Approche entropique pour l'analyse de modèle de chroniques

Nabil Benayadi*, Marc Le Goc*, Philippe Bouché*.

*Laboratoire des Sciences de l'Information et des Systèmes - LSIS
UMR CNRS 6168 - Université Paul Cézanne
Avenue Escadrille Normandie Niemen13397 Marseille Cedex 20 – France
{nabil.benayadi, marc.legoc, philippe.bouche}@lsis.org

Résumé. Cet article propose d'utiliser l'entropie informationnelle pour analyser des modèles de chroniques découverts selon une approche stochastique (Bouché et Le Goc, 2005). Il décrit une adaptation de l'algorithme TemporalID3 (Console et Picardi, 2003) permettant de découvrir des modèles de chroniques à partir d'un ensemble d'apprentissage contenant des séquences d'occurrences d'événements discrets. Ces séquences représentent des suites d'alarmes générées par un système à base de connaissance de monitoring et de diagnostic de systèmes dynamiques. On montre sur un exemple que l'approche entropique complète l'approche stochastique en identifiant les classes d'événements qui contribuent le plus significativement à la prédiction d'une occurrence d'une classe particulière.

1 Introduction

La découverte de connaissances temporelles est un enjeu majeur pour le diagnostic de systèmes dynamiques (Das et al., 1998), (Dousson et Vu Duong, 1999), (Keogh et Smyth, 1997), (Agrawal et al., 1995), (Faloutsos et al, 1994). Récemment, Bouché P. et Le Goc M. (2005) ont proposés une approche stochastique pour découvrir des modèles de chroniques à partir d'une séquence d'événements discrets. Nos travaux visent à compléter cette approche pour identifier les classes d'événements contribuant le plus significativement à la prédiction de l'occurrence d'une classe particulière.

Les arbres de décisions (Breiman, 1984), (Murthy, 1998), sont largement utilisés pour classer des séquences de données (Kadous, 1999), (Geurts, 2001), (Drucker et Hubner, 2002), (Rodriguez et Alonso, 2004). Récemment, l'algorithme ID3 (Quinlan, 1986) a été adapté pour construire des arbres temporels de décision (Console et al., 2003) à partir d'un ensemble de situations. Cette adaptation montre que l'entropie informationnelle permet d'identifier les variables contribuant le plus significativement à une prise de décision.

Nous proposons donc d'utiliser un critère entropique pour analyser des modèles de chroniques. Après un bref rappel sur les arbres temporels de décision, cet article présente une adaptation de l'algorithme proposée par Console pour la déduction de modèles de chroniques à partir d'un ensemble de séquences d'occurrences d'événements discrets et montre sur un exemple comment l'approche entropique peut être utilisée pour compléter l'approche stochastique.

2 Arbres temporels de décision

Un arbre de décision est une structure T= <r, N, E, L> où $N=N_I \cup N_L$ est l'union d'un ensemble $N_I=\{x_i\}$ de nœuds internes désignant une variable x_i et un ensemble $N_L=\{a_i\}$ de nœuds feuilles désignant une décision ai, $r \in N$ est le nœud racine de l'arbre, $E \subseteq N_I \times N$ est un ensemble d'arcs, un arc attribuant une valeur v_j à une variable x_i et L est une fonction d'étiquetage définie sur $N \cup E$, qui retourne le nom de la variable x_i associé au nœud de N_I, la décision a_i associée à une feuille de N_L ou la valeur v_J associée à un arc de E.

L'algorithme ID3 utilise l'entropie informationnelle dans un ensemble de cas Σ pour construire un arbre de décision de profondeur minimal. Un cas e est une collection de valeurs v_j prises par un ensemble de variables x_i conduisant à une décision a_i particulière. A chaque nœud, ID3 choisit la variable x qui minimise l'entropie ξ(x, Σ) dans l'ensemble Σ des cas :

$$\xi(x,\Sigma) = \sum_{j=1}^{k} P(x = v_j) \times \xi(\Sigma|_{x=v_j}) \qquad (1)$$

$$\xi(\Sigma|_{x=v_j}) = -\sum_{i=1}^{n} P(a_i;\Sigma|_{x=v_j}) \times \log_2(P(a_i;\Sigma|_{x=v_j}))$$

$$\Sigma|_{x=v_j} = \{e \in \Sigma \mid \text{la variable x a la valeur } v_j \text{ dans e}\} \qquad (2)$$

où $\xi(\Sigma|_{x=v_j})$ est l'entropie de la partition $\Sigma|_{x=v_j}$ et $P(a_i;\Sigma|_{x=v_j})$ est la probabilité de la décision a_i dans cette partition $\Sigma|_{x=v_j}$.

