ECAI 2008
M. Ghallab et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-891-5-745

Discovering Temporal Knowledge from a Crisscross of Timed Observations

Nabil Benayadi and **Marc Le Goc** [1]

Abstract. This paper is concerned with the discovering of temporal knowledge from a sequence of timed observations provided by a system monitoring of dynamic process. The discovering process is based on the Stochastic Approach framework where a series of timed observations is represented with a Markov chain. From this representation, a set of timed sequential binary relations between discrete event classes is discovered with an abductive reasoning and represented as abstract chronicle models. To reduce the search space as close as possible to the potential relations between the process variables, we propose to characterize a set of series of timed observations with a unique measure of the homogeneity of the crisscross of class occurrences and to use this measure to prune abstract chronicle models.

1 INTRODUCTION

When supervising and monitoring dynamic processes, a very large amount of timed messages (alarms or simple records) are generated and collected in databases. Mining these databases allows to discover the underlying relations between the variables that govern the dynamic of the process.

This paper addresses this problem in the framework of the Stochastic Approach [2] where a timed message is considered as a timed observation that is represented with an occurrence of a discrete event class $C^i = \{(x_i, \delta_i)\}$ linking a variable x_i and a constant δ_i. The *BJT4T* algorithm represents a set of sequences Ω of discrete event class occurrences with a first order Markov chain and uses an abductive reasoning to identify the set of the most probable timed sequential binary relations between classes. A timed sequential binary relation $R(C^i, C^j, [\tau_{ij}^-, \tau_{ij}^+])$ is an oriented relation $C^i \mapsto C^j$) between two classes C^i and C^j that is timed constrained with the interval $[\tau_{ij}^-, \tau_{ij}^+]$. A set $M = \{(C^i \mapsto C^j)\}$ of timed sequential binary relations constitutes an abstract chronicles model which is used by the *BJT4S* algorithm (*BJT for Signatures*) to look for the n-ary relations in Ω. The search space being generally so large, measures of the "interestingness" of a timed relation are required to focus to the minimal set of hypothesis. To this aim, we defined a measure, called the *BJ-measure*, of the homogeneity of the crisscross (i.e. interlacing) of series of timed observations that is an temporal version of the J-measure of [3].

2 The BJ-Measure

Lets Ω a sequence of $|\Omega|$ occurrences of a set of classes $C^i \in C_\omega$, C^i and C^o two classes in C_ω, $N(C^i)$, $N(C^o)$ and $N(C^i, C^o)$ respectively the occurrence number in Ω of the classes C^i and C^o and of the couple (C^i, C^o). According to the memoryless property of a Markov chain, a timed sequential binary relation $C^i \mapsto C^o$ is associated with a discrete memoryless channel [1] that links the values of two random binary variables $X = \{C^i, \neg C^i\}$ and $Y = \{C^o, \neg C^o\}$, where $\neg C^i \equiv C_\omega - \{C^i\}$ and $\neg C^o \equiv C_\omega - \{C^o\}$ so that: $p(C^i) = \frac{N(C^i)}{|\Omega|}, p(C^o) = \frac{N(C^o)}{|\Omega|}, p(C^o|C^i) = \frac{N(C^i,C^j)}{N(C^i)}$, $p(\neg C^o|C^i) = 1 - p(C^o|C^i)$. The "j" function of the J-measure can be adapted to define a *BJL-measure* that evaluates the homogeneity of the crisscross towards the future (i.e. from the C^i class to the C^o class) and a *BJW-measure* that evaluates the homogeneity of the crisscross towards the past (i.e. from the C^o class to the C^i class). These two measures will then be combined to define the *BJ-Measure* of a timed sequential binary relation $C^i \mapsto C^o$.

Definition 1 *Considering a timed sequential binary relation $C^i \mapsto C^o$ such that $p(C^o|C^i) > p(C^o)$, the BJL$(C^i \mapsto C^o)$ measure is given by the following formula :*

$$\begin{aligned} BJL(C^i \mapsto C^o) &= p(C^o|C^i) \times \log_2\left(\frac{p(C^o|C^i)}{p(C^o)}\right) \\ &+ \frac{(1-p(C^o|C^i))}{|C_\omega|-1} \times \log_2\left(\frac{(1-p(C^o|C^i))}{1-p(C^o)}\right) \end{aligned} \quad (1)$$

where $|C_\omega|$ is the number of event classes in ω.

