biologiques ou sur d'autres types de données comme par exemple des données financières (exemple : calcul du risque financier de solvabilité d'un client pour octroyer un crédit bancaire).

Bibliographie

Damian, M.,Frank Hsu D. Combining multiple scoring systems for target tracking using rank-score characteristics. *Information Fusion.* 10 (2009) 124-136.

Detera-Wadleigh, S. D. and McMahon, F. J. (2006). G72/g30 in schizophrenia and bipolar disorder: review and meta-analysis. *Biological Psychiatry*.

Dempster, A. P., Laird, N. M. and Rubin D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J of the Royal Stat Soc* B 39: 1-38.

Geudj, M., Robelin, D., Hoebeke, M., Lamarine, M., Wojcik, J., and Nuel, G. (2006). Detecting Local High-Scoring Segments: a First-Stage Approach for Genome-Wide Association Studies. *Statistical Applications in Genetics and Molecular Biology*, 22.

Jain, A., Nandakumar, K., Ross, A. Score normalization in multimodal biometric systems. *Pattern Recognition* volume 38 Issue 12, Decembre 2005, Pages 2270-2285.

Karlin, S. and Altschul, S. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Science USA* 90, 5873-5877.

Laskey, K. B., MEBN: A language for first-order Bayesian knowledge bases. *Artificial Intelligence* 172 (2008) 140-178.

Leray, PH., Francois, O., Evaluation of structure learning algorithms for Bayesian networks. *Article INSA Rouen – Laboratoire PSI – FRE CNRS 2645*, [En ligne]. Adresse URL : http://www.laas.fr/rfia2004/actes/ARTICLES/123.pdf.

NATIONAL CANCER INSTITUTE (Page consultée le 20 septembre 2009) *Distribution of SNPs by chromosome*, [En ligne]. Adresse URL : http://gai.nci.nih.gov/cgi-bin/histo.cgi?c=13&o=h/.

Prakash, H. N., Guru, D. S. Offline signature verification : An approach based on score level fusion. *International journal of computer applications* (2010) 0975-8887 volume 1 No.18.

Robinson, R. W. Counting unlabeled acyclic digraphs. *Dans Little, C. H. C. (Ed.), Combinatorial Mathematics V* (1977), volume 622 of Lecture Notes in Mathematics, (pp. 28–43)., Berlin.Springer. 14, 64

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., Weir, B. S. Truncated product method for combining P-values *Wiley Inter Science.*