Système d'extraction des connaissances à partir des données temporelles basé sur les Réseaux Bayésiens Dynamiques

Ghada Trabelsi*, Mounir Ben Ayed*, Adel M. Alimi*
*REGIM : Research Group on Intelligent Machines
Université de Sfax, Ecole Nationale des Ingénieurs de Sfax (ENIS)
Sfax, Tunisie
Trabelsi_Ghada@yahoo.fr, mounir.benayed@ieee.org, Adel.Alimi@ieee.org

Résumé. Un grand nombre d'informations qui ont une structure complexe proviennent de diverses sources. Ces informations contiennent des connaissances très utiles pour l'aide à la décision. L'Extraction des Connaissances à partir des Données (ECD), permet d'acquérir des informations pertinentes pour les systèmes interactifs d'aide à la décision (SIAD). Mais, dans plusieurs domaines, les données évoluent d'une manière dynamique et finissent par dépendre de plusieurs dimensions. Les Réseaux Bayésiens dynamiques (RBD) sont des modèles représentant des connaissances incertaines sur des phénomènes complexes de processus dynamiques. Notre objectif revient à fixer les meilleures modèles de connaissances extraites par les RBD et à les utiliser pour la prise de décision dynamique. Ainsi, Nous proposons dans cet article une démarche pour la mise en place d'un processus d'extraction des connaissances à partir des données multidimensionnelles et temporelles.

1 Introduction

Depuis les années soixante, on stocke de plus en plus des données multidimensionnelles et temporelles. Les connaissances sont utilisées pour aider les décideurs à prendre la décision. Certaines de ces connaissances peuvent être extraites à l'aide d'un outil décisionnel qu'est l'ECD [FAY, 96]. Dans plusieurs domaines tels que celui de la santé les données concernant un individu sont saisies à des moments différents d'une manière plus ou moins périodique. D'où la nécessité de développer un SIAD qui évolue au cours du temps appelé un SIAD dynamique (SIADD).

Dans ce travail, nous nous sommes intéressés à la technique des Réseaux Bayésiens Dynamiques comme technique de fouille de données multidimensionnelles et temporelles. Le contexte applicatif de notre projet est la lutte contre les infections nosocomiales (IN) des patients hospitalisés dans le service de réanimation du CHU Habib Bourguiba de Sfax (Tunisie) [KAL, 05]. Une infection est considérée comme telle lorsqu'elle apparaît après un délai de 48 heures d'hospitalisation [GAM, 88]. La durée d'hospitalisation est différente d'un patient à un autre. Ainsi, le nombre des séries de données des différents patients n'est pas la même. Le médecin doit continuellement étudier les données du patient avant de prendre une décision selon son état et son évolution. Les données dont nous disposons sont temporelles et multidimensionnelles. La décision à prendre dépend de l'état actuel du patient et de ses observations antérieurs.

Le présent article est organisé en quatre sections. Dans la section 2, nous présenterons l'état de l'art notre travaille en particulier les SIAD, les données temporelles et la technique d'apprentissages adoptée. Dans section 3, nous expliquons l'aspect dynamique de la décision dans notre travail et le contexte qui est la lutte contre les infections nosocomiales. Dans la

section 4, nous discuterons les résultats obtenues du système SIADDM (SIADD médical). Finalement, une conclusion et plusieurs perspectives seront soulignées.

2 L'état de l'art

2.1 Les SIAD et les bases de données temporelles

Les SIAD : leur concept a tout d'abord été introduit dans les années 70. Il est la traduction du concept de Decision Support Systems (DSS). Un des tous premiers auteurs à l'avoir introduit et diffusé est Scott Morton [MOR, 71] qui a introduit la notion de Systèmes de décision et de Gestion (Management Decision Systems). Sprague et al. ont défini les SIAD comme des systèmes informatisés, interactifs qui aident les décideurs en utilisant des données et des modèles pour résoudre des problèmes mal structurés [SPR, 82].

Certaines décisions nécessitent l'accès à des connaissances plus approfondies que d'autres. Il y a donc un lien entre les systèmes d'aide à la décision et les systèmes de connaissance [LEP, 05].

Une base de données temporelle : est une large collection de séries de temps. Georges et al. ont défini une série de temps comme une séquence temporelle, qui est une suite de couples $<(v_1,t_1),(v_2,t_2),...,(v_i,t_i),...>$ où v_i est une valeur, ou un vecteur de valeurs, prises à un instant t_i [GEO, 07]. Mais dans la plupart des cas, et en particulier dans celui de séries financières et médicales, les séries sont des processus stochastiques, très bruités et non stationnaires.