Temporal ID3 est une extension d'ID3 à des données datées selon une horloge à temps discrète (Console et al, 2003). Un arbre temporel de décision est un arbre de décision où un nœud est un couple (x_i, t_k), x_i désignant une variable et t_k la date d'observation de sa valeur, et un arc défini une valeur v_j de x_i à la date t_k (i.e. $x_i(t_k)=v_j$). Il s'agit donc d'une structure T= <r, N, E, L, Γ> dotée d'une fonction d'étiquetage du temps Γ: $N_I \rightarrow \Re^+$ qui donne la date associée à un nœud interne. L'ensemble d'apprentissage est une collection de situations $S=\{s_{e=0,..,m}\}$. Une situation s_e est l'ensemble des valeurs v_j prises par un ensemble de variables $X=\{x_i\}$ à chaque instant d'observation t_k conduisant à une décision a_n particulière (Table 1). Une situation s_e réfère à une horloge à temps discret où $t_k \equiv kT$, $k \in \aleph$ et $T \in \Re^+$, T est une période d'échantillonnage.

	x_1				x_2				x_3				**Dec**	**Dl**
	t_0	t_1	t_2	t_3	t_0	t_1	t_2	t_3	t_0	t_1	t_2	t_3		
s_1	n	v	n	n	h	h	n	n	l	v	v	v	a_1	t_3
s_2	h	h	v		h	n	n		h	n	n		a_2	t_2

TAB. 1 – *Exemple de table temporelle de décision*.

Dans la table 1, les variables x_1, x_2, x_3 prennent une valeur qualitative n, h, l, ou v aux instants d'observations t_0, t_1, t_2, t_3. Dans la première situation s_1 (resp. s_2), la décision a_1 (resp. a_2) doit être prise au plus tard à la date t_3 (resp. t_2). Cette date est dite « limite » car la connaissance de la valeur des variables au-delà de cette date est inutile à la prise de décision.

$$S_{\approx,t} = \{S_{e=0,1,\ldots,m} \mid \forall S_e \in S, \forall t_k \in [t, Dl(S_e)], \forall s_i, s_j \in S_e, \forall x \in X, \\ Ttd[s_i, x, t_k] = Ttd[s_j, x, t_k] \wedge Dl(S_e) = \min\{Dl(s_l) \mid s_l \in S_e\}\} \quad (3)$$

Une partition S_e est un sous ensemble de S contenant des situations identiques sur un intervalle de temps : $\forall t_k \in [t_{min}, t_{max}]$, $\forall x \in X$, $\forall s_i, s_j \in P$, $Ttd[s_i, x, t_k] = Ttd[s_j, x, t_k] \Rightarrow s_i \equiv s_j$. Ainsi, à chaque instant d'observation t, S est partitionné en un ensemble de partitions $S_{\approx,t} = \{S_{e=0,\ldots,m}\}$ (équation 3). TemporalID3 construit un arbre en recherchant un intervalle de temps qui maximise un critère lié au nombre de partitions. Puis, de la même manière qu'ID3, TemporalID3 choisit la variable $x_i(t)$ qui minimise l'entropie sur cet intervalle (4) et créé le nœud correspondant. Toutes les valeurs des variables à tous les instants précédent t sont éliminées du tableau, y compris celle de $x_i(t)$, puis TemporalID3 recommence son traitement.

$$\xi((x_i, t), S) = \sum_{j=1}^{k} P(x_i(t) = v_j) \times \xi(S|_{x_i(t)=v_j}) \quad (4)$$

3 Représentation des séquences

Nous proposons d'adapter l'algorithme TemporalID3 à un ensemble d'apprentissage $\Omega = \{\omega_n\}$ contenant des séquences ω_n d'occurrences d'événements discrets classées en séquences Ok et Ko selon qu'elle conduisent ou non à une occurrence d'une classe d'événements particulière. L'adaptation consiste pour l'essentiel à représenter un ensemble de séquences dans une table similaire à celle de la table 1, mais où les valeurs sont datées selon une structure de temps continu (i.e. $t_{k+1} - t_k \neq T$) : une occurrence d'événement discret indique le fait qu'une variable change de valeur et la date de ce changement.

$$\forall o_k \equiv (t_k, x, i), o_{k+1} \equiv (t_{k+1}, x, j) \in \omega, \\ (o_k, o_{k+1}) \Rightarrow \forall t \in [t_k, t_{k+1}[, x(t) = i \wedge x(t_{k+1}) = j \quad (5)$$

