Découverte de chroniques à partir de séquences d'événements pour la supervision de processus dynamiques

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

*LSIS, UMR CNRS 6168,
Université Paul Cézanne,
Domaine Universitaire St Jérôme,
13397 Marseille cedex 20, France
{nabil.benayadi,marc.legoc,philippe.bouche}@lsis.org

Résumé. Ce papier adresse le problème de la découverte de connaissances temporelles à partir des données datées, générées par le système de supervision d'un processus de fabrication. Par rapport aux approches existantes qui s'appliquent directement aux données, notre méthode d'extraction des connaissances se base sur un modèle global construit à partir des données. L'approche de modélisation adoptée, dite stochastique, considère les données datées comme une séquence d'occurrences de classes d'événements discrets. Cette séquence est représentée sous les formes duales d'une chaîne de Markov homogène et d'une superposition de processus de Poisson. L'algorithme proposé, appelé BJT4R, permet d'identifier les motifs séquentiels, les plus probables entre deux classes d'événements discrets et les représentent sous la forme de modèles de chroniques. Ce papier présente les premiers résultats de l'application de cet algorithme sur des données générées par un processus de fabrication de semi-conducteur d'un site de production du groupe STMicroelectronics[1].

1 Introduction

Le problème de la découverte des modèles temporels caractérisant le comportement des systèmes dynamiques est un enjeu majeur pour les tâches de contrôle et de surveillance. La raison de base réside dans la difficulté des experts humains d'apprendre et de formuler leurs connaissances sur la dynamique de ces processus. La surveillance est effectuée à partir d'un ensemble d'observations (séquences d'occurrences d'événements discret) produites par le système de pilotage. Les séquences d'observations remontées par le système de supervision sont porteuses de connaissances temporelles sur les relations causales entre les différentes variables du processus.

Notre approche est centrée sur la découverte des séquences particulières d'événements signe d'un comportement particulier. Nous proposons de représenter le comportement du systèmes sous la forme de chroniques (un formalisme graphique pour la représentation des motifs

[1]Ce papier a été effectué sous l'aide financière de la Communauté du Pays d'Aix, de conseil général de Bouches du Rhône, conseil régional de Provence Alpes Côte d'Azur et du STMicroelectronics. Zone Industrielle de Rousset 13106 ROUSSET cedex, France.

temporels où les noeuds sont les classes d'événements et les arcs représentent les contraintes temporelles liant les classes d'événements) (Dousson et Duong (1999), Ghallab (1996)). Ce choix de représentation s'est révélé particulièrement adapté à la représentation des évolutions de systèmes dynamiques, tout en maintenant une complexité raisonnable pour le traitement en temps réel des occurrences d'événements pour la supervision. Notre méthode de découverte des chroniques se déroule en deux phases : la première phase consiste à modéliser la séquence d'événements discrets en intégrant l'approche stochastique proposée par Le Goc et Bouché (2005). Cette approche est basée sur la représentation d'une séquence d'événements discrets sous les formes duales d'une chaîne de Markov homogène et une superposition de processus de Poisson. Dans la deuxième phase, nous proposons un algorithme, appelé BJT4R (*Backward Jump with Timed constraints For Roads*), destiné à la découverte des chroniques à partir du modèle obtenu durant la première phase. L'algorithme BJT4R est une extension de l'algorithme *Viterbi* (Viterbi (1967)) aux chaînes de Markov, basé sur l'application de la relation Chapmann-Kolmogorov comme fonctionnelle de coût.

La section suivante présente les principales approches de découverte des motifs séquentiels à partir de données datées. La section 3 introduit l'approche de modélisation adoptée, dans la section 4 nous présentons l'algorithme BJT4R. Les résultats préliminaires pour la découverte des processus de fabrication des wafers sont présentés dans la section 5. La conclusion de ce papier évoque les prochaines étapes de travail.

2 Contexte

La problématique générale est la suivante : étant donné un ensemble de comportements particuliers ou ordinaires, observés dans une série d'expériences, quelles sont les modèles temporels qui caractérisent au mieux ces comportements ? Des questions similaires ont été traitées dans le domaine de Fouille de Données Temporel.

Introduits pour la première fois dans Agrawal et Srikant (1995), les motifs séquentiels peuvent être vus comme une extension de la notion des règles d'associations (Agrawal et al. (1993)), intégrant la notion de temps. Dans Agrawal et Srikant (1995), les auteurs proposent une approche permettant la découverte des motifs séquentiels à partir de bases de données contenant des séquences de transaction d'achat effectuées par des clients. Une séquence est constituée de plusieurs transactions, réalisées par un client. Une transaction est caractérisée par un identifiant, une date de transaction et l'ensemble des produit achetés, appelés ItemSet. Le problème de la découverte de motifs séquentiels consiste à rechercher l'ensemble des séquences ayant des supports supérieurs à un certain seuil minimal. Un ensemble d'algorithmes ont été tirés de cette approche fréquentielle : *AppioriAll*, *ApprioriSome* et *DynamicSome*. Ces algorithmes de recherche des motifs séquentiels présentent quelques limites concernant la prise en compte des contraintes temporelles. Ce problème a conduit à la recherche des séquences généralisées définies dans (Srikant et Agrawal (1996)). Cette technique de recherche permet d'obtenir des motifs séquentiels respectant certaines contraintes temporelles définies par l'utilisateur (par exemple, regroupement des achats lorsque leurs dates sont assez proches, considération des Itemsets (achats) comme trop rapprochés pour apparaître dans le même motif fréquent).