2.2 Les Réseaux Bayésiens dynamiques

2.2.1 Définition et représentation

Un Réseau Bayésien dynamique RBD est un Réseau Bayésien spécial qui est utilisé pour les modèles de processus dynamique stochastique [MUR, 02].

Un RBD satisfait, presque toujours, les conditions suivantes : (a) il est Time-invariant (b) chaque arc de temps est étendu d'une tranche de t vers la tranche t+1. (c) nous supposons que les variables de chaque tranche sont connectées de la même manière.

Un RBD est un RB qui modélise des distributions de la probabilité sur collections semi-infinies de variables aléatoires, Z_1, Z_2, ….. . Un RBD est défini par une paire, $(B_1, B_{\rightarrow})$, où B_1 est un RB qui définie l'a priori P(Z1), and $B_{\rightarrow}$ représente deux tranches temporelles du Réseau Bayésien (2TRB) par lequel on définie $P(Z_t|Z_{t-1})$ pour un DAG (directed acyclic graph) comme suit :

$$P(Z_t \mid Z_{t-1}) = \prod_{i=1}^{N} P(Z_t^i \mid Pa(Z_t^i))$$

Z^i_t est le $i^{ème}$ nœud au temps t. $Pa(Z^i_t)$ sont les parent de Z^i_t . Les parents d'un nœud, $Pa(Z^i_t)$ peuvent exister dans la même tranche du nœud ou de la tranche de temps précédente.

2.2.2 Inférence dans un RBD

Le problème d'inférence d'un RBD est analogue à celui d'un RB, de telle façon que la quantité désirée est la distribution marginale postérieure d'un ensemble de variables cachées indi-

quant un ordre dans les observations. L'objectif de l'inférence dans un RBD consiste à calculer la probabilité $P(X^i_t|y_{1:\tau})$. Une approche directe pour le calcul d'inférence est effectuée par des messages qui définissent des variables sur une couverture de Markov d-séparée (sépare le passé du futur) et emploie la procédure *forward- backward* [RAB, 78]. Cette procédure est utilisée pour distribuer l'évidence tout au long du RBD [ZWE, 98].

3 Mise en place d'un RBD pour un SIAD basé sur l'ECD

3.1 Modélisation

3.1.1 Construction du modèle de connaissance des données fixes (RB Statique)

Nous avons adopté la méthode basée sur les scores dans l'apprentissage des modèles statiques. Notre construction des modèles est basée sur la théorie de l'information. La notion de d-séparation [BEC, 99] joue un rôle important. La quantité d'écoulement de l'information entre deux nœuds peut être mesurée, en employant l'information mutuelle ou l'information mutuelle conditionnelle. La figure FIG.2-C représente un modèle de connaissance extrait de nos données fixes, en utilisant l'algorithme K2.Ce modèle a pu détecter des relations entre les variables logiques comme la relation entre l'âge et l'antécédent, entre l'âge et cissue (le patient est décédé ou a survécu). Cependant le graphe obtenu, contient des liens « illogiques» entre les nœuds (par exemple, l'âge agit sur la prise d'antibiotique). Nous avons aussi noté des liens manquants qui présentent des relations d'indépendance intéressantes (ex : la relation entre result (nœud cible « infnos » c-à-dire un patient attrape une IN ou non) et cissue).

Pour l'apprentissage des paramètres, nous avons utilisé dans cette phase la méthode de maximum de vraisemblance. La figure FIG.1 présente les probabilités P(age1|result) pour tous ces états et marque la probabilité P(age1=âgé|result=oui).

age1/result	Non	Oui
adulte	0,560693641618497	0,428571428571429
agé	0,271676300578035	0,457142857142857
petit	0,167630057803468	0,114285714285714

FIG. 1- La table de probabilité conditionnelle correspondante au nœud age1

3.1.2 Construction des modèles de connaissance des données temporelles (RBD)

Nous avons développé la première partie de notre système qui s'intéresse à l'apprentissage à partir des données fixes. Le résultat de cette réalisation a donné un RB statique. L'utilisation conjointe des probabilités et du graphe nous offre une famille de modèles de connaissances qui ne sont pas très riches. L'étape suivante consiste à apprendre les données temporelles et à élaborer des modèles de connaissances dynamiques qui sont plus riches que les modèles statiques. Nous allons présenter la structure du RBD obtenu.

Notre réseau vérifie les conditions mentionnées dans la section 2.2.1. L'objectif de notre RBD est de prédire l'événement d'acquisition d'une IN par un patient et ceci tout au long de son hospitalisation, par le calcul de la probabilité quotidienne d'avoir IN sachant son état dans le passé et ses observations (c.à.d. les actes et les examens infectieux effectués) au

cours du jour courant. Un 2TRB représente des relations entre les variables qui sont dépendantes entre elles dans deux tranches de temps données comme l'illustre la figure FIG.2- B.