Une séquence $\omega = \{o_k\}_{k=0\ldots m-1}$ est une suite ordonnée de m occurrences $o_k \equiv (t_k, x, i)$ d'événements discrets $e_k \equiv (x, i)$ où $t_k \in \Gamma \subset \Re^+$ est la date de l'affectation de la valeur i à la variable x (Bouché P et Le Goc, 2005). Un couple (o_k, o_{k+1}) de deux occurrences successives liées à une même variable x décrit l'évolution temporelle de la fonction discrète x(t), définie sur $\aleph$ (équation 5). L'ensemble des événements discrets $E_d = \{e_k \equiv (x, i)\}$ est partitionné en un ensemble de classes $C^j = \{e_k\}$. La notation "$o_i :: C^k$" signifie que l'occurrence o_i appartient à la classe C^k. Une fonction d fournit la date d'une occurrence en sorte qu'une séquence ω_n d'un ensemble $\Omega = \{\omega_n\}$ définie son propre sous ensemble $\Gamma_{\omega n} = \{t_j\}$ de date inclus dans l'ensemble des dates Γ défini par Ω (équation 6 où O désigne l'ensemble des occurrences de Ω). Un modèle de chroniques est un ensemble de relations binaires temporellement contraintes entre des classes d'événements (équation 7).

$$d : O \to \Gamma, \quad \forall o_k \equiv (t_j, x, i) \in \omega_n, d(o_k) = t_j \quad (6)$$

$$\forall \omega_n \in \Omega, o_k \in \omega_n \Rightarrow d(o_k) \in \Gamma_{\omega_n} \wedge \Gamma_{\omega_n} \subseteq \Gamma$$

$$R(C^i, C^o, [\tau^-, \tau^+]) \Leftrightarrow \exists o_n, o_k \in \omega \subseteq \Omega, \\ (o_n :: C^o) \wedge (o_k :: C^i) \wedge (d(o_n) - d(o_k) \in [\tau^-, \tau^+]) \quad (7)$$

Une séquence classée Ok, notée ω_n^{ok}, est un exemple qui respecte toutes les contraintes logiques et temporelles d'un modèle de chroniques. Une séquence classée Ko, notée ω_n^{ko}, est un contre exemple. Un contre exemple respecte toutes les contraintes d'un modèle de chroniques sauf la dernière relation binaire qui concerne la classe à prévoir.

Par exemple, considérons une séquence ω_0 de 70 occurrences des classes A, B, C, D décrivant l'évolution des variables v_a, v_b, v_c et v_d sur une période de temps donnée et un modèle de chroniques déduit sur cette séquence selon une approche stochastique (P Bouché et Le Goc, 2005) (figure 1). Ce modèle de chroniques défini un ensemble d'apprentissage $\Omega=\{\omega_n\}$ contenant 8 sous séquences de ω_0 classées Ok et 16 classées Ko (figure 2). Pour pouvoir être comparées, les occurrences doivent être datées en temps relatif, en prenant pour référence la date de la dernière occurrence. Soit tmax la date de la dernière occurrence dans une séquence ω_i définissant un ensemble $\Gamma_{\omega i}$ de dates : $tmax = \max\{t_{k=0,\ldots,n} / t_k \in \Gamma_{\omega i}\}$. L'inversion des dates d'occurrences des classes dans une séquence ω_i s'effectue simplement de la manière suivante : $\forall o_k \in \omega_i$, $d(o_k) = tmax - d(o_k)$. La date limite $Dl(\omega_n)$ d'une séquence ω_n est la date la plus grande dans une séquence inversée : $Dl(\omega_n)= \max\{t_k | \exists o \in \omega_n, d(o)=t_k\}$. Par extension, la date limite globale $Dl(\Omega)$ d'un ensemble de séquences $\Omega = \{\omega_n\}$ est la plus petite des dates limites de chaque séquence : $Dl(\Omega) = \min\{Dl(\omega_n) \mid \omega_n, \in \Omega\}$.