Dans Mannila et al. (1997), les auteurs abordent le problème de la découverte des motifs temporels, appelés épisodes, fréquemment contenus dans une séquence d'occurrence d'événe-

ments. Dans cette approche, un événement est le couple (A,t), où $A \in E$ est un type d'événement appartenant à l'ensemble des types d'événements E, et t représente la date d'occurrence de type d'événement A. Une séquence d'événements dans E est un triplet (s,T_s,T_e) où $s = \langle (A_1,t_1),(A_2,t_2),\ldots,(A_n,t_n)\rangle$ est une suite ordonnée d'événements $A_i \in E$ pour tout $i = 0,\ldots,n-1$, $t_i \leq t_{i+1}$ et T_s et T_e représentent la date de début et de fin de la séquence tel que :$T_s \leq t_i \leq t_{i+1} \leq T_e$ pour tout $i < n$. Un épisode est une collection d'événements apparaissant relativement proches entre eux dans un ordre partiel. Deux types d'épisode ont été définis : les épisodes parallèles où l'ordre des occurrences des événements dans les épisodes n'est pas porteur de sens, et les épisodes séquentiels ou les occurrences des événements sont complètement ordonnées par leurs dates d'occurrence. La méthode de découverte des épisodes fréquents est basée sur la notion de *fenêtre temporelle* définie à travers une séquence (s,T_s,T_e). Une fenêtre temporelle est un triplet (w,t_s,t_e) où $t_s \leq T_e$ et $t_e \geq T_s$, et w contient des couples (A,t) de la séquence s ou $t_s \leq t \leq t_e$. La durée $t_e - t_s$ est appelée la largeur de la fenêtre temporelle. Étant donnée une fenêtre temporelle et un épisode α, la fréquence de l'épisode α est le rapport du nombre des fenêtres où α apparaît sur le nombre total des fenêtres dans la séquence. Le problème de la découverte des motifs temporels consiste à extraire les épisodes ayant une fréquence d'apparition supérieure à un certain seuil minimal (Algorithme *Winapi*). Par contre, la taille des épisodes ainsi découverts est liée à la largeur de la fenêtre temporelle. Pour éviter ce problème, une notion d'occurrence minimale a été introduite par Mannila et al. (1997)(Algorithme *Minepi*). Étant donné un épisode α, l'intervalle $[t_s,t_e]$ est une occurrence minimale de α dans la séquence (s,T_s,T_e) si α apparaît dans $[t_s,t_e]$ et α n'appartient à aucun sous intervalle de $[t_s,t_e]$. Lorsque les épisodes ont été découverts, des règles sont déduites afin de décrire ou de prédire toute ou partie d'une séquence. Cette méthode proposée traite le temps d'une manière implicite : la seule contrainte temporelle qu'elle autorise est une borne maximale (la largeur de la fenêtre temporelle) sur la durée des épisodes, celle-ci devant être fixée par l'utilisateur. Par contre, les contraintes temporelles liant les éléments des épisodes ainsi découverts sont ignorées. Ghallab (1996) propose une méthode permettant la découverte des modèles de chroniques à partir d'un ensemble de séquences d'alarmes, divisées en deux sous ensembles : un ensemble de séquences positives (exemples) et un ensemble de séquences négatives (contre exemples). L'idée générale de la méthode est de considérer dans un premier temps les séquences d'événements comme des suites ordonnées, sans tenir compte l'aspect temporel. Ceci revient à conserver l'ordonnancement des événements, et à supprimer toute notion de temps. Une fois les séquences dépourvues des écarts temporels, on détermine les chroniques les plus longues qui sont communes à tous les exemples, et qui ne sont pas reconnues par les contres exemples. Pour les motifs séquentiels découverts, les contraintes temporelles entre les éléments de modèles peuvent être déterminées par les experts ou par le calcul des écarts minimaux et maximaux entre chaque couple de deux alarmes afin d'englober toutes les occurrences de ces deux types d'alarmes. Une autre approche proposée par Dousson et Duong (1999) permettant la découverte des modèles de chroniques à partir d'une séquence d'alarmes, appelée journal d'alarmes. Cette approche repose sur une analyse fréquentielle de la séquence d'alarmes, visant à identifier les formes temporelles récurrentes. Cette méthode est une extension de l'approche proposée par Mannila et al. (1997) (Algorithme Minepi) par l'introduction des contraintes temporelles entre les alarmes.

Dans un article plus récent, Mannila (2002) dresse un bilan des principales limites des approches présentées et identifie le principal défaut : les relations entre les éléments des modèles

que ces algorithmes permettent de découvrir sont trop locales pour constituer une véritable représentation de la séquence étudiée. Il invite donc à rechercher des algorithmes adoptant un point de vue plus global. Le Goc et Bouché (2005) proposent ainsi une approche stochastique globale permettant la modélisation des séquences d'événements discrets générées par un système à base de connaissances, de surveillance et de diagnostic de processus dynamiques. Cette approche est basée sur la représentation d'une séquence d'événements discrets sous les formes duales d'une chaîne de Markov homogène et d'une superposition de processus de Poisson.