La figure FIG.2- B montre que 2TRB est composée par deux types d'arcs : (1) Les arcs intra-tranche qui représentent l'interdépendance entre les variables d'une tranche. Nous avons utilisé pour cette structure celle du RB naïf puisque les variables de chaque série de temps (acte1, ….. acte10, exinf1…. exinf30) n'ont aucune dépendance entre elles et qu'elles sont toutes connectées directement avec les deux nœuds cibles (result et cissue). (2) les arcs inter-tranche qui représentent les arcs de temps. En effet, ces arcs symbolisent les interdépendances qui existent entre la variable elle-même mais pour des tranches de temps différentes et successives. Notre système a appris que le résultat de prédiction d'un jour t dépend directement du résultat de t-1. Aussi, nous avons inséré dans notre modèle dynamique des connaissances données par l'expert en ce qui concerne les examens infectieux.

Notre RBD modélise des distributions de la probabilité sur les collections semi-infinies de variables aléatoires ($result_i$, act_i,……, $exminf_i$, ….., $cissue_i$). Le principe de notre RBD peut être défini par "dérouler" les 2TRB jusqu'à ce que nous ayons tous les T jours d'hospitalisation du patient. Le résultat de la distribution des probabilités jointes est donné par:

$$P(result_{1:T}) = \prod_{t=1}^{T} \prod_{i=1}^{N} P(result_t^i \mid Pa(result_t^i))$$

Où T est l'intervalle de temps d'hospitalisation, N est le nombre total des variables pour chaque tranche de temps. Nous obtenons un Réseau Bayésien dynamique final ayant comme structure celle représentée par la figure FIG.2-A:

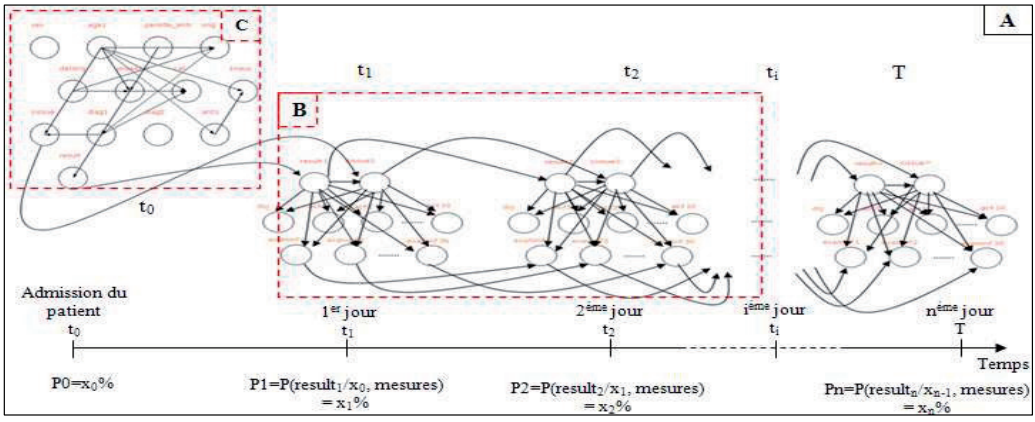

FIG. 2- A) Le graphe causal de notre RBD. B) Représentation de 2TRB de notre réseau dynamique. C) Modèle statique extrait par l'algorithme K2.

4 Expériences et résultats

Notre étude concerne la prédiction de l'état du patient. Cette prédiction est dynamique elle évolue tout le long de l'hospitalisation du patient par des nouvelles mesures. Ces mesures enrichissent les modèles pour donner d'autre prédiction. A chaque jour^{t} d'hospitalisation du patient nous avons pu prévoir son état au futur par une probabilité, qui sera utilisée, dans la prédiction du jour^{t+1}, avec ces observations mesurées. Nous avons utilisé une base d'apprentissage qui contient 200 cas (patients) et une base de test qui contienne 23 cas, pour l'évaluation de la performance de notre système. Nous avons obtenu les résultats donnés par la matrice de confusion suivante :

Les résultats observés/ Les résultats prédits	Oui	Non	Totaux
Oui	5	2	7
Non	3	13	16
Totaux	8	15	23

TAB. 1- La matrice de confusion des résultats donnés par notre Réseau Bayésien dynamique

Nous avons calculé les taux d'évaluation à partir des résultats de prédiction obtenus par notre structure élaborée par le RBD. Nous avons trouvé que le taux des classifications dynamique étaient correctes à 0.78±0.17.Les autres taux sont illustrés par la figure FIG.3.

