Une J-mesure orientée pour élaguer des modèles de chroniques

Nabil Benayadi* and Marc Le Goc*

*LSIS, UMR CNRS 6168, Université Paul Cézanne
Domaine Universitaire St Jérôme
{nabil.benayadi,marc.legoc}@lsis.org

1 Introduction

Les systèmes de supervision de la plupart des applications industrielles génèrent une très grande quantité d'informations et les collectent dans des bases de données. Ce papier concerne la découverte de modèles de chroniques à partir de séquences d'événements. Chaque événement appartient à une certaine classe. Selon l'approche stochastique (Le Goc et al. (2005)), un ensemble de séquences est représenté sous la forme d'une chaîne de Markov afin de l'utiliser par la suite pour générer un modèle de chroniques (Le Goc et al. (2005)) sous forme de relations binaires entre classes d'événements $C^i \mapsto C^o$. Le nombre des relations binaires peut être très grand, par conséquent une réduction de ce nombre est nécessaire. Pour cela, nous proposons une adaptation de la J-Measure de la théorie de l'information aux chaînes de Markov, la BJ-Measure, pour formuler des heuristiques d'élimination d'hypothèses.

2 Élagage d'un modèle de chroniques

Considérant la propriété d'absence de mémoire de la chaîne de Markov, la relation $C^i \mapsto C^o$ entre deux classes C^i et C^o peut être considérée comme l'une des quatre relations entre deux variables aléatoires binaires $y = \{C^i, \neg C^i\}$ et $x = \{C^o, \neg C^o\}$, connectées à travers un canal binaire discret sans mémoire (Shannon (1948)), avec $\neg C^i \equiv C_\omega - \{C^i\}$ et $\neg C^o \equiv C_\omega - \{C^o\}$. Les occurrences de la classe d'événement C^i portent de l'information sur les occurrences de C^o dans la séquence ω si et seulement si $p(C^o|C^i) > p(C^o)$. La relation binaire entre C^i et C^o dépend de l'écart entre $p(C^o)$ et $p(C^o|C^i)$. Nous mesurons cet écart par la formule suivante :

$$BJM(C^i \mapsto C^o) = p(C^o|C^i) \cdot \log_2(\frac{p(C^o|C^i)}{p(C^o)}) + \frac{1}{\|\neg C^o\|} \cdot p(\neg C^o|C^i) \cdot \log_2(\frac{p(\neg C^o|C^i)}{p(\neg C^o)}) \quad (1)$$

Soit $S = \{C^i \mapsto C^o\}$ un ensemble de relations binaires construites à partir de la séquence ω. Selon la propriété d'absence de mémoire de la chaîne de Markov, les relations binaires contenues dans S sont indépendantes. L'ensemble S est vu comme une succession de plusieurs canals binaires de transmissions sans mémoire. La BJ-Measure d'un chemin $M = \{C^i \mapsto$

$C^{i+1}\}_{i=0\ldots n-1}$ est le produit de nombre de relations binaires et la somme des BJ-Measure de chaque relation binaire $C^i \mapsto C^{i+1}$ de M.

$$BJM(M) = n \cdot \sum_{i=0,\ldots,n-1} BJM(C^i \mapsto C^{i+1}) \tag{2}$$

La probabilité $p(M)$ d'un chemin $M = \{C^i \mapsto C^{i+1}\}_{i=0\ldots n-1}$ dans une matrice de probabilités de transitions d'une séquence ω peut être calculée en utilisant la relation de Chapmann-Kolmogorov. L'élagage consiste à trouver un bon compromis entre la probabilité d'un chemin M et sa quantité d'information qui le traverse. Pour cela, nous utilisons l'heuristique $L(M) = p(M) \cdot BJM(M)$. Notre approche a été appliquée sur les séquences générées par le système à base de connaissances SACHEM. Nous nous sommes intéressés à la prédiction

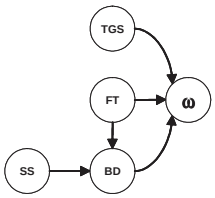

FIG. 1 – *Expertise (1995)*

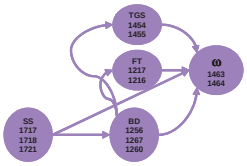

FIG. 2 – *Relations observées en 2007*

des occurrences associées à la variable appelée $omega$. La séquence étudiée contient 7682 occurrences de classes d'événements. Le nombre des relations binaires générées est 3199999. L'application de l'heuristique $L(M)$ permet d'élaguer l'ensemble des relations afin de garder que $195-1$ relations binaires. Grâce à la définition de la notion de classe, nous avons construit un modèle fonctionnel en substituant chacun des identifiants de classe par la variable associée. Le graphe de la figure 2 indique les variables ayant un impact sur la variable $omega$. Ce graphe peut être comparé avec les connaissances *a priori* formulées par les experts en 1995 (cf Figure 1). Le graphe (Figure 1) donné par les connaissances des experts est inclus dans celui donné par l'Approche Stochastique (figure 2) sauf en ce qui concerne le sens de la relation entre les variables FT et BD.

Références

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) Pacific Grove, California, USA*.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal 27*, 379–423.

Summary

In this paper, we propose to adapt the Information Theory J-Measure to Markov chains, the BJ-Measure, to define heuristics to prune the set of binary relations generated by the stochastic approach.