3 Modélisation de séquences -Approche Stochastique-

Soit $X = \{x_k(t)\}_{k\in\aleph}$ est un ensemble de variables, $x_k(t)$ est une fonction définie sur $\Re$. Dans l'approche stochastique, 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 x est une variable, $i \in I_x \subseteq \aleph$ est une valeur discret de $x(t)$, et $t_k \in \Gamma = \{t_i\}, t_i \in \Re$, est le temps d'affectation de la valeur i à la variable x tel que une occurrence $o_k \equiv (t_k, x, i)$ correspond à l'affectation $x(t) = i$. Les occurrences sont datées selon une horloge à temps continu (i.e $t_{k-1} - t_{k-2} \neq t_k - t_{k-1}$).

$$\begin{array}{l} \forall t_k \in \Re, \forall i \in I_x \in \aleph, \exists t_{k-1} < t_k, \\ x(t_{k-1}) \notin i \wedge x(t_k) \in i \Rightarrow o_k \equiv (t_k, x, i) \end{array} \tag{1}$$

Soit $E = \{e_k\}_{k\in\aleph}$ l'ensemble des événements définit dans $X \times I_x$ et $\Gamma = \{t_i\}_{t_i\in\Re}$ l'ensemble des dates d'occurrences définit dans $\Re$. Nous notons $E_o = \{o_k\}_{k\in\aleph}, o_k \equiv (t_k, x, i)$, l'ensemble des occurrences d'événements discrets défini dans $\Gamma \times E$. Soit d une fonction qui renvoie la date d'occurrence d'un événement :

$$d : E_o \mapsto \Gamma, \forall o_k \equiv (t_k, x, i) \in E_o, d(o_k) = t_k \tag{2}$$

Le couple (o_k, o_{k+1}) de deux occurrences successives liées à une même variable x décrit l'évolution temporelle de la fonction $x(t)$, définie dans $[t_k, t_{k+1}[$:

$$\begin{array}{l} \forall o_k \equiv (t_k, x, i), o_{k+1} \equiv (t_{k+1}, x, j) \in \omega \\ (o_k, o_{k+1}) \Rightarrow \forall t \in [t_k, t_{k+1}[, x(t) = i \wedge x(t_{k+1}) = j \end{array} \tag{3}$$

Par conséquence, une séquence $\omega = \{o_k\}_{k=0,\ldots,m-1}$ d'occurrences d'événements discrets $o_k \equiv (t_k, x, i)$ qui concerne la variable x décrit l'évolution temporelle de la fonction discrète $x(t)$ définie dans $\aleph$.

Une classe d'événement discret est l'ensemble $C^j = \{e_i\}$ d'événements discrets $e_i \equiv (x, i)$. Nous allons utiliser la notation "$e_i :: C^j$" pour noter que l'événement discret e_i appartient à la classe d'événement C^j. Par extension, nous notons "$o_i :: C^j$" une occurrence d'un événements discret appartenant à la classe C^j. La relation binaire $R\left(C^i, C^o, [\tau^-, \tau^+]\right)$ décrit la relation orientée entre deux classes d'événements discrets contraintes temporellement. "$[\tau^-, \tau^+]$" est un intervalle de temps pour observer une occurrence de la classe de sortie C^o après l'occurrence de la classe d'entrée C^i.

$$\begin{array}{l} R(C^i, C^o, [\tau^-, \tau^+]) \Leftrightarrow \exists o_n, o_k \in \omega \subseteq \Omega, \\ (o_n :: C^o) \wedge (o_k :: C^i) \wedge (d(o_n) - d(o_k) \in [\tau^-, \tau^+]) \end{array} \tag{4}$$

Dans ce contexte, un modèle de chroniques est un ensemble de relations binaires temporellement contraintes entre des classes d'événements discrets. Le modèle de chronique $M_{123} =< R_{12}(C^1, C^2, [\tau_{12}^-, \tau_{12}^+]), R_{23}(C^2, C^3, [\tau_{23}^-, \tau_{23}^+]) >$ définit deux relations binaires entre trois classes d'événements discrets vérifiant la relation suivante :

$$\begin{array}{l} \exists o_k, o_n, o_m \in \Omega, (o_k :: C^1) \wedge (o_n :: C^2) \wedge (o_m :: C^3) \wedge \\ (d(o_n) - d(o_k) \in \left[\tau_{12}^-, \tau_{12}^+\right]) \wedge (d(o_m) - d(o_n) \in \left[\tau_{23}^-, \tau_{23}^+\right]) \end{array} \tag{5}$$

La figure 1 montre une représentation graphique de modèle de chronique M_{123} représenté dans le "Langage ELP" (Frydman et al. (2001) ; Le Goc et al. (2006)).