The *BJL-measure* has the following properties:

- if $p(C^o|C^i) \leq p(C^o)$ then BJL$(C^i \mapsto C^o) = 0$
- if sequence ω consists only of two classes occurrences C^i and C^o,$|C_\omega| = 2$, the BJL$(C^i \mapsto C^o)$ behaves like j-measure.
- for $p(C^o|C^i) = p(C^o)$, BJL$(C^i \mapsto C^o)$ increase when $N(C_\omega)$ increase.
- for $p(C^o|C^i) = 1$, BJL$(C^i \mapsto C^o)$ is maximal $(= \log_2(\frac{1}{p(C^o)}))$.

Definition 2 *Considering a timed sequential binary relation $C^i \mapsto C^o$ such that $p(C^i|C^o) > p(C^i)$, the BJW$(C^i \mapsto C^o)$ measure is given by the following formula :*

$$\begin{aligned} BJW(C^i \mapsto C^o) &= p(C^i|C^o) \times \log_2\left(\frac{p(C^i|C^o)}{p(C^i)}\right) \\ &+ \frac{(1-p(C^i|C^o))}{|C_\omega|-1} \times \log_2\left(\frac{(1-p(C^i|C^o))}{1-p(C^i)}\right) \end{aligned} \quad (2)$$

A noticeable property is that BJW$(C^i \mapsto C^o)$ is null at the same point as BJL$(C^i \mapsto C^o)$. This property of symmetry is a consequence of Bayes' rule: $\frac{p(C^o|C^i)}{p(C^o)} = \frac{p(C^i|C^o)}{p(C^i)}$. The Figure 1 shows the BJW$(C^i \mapsto C^o)$ (abscissa) and the corresponding BJL$(C^i \mapsto C^o)$ (ordinate) for different ratio $\theta = \frac{N_{Ci}}{N_{Co}}$. When the numbers of the occurrences of the classes C^i and C^o are equals (i.e. $\theta = \frac{N_{Ci}}{N_{Co}} = 1$), BJL$(C^i \mapsto C^o)$ = BJW$(C^i \mapsto C^o)$ and the corresponding curve is

[1] LSIS- University AIX-Marseille III France email: {nabil.benayadi, marc.legoc}@lsis.org

the diagonal. The maximum point of the diagonal corresponds to a perfectly homogeneous crisscross of occurrences with $N(C^i,C^o) = N(C^i) = N(C^o)$: each occurrence of the C^i class is followed with an occurrence of the C^o class and each occurrence of the C^o is preceded with an occurrence of the C^o class. The minimum point of the diagonal (i.e. the origin) corresponds to $\text{BJL}(C^i \mapsto C^o) = \text{BJW}(C^i \mapsto C^o) = 0$: the occurrences of the C^i and the C^o classes are not interlaced. It is to note also that the curves of Figure 1 corresponding to θ and θ^{-1} are symmetric according to the diagonal (i.e. $\text{BJL}(C^i \mapsto C^o) = \text{BJW}(C^i \mapsto C^o)$).

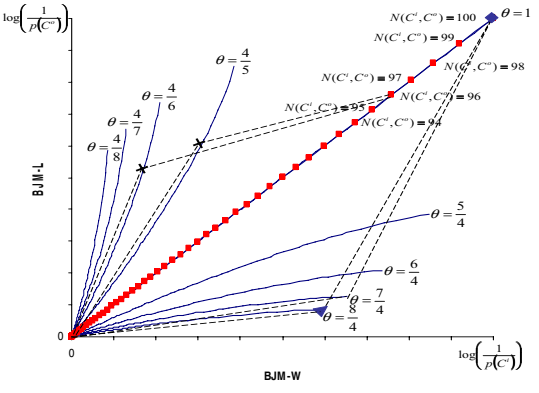

Figure 1: BJL and BJW measures with different ratios $\theta = \frac{N_{C^i}}{N_{C^o}}$

The *BJ-Measure* aims to provide a general mean to evaluate and to represent the homogeneity of the crisscross of any series of classes occurrences.

Definition 3 *The BJ-measure of a timed sequential binary relation* $C^i \mapsto C^o$ *is the norm of the vector* $\begin{pmatrix} BJL(C^i \mapsto C^o) \\ BJW(C^i \mapsto C^o) \end{pmatrix}$ *:*

$$BJM(C^i \mapsto C^o) = \sqrt{BJL(C^i \mapsto C^o)^2 + BJW(C^i \mapsto C^o)^2} \quad (3)$$

The *BJ-measure* depends only of the rates $\theta = \frac{N_{C^i}}{N_{C^o}}$, which made difficult the comparison between two crisscross. This is the aim of the $\alpha(C^i \mapsto C^o)$ function.