ω_0 = {(0.8774, B), (1.9313, A), (2.8625, C), (3.8718, A), (4.4063, B), (4.7837, D), (6.0282, B), (6.0874, C), (6.2531, A), (8.0034, D), (8.4572, A), (9.2311, A), (9.4742, C), (9.5447, B), (9.8285, A), (10.6631, B), (11.3967, A), (12.9826, B), (13.4464, A), (13.621, B), (13.7333, C), (14.1756, D), (14.6806, A), (15.9598, B), (16.0240, A), (16.7736, C), (18.2447, D), (18.3639, A), (18.9228, B), (19.0749, A), (19.3406, C), (21.1271, B), (21.3377, A), (22.6778, A), (23.0197, D), (23.8478, C), (24.0392, B), (24.3164, A), (25.7974, A), (26.1294, C), (26.5961, A), (26.882, B), (28.3387, B), (28.9188, D), (29.1697, A), (29.7840, B), (31.1968, C), (31.2985, A), (32.1786, A), (33.3798, B), (33.798, A), (33.8452, C), (34.6423, B), (35.1186, A), (36.0294, A), (37.0752, A), (37.4236, B), (37.4543, D), (38.31, B), (39.0201, C), (39.1335, A), (40.2009, B), (41.1502, A), (41.992, C), (42.5301, A), (43.0774, B), (43.4958, A), (43.8625, B), (46.3524, C), (49.6673, C)}.

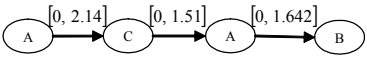

FIG. 1 – *Séquence ω_0 et modèle de chroniques déduit.*

Ω défini un ensemble $X=\cup X_{\omega i=0,..,n}$ de variables x_i et un ensemble $\Gamma=\cup\Gamma_{\omega i=0,..,n}$ d'instants d'observation t_k. Dans l'exemple, X contient 4 variables et Γ 174 dates. La construction d'une table temporelle de décision requiert la connaissance des valeurs de toutes les variables x_i de X à tous les instants d'observation t_k de Γ. L'équation (5) ne permettant pas de déduire toutes les valeurs prises par une variable dans une séquence ω_i donnée, la table de décision doit être complétée en introduisant une occurrence d'événement de la forme $o_k \equiv (t,x, ?)$, $o_k::C^?$ où «$C^?$» désigne la classe des valeurs inconnues.

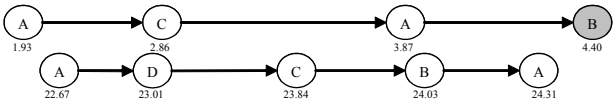

FIG. 2 – *Exemple de séquences Ok (haut), KO (bas).*

4 Extraction d'un modèle de chroniques

La table temporelle à temps continu est une matrice H définie sur $\Omega \times X \times \Gamma$ où un élément $h[\omega_i, x, t]$ défini la classe liée à la valeur de la variable x à l'instant t dans la séquence ω_i tel que, si $o_k \equiv (t, x, j) \in \omega_i$ et $o_k::C^l$, alors $h[\omega_i, x, t] = C^l$, sinon $h[\omega_i, x, t]$ est donné par l'équation 8.

$$\forall o_k \equiv (t_k, x, i), o_{k+1} \equiv (t_{k+1}, x, j) \in \omega,$$
$$o_k :: C^l \wedge o_{k+1} :: C^m \wedge (o_k, o_{k+1}) \Rightarrow \forall t \in [t_k, t_{k+1}[, h[\omega_i, x, t] = C^m \quad (8)$$

La date limite d'une séquence est la date de la dernière occurrence sans prendre en compte les occurrences ajoutées par complétion. La figure 3 montre le résultat de l'application de TemporalID3 sur Ω.

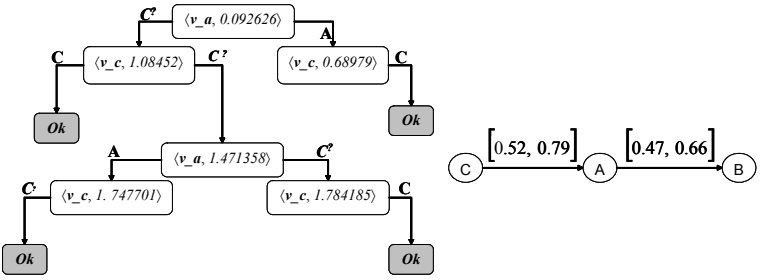

FIG. 3 - *Arbre temporel de décision et modèle de chronique déduit.*

Le modèle de chroniques est construit en ne considérant que les branches menant à une décision Ok en parcourant l'arbre en profondeur d'abord depuis les feuilles terminales à la racine de l'arbre (les dates ont été inversées). Le modèle est initialisé par la classe C^1 correspondant à la décision Ok (i.e. « B »). Pour tout arc $(n_1 \rightarrow n_2) \in E$ de l'arbre, la classe C^2 associée n_2 est créée dans le modèle de chroniques et une relation $R(C^2, C^1, [\tau^1, \tau^2])$ est créée avec $\tau^1 = (\Gamma(n_2) - \Gamma(n_1)) - dm(C^2)$ et $\tau^2 = (\Gamma(n_2) - \Gamma(n_1)) + dm(C^1)$, où $dm(C^i)$ est la moyenne des durées écoulées entre les occurrences de la même classe C^i dans les séquences ω_n de Ω. Le modèle de chroniques de la figure 3 est construit à partir de la branche ((v_a, 0.092626)→(v_c, 0.68979)→(Ok)). La relation R(A, C, [0,2.14]) du modèle de la figure 1 n'apparaît pas dans le modèle produit par l'approche entropique. Cela induit l'idée que cette relation n'apporte que peu d'information pour prédire une occurrence de la classe B.