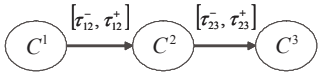

FIG. 1 – *Représentation graphique de* M_{123}

Un modèle de chroniques peut être utilisé dans une tâche de diagnostic pour la prédiction d'une occurrence d'une classe particulière dans une séquence. Comme la classe d'événement C^3 dans le modèle M_{123}.

$$\begin{array}{l} \forall \omega \subseteq \Omega, \forall o_k, o_n \in \omega, \\ (o_k :: C^1) \wedge (o_n :: C^2) \wedge (d(o_n) - d(o_k) \in \left[\tau_{12}^-, \tau_{12}^+\right]) \\ \Rightarrow \exists o_m \in \omega, (o_m :: C^3) \wedge (d(o_m) - d(o_n) \in \left[\tau_{23}^-, \tau_{23}^+\right]) \end{array} \tag{6}$$

L'approche stochastique considère une séquence ω des occurrences de classes d'événements discrets comme une séquence de transition d'états d'un automate stochastique temporel. Lorsque les classes d'événements sont indépendantes, et la distribution des durées inter-occurrences suit une distribution exponentielle $f(t) = 1 - e^{-\lambda t}$, où λ est le taux de poisson (nombre moyen d'occurrences des événements par unité de temps), alors l'automate stochastique peut être représenté sous les formes duales d'une chaîne de Markov homogène et d'une superposition de processus de poisson. Afin de représenter une séquence comme une chaîne de Markov $X = (X(t_k); k \geq 0)$, l'ensemble des classes d'événements discrets $C_\omega = \left\{C^i\right\}_{i=0,\ldots,n-1}$ de la séquence $\omega = (o_k :: C^i)_{k=0\ldots m}$ est confondu avec l'espace d'états $S_M = \{i\}_{i=0,\ldots,n-1}$ de la chaîne de Markov. La relation binaire $\omega' = (o_{k-1} :: C^i, o_k :: C^j)$ dans la séquence ω est correspond à la transition d'états dans la chaîne de Markov $X(d(o_{k-1} :: C^i)) = i \rightarrow X(d(o_k :: C^j)) = j$. Lorsque la chaîne de Markov est homogène, la probabilité de transition d'un état i à la date t_{k-1} vers l'état j à la date t_k dépend uniquement des états i et j :

$$\forall k > 0 \in \aleph, P[X(t_k) = j | X(t_{k-1}) = i] = P[j|i] = p_{ij} \tag{7}$$

Le processus de comptage des transitions d'états dans une chaîne de Markov homogène est un processus de Poisson $(N_{ij}(t); t \geq 0)$ où $N_{ij}(t)$ compte le nombre des transitions $X(t_{k-1}) = i \rightarrow X(t_k) = j$ de l'état i vers l'état j. C'est le nombre de sous séquences $\omega' = (o_{k-1} :: C^i, o_k :: C^j)$ dans la séquence ω. Le processus de Poisson $N_{ij}(t)$ est entièrement défini par l'unique

paramètre λ_{ij}, appelé le taux de Poisson, qui correspond au nombre de transitions $(X(t_{k-1}) = i) \rightarrow (X(t_k) = j)$ par unité de temps. Dans ce cas, la probabilité de transition dans la chaîne de Markov homogène est donnée par :

$$\forall i \neq j \in S_M, p_{ij} = \frac{\lambda_{ij}}{\sum_{j \neq i} \lambda_{ij}} \tag{8}$$

Les contraintes temporelles sont évaluées à partir des délais $d(o_k :: C^j) - d(o_{k-1}C^i)$ où $o_k :: C^j$ est l'occurrence d'un événement de la classe C^j qui suit une occurrence d'un événement de la classe C^i. Celle ci mène à considérer un nouveau type de séquence ω_{i-j} contenant seulement les occurrences de deux classes C^i et C^j dans ω .

Pour évaluer le délai moyen entre les occurrences successives des classes $o_{k-1} :: C^i$ et $o_k :: C^j$ dans ω_{i-j}, deux processus de poisson sont définis (C^i et C^j sont deux classes d'événements mutuellement indépendantes) :

- Un processus de Poisson $(N_{i-j}(t); t \geq 0)$ qui compte le nombre de sous séquences $\omega' = (o_{k-1} :: C^i, o_k :: C^j)$.
- Un processus de Poisson composé $(N^D_{i-j}(t); t \geq 0)$ associé à chaque processus de Poisson $(N_{i-j}(t); t \geq 0)$ pour le calcule des délais entre chaque deux instances successives $(o_{k-1} :: C^i, o_k :: C^j)$:

$$\begin{array}{l} \forall k = 1, \ldots m-1, \forall \omega' \equiv (o_{k-1} :: C^i, o_k :: C^j) \subseteq \omega_{i-j}, \\ N^D_{i-j}(d(o_k :: C^j)) = N^D_{i-j}(d(o_{k-1} :: C^i) + d(o_k :: C^j) - d(o_{k-1} :: C^i) \end{array} \tag{9}$$

Le délai moyen D_{ij} entre deux occurrences de classe C^i et C^j dans ω est donné par :

$$D_{ij} = E\left[d(o_k :: C^j) - d(o_{k-1} :: C^i)\right] = \frac{N^D_{i-j}(t_m)}{N_{i-j}(t_m)} \tag{10}$$

Alors, $n \times (n-1)$ processus de Poisson composés $(N^D_{i-j}(t); t \geq 0)$ sont utilisés pour évaluer les contraintes temporelles entre chaque couple de classes d'événements dans ω. Par définition, D_{ij} est égale à $\frac{1}{\lambda_{ij}}$ où λ_{ij} est l'intensité de la distribution exponentielle des temps inter-occurrences entre les classes C^i et C^j. La contrainte temporelle entre chaque couple de classes (C^i, C^j) est directement déduite à partir de la surface de la loi exponentielle. Dans notre application, les contraintes temporelles sont des intervalles de la forme $[0, 2/\lambda_{ij}]$. Ceci permet de couvrir 59% des occurrences des couples (C^i, C^j).