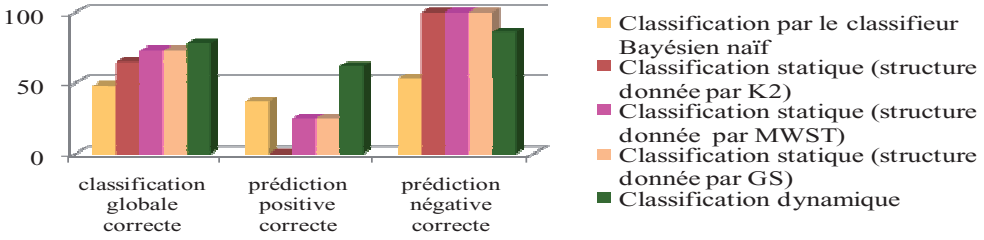

FIG.3- Les taux globaux de classification des algorithmes

L'histogramme ci-dessus présente des taux élevés pour la classification dynamique. Nous avons obtenu d'autres taux relativement acceptables pour la classification statique en utilisant les structures des réseaux donnés par les algorithmes d'apprentissage de structure comme K2, MWST et GS. Et enfin, nous avons obtenu des taux inférieurs ou égaux à 50% pour la classification donnée par le classificateur Bayésien naïf.

5 Conclusion

Dans ce travail nous avons démontré l'efficacité des RBD pour l'extraction des connaissances à partir de données temporelles et multidimensionnelles. Nous avons proposé une démarche ayant un aspect dynamique composée des algorithmes d'apprentissage et d'inférence. Ces algorithmes sont appliqués sur des données réelles provenant d'un service de réanimation. Nous avons obtenu des résultats quantitatifs (probabilistes) et qualitatifs pour la prédiction. Les résultats de prédiction de notre système sont fiables à 78%.

Une extension de la phase de prédiction sera faite, d'une prédiction offline à online. Cette dernière sera effectuée à chaque observation détectée et à chaque instant.

Références

[BEC, 99] Becker A., P. Naïm (1999), "*les Réseaux Bayésiens: Modèles graphiques de connaissance*", Editions Eyrolles, ISBN 9782212090659.

[FAY, 96] Fayyad U.M. (1996), Djorgovski S.G., et Weir N., "*Automating the Analysis and Cataloging of Sky Surveys*", Advances in Knowledge Discovery and Data Mining, MIT Press, pages 471-494.

[GAM, 88] Garner J.S., Jarvis W.R, Emori T.G., Hogan T.C., Hugues J.M., "*CDC definitions for nosocomial infections*". Am. J. of Inf. Contr., 16 (3), pp. 128-140, 1988.

[GEO, 07] Georges. J, Luciano. M, (2007)," *Prediction of financial time series with Time-Line Hidden Markov Experts and ANNs*", Wseas Transactions on Business and Economics, Issue 9, Vol4, pp140-146.

[KAL, 05] Kallel H., Bouaziz M., Ksibi H., Chelly H., Hmida C.B., Chaari A., Rekik N., Bouaziz M. "*Prevalence of hospital-acquired infection in a Tunisian Hospital*", Journal of Hospital Infection, 59, pp. 343-347, 2005.

[LEP, 05] Lepreux. S. (2005), "*Approche de Développement centré décideur et à l'aide de patrons de Systèmes Interactifs d'Aide à la Décision*", Thèse Doctorat, l'Université de Valenciennes et du Hainaut-Cambrésis.

[MOR, 71] Morton Scott M (1971), "*Computer based support for decision making*". Management decision systems, Harvard University, Boston, MA, USA.

[MUR, 02] Murphy. K. P (2002). "*Dynamic Bayesian Networks: Representation, Inference and Learning*", PhD thesis. UC Berkeley, Computer Science Division.

[RAB, 89] Rabiner L.R. (1989), "*A tutorial on Hidden Markov Models and Selected applications in speech Recognition*", Proceedings of the IEEE, Vol77, NO2.

[SPR, 82] Sprague R. et Carlson E. (1982), "*Building Effective Decision Support Systems*". Prentice-Hall, Inc, Englewood Cliffs.

[ZWE, 98] Zweig. G (1998). "*Speech Recognition with Dynamic Bayesian Networks*". Ph.D. Thesis, University of California, Berkeley.

Summary

A big number of information that has a complex structure comes from various sources. The knowledge Discovery from data (KDD) permits to acquire some useful information to help making decision. But, in several domains, this data evolve in a dynamic manner and depends on many dimensions. The Dynamic Bayesian Networks (DBN) are models representing some uncertain knowledge on the complex phenomena of dynamic processes. Our objective comes back to fix the best models of knowledge extracted by the DBN and to use in the dynamic decision making. Thus, we propose in this paper a gait for the setting up of KDD process from the multidimensional and temporal data.