5 Conclusion

Cet article propose une adaptation de l'algorithme TemporalID3 pour la découverte des modèles de chroniques à partir d'un ensemble de séquences d'occurrences d'événements discrets. L'intérêt de cette approche est d'utiliser un critère de minimisation entropique pour identifier les classes des modèles de chroniques les plus signifiantes afin de prédire l'occurrence d'une classe dans une tâche de diagnostic de systèmes dynamiques. Les premiers résultats obtenus invitent à envisager une combinaison des approches entropiques et stochastiques pour découvrir des connaissances temporelles avec un fort pouvoir anticipatif.

Références

Agrawal, R., K. I. Lin, H. S. Sawhney, et K. Shim (1995). *Fast similarity search in the presence of noise, scaling, and translation in time-series databases*. In Proc. of the 21st Int'l Conf. on Very Large Databases, pp 490-50.

Bouché, P., et M. Le Goc (2005). Analyse stochastique de séquences d'événements discrets pour la découverte de signatures. *Revue des Nouvelles Technologies de l'Information, RNTI-E-3, Extraction et gestion des connaissances (EGC)*, Volume 1, p. 103-114.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification and Regression Trees. Belmont*, CA: Chapman & Hall.

Console, L., C. Picardi et D. Theiser Dupré (2003). Temporal Decision Trees: Model-based Diagnosis of Dynamic Systems On-Board. *Journal of Artificial Intelligence Research,* 19, pp 469-512.

Das, G., K. Lin, H. Mannila, G. Renganathan, et P. Smyth (1998). *Rule Discovery from Time Series*. International Conference on Knowledge Discovery & Data Mining p. 16-22.

Dousson, C., T.Vu Duong (1999). *Discovering Chronicles with Numerical Time Constraints from Alarms Logs for Monitoring Dynamic Systems*. the 1-th International Join conference on Artificial Intelligence (IJCAI'99), pp. 620-626.

Drucker, C., S. Hubner, U. Visser, et H. G. Weland (2001). "*As Time Goes by"- Using Time Series Based Decision Tree Induction to Analyze the Behaviour of Opponent Players*. RoboCup 2001: Robot Soccer World Cup V, LNAI. 2377 (pp. 325–330). Berlin: Springer-Verlag.

Faloutsos, C., M. Ranganathan, et Y. Manolopoulos (1994). *Fast Subsequence Matching in Time-Series Databases*. ACM SIGMOD Int. Conf. on Management of Data. 419-429

Geurts, P. (2001). Pattern extraction for time series classication. Principles of Data Mining and Knowledge Discovery, LNAI 2168 . pp. 115–127. Berlin: Springer-Verlag.

Kadous, M. (1999). *Learning comprehensible descriptions of multivariate time series*. Proceedings of the Sixteenth International Conference on Machine Learning, pp. 454–463. San Francisco: Morgan Kaufmann.

Keogh, E., et P. Smyt (1997). *A probabilistic approach to fast pattern matching in time series databases*. In Proc third international conference on knowledge discovery and data mining ,California,AAAI Press,Menlo Park,California 24-30.

Le Goc, M., P. Bouché, et N. Giambiasi (2005). *Stochastic Modeling of Continuous Time Discrete Event Sequence for Diagnosis*. 16th International Workshop on Principles of Diagnosis, DX'05, Monterey, California, USA, June 1-3, 2005, p. 133-138.

Murthy, S. K. (1998). Automatic Construction of Decision Trees from Data: A Multi-disciplinary Survey. *Data Mining and Knowledge Discovery,* vol. 2, no. 4, pp. 345-389.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning* 1, 81-106.

Rodriguez J J., et C. J. Alonso (2004). *Interval and Dynamic Time Warping-based Decision Trees*. Symposium on Applied Computing, Proceedings of the 2004 ACM symposium on Applied computing . 548-552

Summary

This article proposes an adaptation of the TemporalID3 algorithm to analyze chronicle models discovered by a stochastic approach where discrete event sequences are alarms generated by a knowledge base system for monitoring and diagnosis of dynamic systems.