4 Algorithme BJT4R

L'algorithme BJT4R (*Backward Jump with Timed constraints for Roads*) est un algorithme de découverte des chroniques liant deux classes d'événements dans une séquence d'événements discrets. Le problème peut être formalisé de la manière suivante : étant donnée une séquence d'événements discrets, quelle sont les chroniques les plus probables liant une classe d'événement d'entrée à une classe d'événement de sortie ? La recherche se fait à partir de la classe de sortie en cherchant itérativement les prédécesseurs permettant d'obtenir les chroniques les plus probables (*Backward Jump*). Sa complexité algorithmique est exponentielle aux nombres des classes d'événement discret.

L'algorithme BJT4R, basé sur l'approche stochastique, est une extension de l'algorithme *Viterbi* pour la recherche des chroniques dans un espace d'états Markovien. La relation de Chapmann-Kolmogorov est utilisée pour définir la fonction de coût des chroniques, afin de sélectionner les chroniques les plus probables. Cet algorithme opère en deux étapes :

1. Identification des chroniques sans contraintes temporelles, l'idée est d'identifier un ensemble des motifs séquentiels les plus probables liant une classe d'événement d'entrée à une classe d'événement de sortie. Cette étape est basée sur l'utilisation de la matrice de probabilité de transition construite à partir de la séquence ω et l'utilisation de la relation de Chapmann-Kolmogorov comme fonction de coût dans l'algorithme *Viterbi*.
2. Établissement des contraintes temporelles en utilisant la superposition des processus de Poisson, l'idée est d'utiliser directement les contraintes temporelles estimées à partir des processus de Poisson composés.

4.1 Découverte des chroniques sans contraintes temporelles

Soit $X = (X(t_k)), k \geq 0$, une chaîne de Markov homogène correspondant à la séquence ω. Soit S_M l'ensemble des états de X correspond à l'ensemble des classes d'événements discrets ayant une occurrence dans ω, et $P = [p_{ij}], (i, j) \in S_M^2$, sa matrice de transition des occurrences binaires de sous séquences (C^i, C^j) dans ω . Une chronique liant l'état i_0 à l'état i_k dans X est une suite d'états $\delta(i_0, i_1, \ldots, i_k)$ telle que :

$$P[\delta(i_0, i_1, \ldots, i_k)] = P[X(t_k) = i_k, X(t_{k-1}) = i_{k-1}, \ldots, X(t_0) = i_0] > 0, \tag{11}$$

La longueur d'une chronique $|\delta| = k$ est le nombre de relations binaires contenues dans $\delta(i_0, i_1, \ldots, i_k)$. Une chronique réduit à un seul état a une longueur de 0. Selon la propriété d'absence de mémoire de chaîne de Markov, la probabilité $P[\delta(i_0, i_2, ..., i_k)]$ d'une chronique $\delta(i_0, i_2, \ldots, i_k)$ est donnée par :

$$\begin{aligned} P[\delta(i_0, i_1, \ldots, i_k)] &= P[X(t_k) = i_k, X(t_{k-1}) = i_{k-1}, \ldots, X(t_1) = i_1 | X(t_0) = i_0] \\ &= \prod_{n=1}^{k} (P[X(t_n) = i_n | X(t_{n-1}) = i_{n-1}]) \\ &= \prod_{n=1}^{k} (P[\delta(i_{n-1}, i_n)]) \end{aligned} \tag{12}$$

Comme la chaîne de Markov est homogène, la probabilité $P[\delta(i_0, i_1, \ldots, i_k)]$ de chronique $\delta(i_0, i_1, \ldots, i_k)$ est égale à la probabilité de transition d'un état i_0 vers une autre état i_k en passant par $(k-1)$ états intermédiaires. Elle est donnée par le produit des probabilités des transitions d'un état à l'autre le long du chemin :

$$P[\delta(i_0, i_1, \ldots, i_k)] = \prod_{n=1}^{k} p_{i_{n-1}, i_n} \tag{13}$$

La probabilité p_{ij}^k d'aller d'un état i vers l'état j en k transitions est égale à la somme de toutes les probabilités des chroniques de longueur k liant l'état i à l'état j, donnée par la relation de Chapmann-Komologrov :

$$\begin{aligned} p_{ij}^k &= P[X(t_k) = j | X(t_0) = i] \\ &= \sum_{\forall(i=i_0,i_1,i_2,\ldots,i_{k-1},i_k=j)\in S_M} \prod_{n=1}^{k} (P[X(t_n) = i_n | X(t_{n-1}) = i_{n-1}]) \\ &= \sum_{\forall(i=i_0,i_1,i_2,\ldots,i_{k-1},i_k=j)\in S_M} (P[\delta(i_0, i_1, \ldots, i_k)]) \end{aligned} \tag{14}$$

Si la chaîne de Markov est homogène, on a :

$$p_{ij}^k = \sum_{\forall(i=i_0,i_1,i_2,\ldots,i_{k-1},i_k=j)\in S_M} (\prod_{n=1}^{k} P_{i_{n-1},i_n}) \tag{15}$$

La relation 15 donne la probabilité totale de tous les chroniques possibles de longueur k menant de l'état i vers l'état j, y compris les chroniques de faible probabilité, qui correspondent à des comportements peu fréquents dans le processus modélisé par la chaîne de Markov.