Definition 4 *The* $\alpha(C^i \mapsto C^o)$ *function provided the value, projected in the interval* $[0.5, 1]$*, corresponding to the* $BJM(C^i \mapsto C^o)$ *if* $\theta = \frac{N_{C^i}}{N_{C^o}}$ *was equal to* 1.

$$\alpha(C^i \mapsto C^o) = \frac{BJM(C^i \mapsto C^o)}{2 \times max(BJM(C^i \mapsto C^o))} + 0.5 \quad (4)$$

where $max(BJM(C^i \mapsto C^o))$ *is the maximal value of the BJ-measure for a given* θ *(i.e. when* $N(C^i,C^o) = min(N(C^i),N(C^o))$ *for any* N_{C^i} *and* N_{C^o}*).*

The $\alpha(C^i \mapsto C^o)$ is illustrated with the red squares along the diagonal of Figure 1 when $N_{C^i} = N_{C^o} = 100$:

- $\alpha(C^i \mapsto C^o) = 1$ when each of the 100 occurrences of the class C^i is followed by a one and only one of the 100 occurrences of class C^o and inversely (perfect crisscross).
- $\alpha(C^i \mapsto C^o) = 0.99$ when 99 of the 100 occurrences of class C^i is followed by one of the 100 occurrences of class C^o.
- $\alpha(C^i \mapsto C^o) = 0.75$ when 75 of the 100 occurrences of class C^i is followed by one of the 100 occurrences of class C^o.
- $\alpha(C^i \mapsto C^o) = 0.5$ when 50 of the 100 occurrences of class C^i is followed by one of the 100 occurrences of class C^o.

The α function provides then a simple mean to interpret the *BJ-measure* of a crisscross of a series of timed observations.

3 Application to SACHEM system

Our approach has been applied to sequences generated by the SACHEM knowledge-based system developed at the end the 20th century to help the operators to monitor, diagnose and control the blast furnace [2]. We are interested with the *omega* variable that reveals the quality of the management of the whole blast furnace. The studied sequence contains 7682 occurrences of 45 discrete

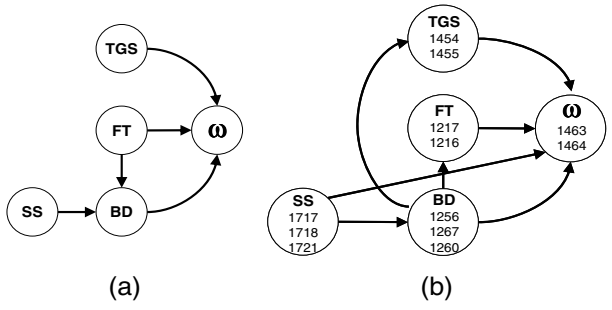

Figure 2: Expert's (1995, a) and discovered relations (2007, b)

event classes of the SACHEM system at Fos-Sur-Mer (France) from 08/01/2001 to 31/12/2001. For the 1463 class linked to the *omega* variable, the *BJT4T* algorithm provides a chronicle model with $20^5 = 3,200,000$ sequential binary relations. Applying the BJ-measure to prune this tree, the *BJT4P* algorithm produces a tree with 195 nodes (the the pruning method is given in [2]). The reduction factor is greater than 16,000 and the pruned tree can then be used by the *BJT4S* algorithm to look for the set of n-ary relations observed in the sequence. When substituting a class withe the corresponding variable, this set becomes the graph (b) of Figure 2. The only difference with the Expert's knowledge formulated in 1995 (graph a) is the direction of the relation between the variables *FT* and *BD*. This result shows that the branches with a high BJ-measure have a strong potentiality to be reveal some knowledge about the relations between the variables of a process.

It is to note that the same result is observed with the Apache system, a clone of Sachem design to monitor and diagnose a galvanization bath.

REFERENCES

[1] C.E.Shannon and W. Weaver, 'The mathematical theory of communication', *University of Illinois Press*, **27**, 379–423, (1949).
[2] M. Le Goc and N. Benayadi, 'Discovering experts knowledge from sequences of discrete event class occurrences', *Proceedings of the 10th International Conference on Enterprise Information Systems (ICEIS08)*, (June 12-16 2008).
[3] P. Smyth and R. M. Goodman, 'An information theoretic approach to rule induction from databases', *IEEE Transactions on Knowledge and Data Engineering 4*, 301–316, (1992).