L'objectif de l'algorithme est d'identifier les m chroniques les plus probables liant deux états i et j dans une chaîne de Markov $X = (X(t_k), k \geq 0)$ construite à partir d'une séquence ω.

Déclaration	1.	n (langueur de chemin)
	2.	$\delta(k)$ (chemin de taille k qui se termine par l'état i_k)
	3.	$\Sigma(k)$ (l'ensemble des chemins de taille k)
	4.	$\xi(\delta(k))$ (la probabilité de chemin $\delta(k)$)
Initialisation	5.	$\delta(0) \leftarrow i$
	6.	$\Sigma(0) \leftarrow \delta(0)$
	7.	$\xi(\delta(0)) \leftarrow 0$
Boucles	8.	$\forall$ k = 1,…n-1
	9.	$\forall s \in S_M$-{i, j}
	10.	$\forall \delta(k-1) \in \Sigma(k-1) \wedge s \notin \delta(k-1)$
	11.	$\Sigma(k-1) \leftarrow \Sigma(k-1)$- $\delta(k-1)$
	12.	$\delta(k) \leftarrow \delta(k-1) \cup s$
	13.	$\xi(\delta(k)) \leftarrow \xi(\delta(k-1)) \times b_{i_k,s}$
	14.	si $\xi(\delta(k)) \geq$ *min_seuil* alors $\Sigma(k) \leftarrow \Sigma(k) \cup \delta(k)$
Finalisation	15.	$\forall \delta(n-1) \in \Sigma(n-1)$
	16.	$\Sigma(n-1) \leftarrow \Sigma(n-1)$- $\delta(n-1)$
	17.	$\delta(n) \leftarrow \delta(n-1) \cup i$
	18.	$\xi(\delta(n)) \leftarrow \xi(\delta(n-1)) \times b_{i_{n-1},i}$
	19.	si $\xi(\delta(n)) >$ min alors $\Sigma(n) \leftarrow \Sigma(n) \cup \delta(n)$
Terminer	20.	return $\Sigma(n)$

FIG. 2 – *Génération des chroniques sans contraintes*

L'algorithme (Fig. 2) opère sur un graphe d'états fini dont les transitions sont étiquetées par une fonction de coût. L'algorithme énumère tout les chroniques possibles menant d'un état d'entrée à un état de sortie, et maintient seulement ceux qui maximise la fonction de coût. Notre méthode de recherche est basée sur l'algorithme *Viterbi* comme un algorithme de recherche des chemins, les chemins ayant une probabilité au dessous d'un certain seuil sont éliminées. Pour

chaque profondeur k (ligne 8), l'algorithme cherche un sous ensemble d'états $\{s\} \in S_M$ (ligne 10-14) vérifiant les conditions suivantes :

- Un état est ajouté s'il n'apparaît pas dans les chroniques sous la construction. Cela signifie que l'algorithme interdit l'apparition de plus q'une classe d'événement dans une chronique (ligne 10).
- Un état est ajouté dans la chronique s'il conduit à un nouveau chronique de probabilité supérieure à un seuil minimal (ligne 13-14).

4.2 Établissement des contraintes temporelles

La seconde phase de l'algorithme consiste à calculer les contraintes temporelles entre les classes d'événements des chroniques identifiées. L'idée est d'utiliser directement les contraintes temporelles estimées à partir des processus de Poisson composés déduits de la superposition des processus de Poisson (paragraphe 3, équations 9 et 10).

La contrainte temporelle entre chaque couple de classes (C^i, C^j) est un intervalle de la forme $\left[0, \frac{2}{\lambda_{ij}}\right]$ où λ_{ij} est le taux de Poisson, qui correspond au nombre de transitions $(o_{k-1} :: C^i, o_k :: C^j)$.

L'algorithme BJT4R est un des outils développés au sein du " Laboratoire ELP ", un environnement Java dédié à l'analyse des séquences d'événements discrets (Frydman et al. (2001)). La section suivante présente l'application industrielle de l'algorithme BJT4R pour la découverte des routes de fabrication des wafer dans la société STMicroelectronics.

5 Application

L'application proposée dans cette section concerne les routes de fabrication des wafers dans le site de production de STMicroelectronics. Un wafer est une galette de silicium sur laquelle sont gravées des puces électroniques pour la télécommunication. Le système de supervision de processus de fabrication génère une large quantité d'informations ($\approx$ 10.000 alarmes par jour). Ces informations décrivent les différentes étapes de processus de fabrication sous la forme d'occurrences d'événements discrets correspondant aux débuts et fins de chacun des traitements appliqués sur la plaque. Cette suite de traitements (appelée route) transforme les plaques de silicium en wafers contenant des puces électroniques. Une route de production d'un wafer particulier, décrite par le système de supervision, est donc une séquence d'occurrences d'événements discrets. Deux classes d'événements spécifiques sont utilisées pour décrire le début et la fin d'une route. La durée moyenne d'une route est d'environ 1 mois (hors tests finaux), et compte environ 250 occurrences de recettes. Une occurrence d'événement discret peut être définit en 3 niveaux de granularités :

- Équipement. Une occurrence est alors un couple (e,t) où t est la date de début ou de fin de d'une recette effectuée sur l'équipement e. Ce niveau de granularité est le plus bas. Il permet de définir une route comme une suite d'occurrences d'équipements (il existe 309 types (classes) d'équipements dans le site).
- Opération. Une occurrence est le couple (o,t) où la date t est le début ou la fin d'une recette contenue dans une opération o. A ce niveau de granularité intermédiaire, une route est une suite d'occurrences d'opérations parmi les 1375 types (classes) d'opérations.

– Recette. A ce niveau de granularité le plus élevé, une occurrence est un couple (r,t) où t est la date de début de la recette r. Dans ce cas, une route est une suite d'occurrences de recettes parmi les 5189 types (classes) de recettes.

L'application présentée dans cet article concerne une première approche du problème, nous nous sommes intéressées au niveau de granularité plus bas : le niveau Équipement. L'objectif est de montrer la faisabilité de l'approche aux processus de fabrication exploités sur le site Rousset de la société STMicroelectronics. Dans cette application, un modèle de chroniques est un ensemble de relations binaires les plus probables liant les équipements deux à deux en satisfaisant les contraintes temporelles. Une contrainte temporelle est le temps moyen entre deux traitements successifs effectués sur les équipements de la relation binaire.

Pour appliquer l'approche stochastique, deux conditions doivent être satisfaites :

1. Les occurrences des événements doivent être indépendantes.
2. Le processus de génération des occurrences doit se comporter comme une superposition de processus de Poisson.

Dans notre application, ces conditions sont vérifiées. La première est assurée par la définition du système de supervision (les alarmes sont générées indépendamment les unes des autres). Selon les experts de STMicroelectronics, le taux d'occurrences des événements discrets par jour est globalement stable durant toute la durée de fabrication, sauf incidents exceptionnels. La figure 3 montre six processus de Poisson correspondant à six équipements sélectionnés aléatoirement, cela garantie la deuxième condition.

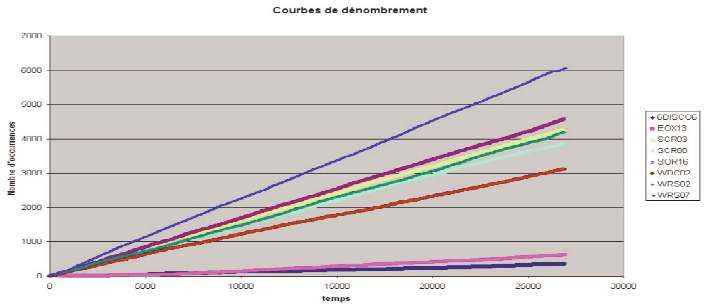

FIG. 3 – *Superposition des processus de Poisson*

La chaîne de Markov associée comprend 309 états(i.e 95481 transitions). La figure 4 présente une partie de la matrice P de probabilité de transitions entre les classes d'événements discrets. La séquence d'événements analysée correspond à une superposition de $309 \times (309-1) = 95172$ processus de Poisson composés. Pour chaque processus de Poisson, le temps inter-occurrences est constant, et correspond à la probabilité maximale de la loi exponentielle.

L'application de l'algorithme BJT4R à la séquence ω produit un graphe (Fig 5) liant la classe d'événement d'entrée à la classe d'événement de sortie. La figure 5 montre partiellement l'ensemble des chroniques les plus probables menant de l'équipement 1130 à l'équipement 1206, la classe d'événement spéciale 0 dénote le début de la route.

L'algorithme BJT4R est paramétré afin de produire les cinq chroniques les plus probables de longueur 15. Par souci de lisibilité, seulement les débuts et les fins des chemins sont montrés dans la figure 5.

	1190	1191	1192	1193	1194	1195	1196	1197	1198	1199	1200	1201	1202	1203	1204	1206	1208	1209
1118	3,05	3,05	7,93	7,32	4,27	2,44	0,00	8,54	20,73	4,27	3,05	3,66	2,44	1,83	1,22	3,66	0,00	0,00
1119	8,58	10,45	7,09	3,73	5,60	1,12	0,00	8,96	5,60	3,73	3,36	4,85	5,97	2,99	0,00	4,10	0,00	0,00
1120	3,40	6,04	9,43	3,40	3,02	0,75	0,00	5,66	6,79	8,68	5,66	3,40	8,30	10,19	0,75	3,77	0,00	0,00
1121	5,02	10,50	7,31	5,48	3,65	2,74	0,00	4,11	8,22	6,85	6,39	5,02	3,20	4,11	1,83	8,68	0,00	0,00
1122	0,73	0,92	1,65	0,37	0,00	0,00	0,00	2,75	2,20	0,18	0,18	1,10	0,37	0,55	0,55	0,18	0,00	0,00
1123	4,83	0,45	1,66	0,76	0,00	0,00	0,00	2,27	1,21	0,60	0,00	0,91	0,00	0,00	0,15	0,60	0,00	0,00
1124	0,00	0,00	0,95	0,38	0,00	0,00	0,00	0,38	0,38	0,19	0,00	0,19	0,00	0,00	0,00	0,57	0,00	0,00
1125	2,22	2,22	4,44	5,19	1,48	5,19	0,00	4,44	2,22	8,15	4,44	4,44	2,96	4,44	1,48	18,52	0,00	0,00
1126	2,94	1,47	4,41	5,15	1,47	0,00	0,00	4,41	2,21	2,21	3,68	5,88	4,41	5,15	0,00	14,71	0,00	0,00
1127	2,89	2,10	6,82	3,15	1,31	0,26	0,00	9,97	12,86	4,46	3,94	2,10	8,40	5,25	0,79	2,10	0,00	0,00
1128	5,98	1,42	5,98	3,70	0,57	0,57	0,00	12,82	11,40	5,41	4,84	2,85	5,98	5,41	0,85	4,84	0,00	0,00
1129	3,87	2,73	8,20	5,92	1,14	1,37	0,00	8,66	6,61	2,28	2,96	4,56	2,96	3,42	0,23	11,85	0,00	0,00
1130	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

FIG. 4 – *Matrice de Probabilité de Transition*

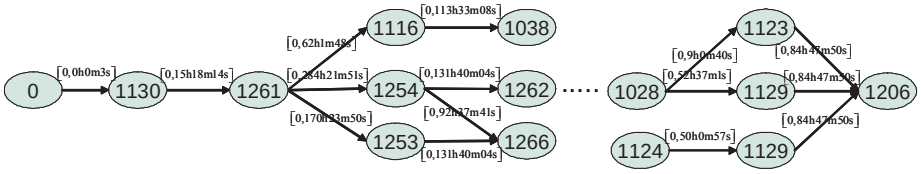

FIG. 5 – *L'ensemble des motifs produits par BJT4R*

6 Discussions et Conclusions

Cet article présente l'algorithme BJT4R pour la découverte des motifs temporels, appelés chroniques, à partir de données datées générées par le système de supervision des processus dynamiques. Dans le domaine de l'extraction des motifs séquentiels, la plupart des algorithmes proposés sont de type Apriori-Like. La complexité reste toujours un obstacle pour ces algorithmes. Notre approche d'extraction des motifs séquentiels se base sur la modélisation de la séquence sous la forme d'une chaîne de Markov. L'estimation des probabilités des transition nécessite une seule passe dans la séquence. la recherche des motifs est faite au niveau de la matrice de probabilité de transitions. L'algorithme de recherche est inspiré de l'algorithme Viterbi en utilisant l'équation de Chapmann-Kolmogorov comme fonction de coût. La méthode de découverte des chroniques est basée sur l'idée qu'une chronique est significative si sa probabilité est supérieure à un certain seuil. Par contre, les contraintes temporelles sont déduites de la superposition des processus de Poisson associée à la chaîne de Markov.

L'application présenté dans cet article montre que l'algorithme est capable d'identifier des chroniques qui modélisent par partie différentes routes de fabrication des galettes de puces électroniques de la société STMicroelectronics.

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 207–216.

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. *Proceedings of the 11th International Conference on Data Engineering (ICDEŠ95)*, 3–14.

Dousson, C. et T. V. Duong (1999). Discovering chronicles with numerical time constraints from alarm logs for monitoring dynamic systems. *In D. Thomas, editor, Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI-99) 1*, 620–626.

Frydman, C., M. L. Goc, N. Giambiasi, et L. Torres (2001). Knowledge-based diagnosis in sachem using devs models. *Special Issue of Transactions of SCS on Recent Advances in DEVS Methodology 18*, 148–159.

Ghallab, M. (1996). On chronicles: Representation, on-line recognition and learning. *Proc. Principles of Knowledge Representation and Reasoning, Aiello, Doyle and Shapiro (Eds.) Morgan-Kauffman,*, 597–606.

Le Goc, M. et P. Bouché (2005). Analyse stochastique de séquences d'événements discrets pour la découverte de signatures. *Revue des Nouvelles Technologies de l'Information - Extraction et gestion des connaissances ISSN :1764-1667 1*, 103–114.

Le Goc, M., P. Bouché, et N. Giambiasi (2006). Devs a formalism to operationnalize chronicle models in the elp laboratory. *in DEVS06, DEVS Integrative M S Symposium, Part of the 2006 Spring Simulation Multiconference (SpringSim06)Van Braun Convention, Huntsville, Alabama, USA 1*, 143–150.

Mannila, H. (2002). Local and global methods in data mining: Basic techniques and open problems. *9th International Colloquium on Automata, Languages and Programming, Malaga, Spain, 2380*, 57–68.

Mannila, H., H. Toivonen, et A. I. Verkamo (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 259–289.

Srikant, R. et R. Agrawal (1996). Knowledge discovery from telecommunication network alarm databases. *Proceedings of the 5th International Conference on Extending Database Technology (EDBTŠ96)*, 3–17.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 260–269.

Summary

This paper addresses the problem of discovering of temporal knowledge from the timed data contained in a monitoring database. Contrary to the existing approaches which apply directly to the data, our method of knowledge discovery is based on a global model built from the data. The modeling approach, said stochastic, considers the timed data as occurrences of discrete event classes. The temporal knowledge are represented under the form of chronicle models (temporal patterns). These models are elaborated from the representation of a a sequence under the form of a homogenous Markov chain and its corresponding superposition of Poisson processes. The BJT4R algorithm identifies the most probable patterns linking two event classes and represents them under the form of a chronicle model. This paper shows the first results of the application of this algorithm to the temporal data generated by the manufacturing process of a production of the STMicroelectronics Company.