Aide à la décision pour la maintenance ferroviaire préventive Rapport de recherche

Julien Rabatel[*,**], Sandra Bringay[*,***], Pascal Poncelet[*]

[*]LIRMM, Université Montpellier 2, CNRS
161 rue Ada, 34392 Montpellier Cedex 5, France
[**]Fatronik France Tecnalia, Cap Omega, Rond-point Benjamin Franklin - CS 39521
34960 Montpellier, France
[***]Dpt MIAP, Université Montpellier 3, Route de Mende
34199 Montpellier Cedex 5, France
({rabatel,bringay,poncelet}@lirmm.fr

Résumé. La maintenance de trains est un problème particulièrement délicat lié à de nombreux enjeux à la fois financiers, sécuritaires et énergétiques. Une stratégie de maintenance curative, consistant à restaurer les équipements après l'apparition d'une panne, ne répond pas à ces trois enjeux. De même, une maintenance planifiée, visant à remplacer périodiquement et systématiquement les équipements avant l'émergence d'une panne est extrêmement coûteuse. Pour ces raisons, il est nécessaire de proposer de nouvelles approches afin d'assister les experts dans la mise en place d'opérations de maintenance préventive en détectant tout comportement anormal, susceptible de provoquer un problème majeur (panne, dysfonctionnement, immobilisation du train) dans un futur proche. Pour satisfaire ces besoins, nous proposons un outil d'aide à la décision afin de *(i)* dégager des connaissances utiles pour les utilisateurs sur l'historique des trains, et *(ii)* détecter les anomalies comportementales pouvant être employées pour alerter l'expert. De plus, cet outil apporte de nombreuses informations à l'expert pour analyser les causes possibles ou la gravité des anomalies détectées afin de prendre des décisions optimales en termes de maintenance et de résoudre les incidents.

1 Introduction

Minimiser les coûts liés à la maintenance d'équipements complexes est un enjeu important pour les entreprises du secteur industriel. En effet, les coûts sont continuellement à la hausse. Par exemple, Terry Wireman a montré en 1990 qu'aux Etats-Unis, les frais de maintenance de telles entreprises ont augmenté de 10-15% tous les ans depuis 1979. Plusieurs stratégies de maintenance existent pour répondre à ces besoins. La solution la plus courante est la maintenance curative. Il s'agit de remplacer ou de réparer les équipements après l'apparition d'une panne. Cette solution est particulièrement coûteuse. Par exemple, quelques heures voire minutes d'arrêt sur une chaîne de montage du secteur automobile peuvent se traduire en perte

de plusieurs millions de dollars. Par conséquent, il est généralement beaucoup moins coûteux d'essayer de prévenir ce type de panne. D'autre part, cette maintenance curative pose également des problèmes sur le plan de la sécurité. Par exemple, il est estimé qu'environ 5% des accidents de véhicules motorisés sont inhérents à des dysfonctionnements d'équipement ou à un manque de maintenance[1]. Un dernier aspect problématique de la maintenance curative est lié à environnement et à l'économie d'énergie. Un équipement usé ou sujet à des dysfonctionnements consomme généralement plus d'énergie qu'un équipement neuf. Pour ces diverses raisons, la maintenance curative n'est pas toujours une stratégie appropriée. En outre, une maintenance systématique et planifiée des équipements avant l'émergence d'une panne ne représente pas une solution satisfaisante non plus, car elle engendre des frais très important. En effet, les équipements sont remplacés périodiquement, sans tenir compte de l'état réel du matériel.

Il est donc important de proposer des moyens pour rendre la maintenance à la fois plus rapide et plus efficace en anticipant les pannes et dysfonctionnements. Dans cet article, nous nous intéressons à la maintenance ferroviaire. La maintenance de trains est un problème particulièrement délicat lié à de nombreux enjeux à la fois financiers, sécuritaires et énergétiques. Pour satisfaire ces besoins, nous proposons un outil d'aide à la décision afin de *(i)* dégager des connaissances utiles pour les utilisateurs à propos des trajets de trains passés, et *(ii)* détecter les anomalies comportementales pouvant être employées pour alerter l'expert. De plus, cet outil apporte de nombreuses informations pour permettre à l'expert d'analyser les causes possibles ou la gravité des anomalies détectées, afin de prendre des décisions optimales en termes de maintenance et résoudre les incidents survenus.

La suite de l'article est organisé de la manière suivante. La section 2 discute les travaux existants relatifs à la détection d'anomalies pour la maintenance préventive, ainsi que la nécessité de proposer de nouvelles approches. La section 3 décrit l'approche développée pour extraire des connaissances et détecter des anomalies. La section 4 expose l'outil implémentant cette approche. Enfin, nous concluons dans la section 5.

2 Discussion des travaux existants

La maintenance préventive se décompose en trois étapes (Lee et al. (2004)). L'objectif de la première étape est de caractériser le comportement du système étudié. Cette étape est bien entendu dépendante de l'application et dans la section 3, nous décrivons son application dans le cadre de la maintenance ferroviaire. Puis vient l'étape de traitement de ces données pour l'analyse des comportements suivis. Enfin, la dernière étape est celle de la prise de décisions pour préconiser une politique de maintenance adaptée à l'état de l'équipement. C'est sur cette dernière étape que nous focalisons nos travaux.

Les techniques pour assister la prise de décision dans le domaine de la maintenance préventive conditionnelle peuvent traditionnellement être divisées en deux catégories principales : le *diagnostic* et le *pronostic* (Jardine et al. (2006)). Le diagnostic aborde la détection et l'identification de comportements anormaux, tandis que le pronostic vise à prédire ces comportements avant leur apparition.

[1] http://www.smartmotorist.com

L'établissement de diagnostics s'apparente au problème de la détection d'anomalies dans des données, qui a fait l'objet de nombreux travaux ces dernières années. Défini par Grubbs (1969), une anomalie ou outlier est une observation qui dévie nettement des autres membres de l'échantillon dans lequel elle apparaît. Plus récemment, différentes définitions ont été proposées (Hawkins (1980), Barnett et Lewis (1994), Ramaswamy et al. (2000)) et de nombreuses approches ont été développées pour répondre aux besoins de domaines d'applications divers, tels que le suivi de matériel industriel (Keogh et al. (2006)), mais également la détection d'intrusion (Shaikh et al. (2008), Wang et al. (2008)), le suivi médical (Lin et al. (2005)), la détection de fraudes (Phua et al. (2004)), etc.

La constitution de pronostics est un problème qui a reçu moins d'attention dans les travaux existants. La principale stratégie pour l'établissement de pronostics consiste à prédire le temps qu'il reste avant l'apparition d'une panne. Elle tient généralement compte du comportement actuel de l'équipement et de son âge. Les approches proposées se divisent en trois catégories : les approches statistiques Li et Nilkitsaranont (2009), d'intelligence artificielle Samanta et Nataraj (2008) et celles fondées sur des modèles Heng et al. (2009).

La problématique que nous abordons s'appuie sur l'établissement de diagnostics de comportements anormaux. Néanmoins, nous proposons de repérer ces comportements déviants suffisamment tôt pour anticiper les dysfonctionnements importants. De plus, il est indispensable dans notre cas de proposer une approche qui fournit des résultats faciles à interpréter pour les utilisateurs. En effet, les experts ont peu de connaissances relatives aux outils informatiques utilisés. L'approche proposée doit ainsi permettre d'expliquer pourquoi le système a détecté une anomalie en exposant précisément les raisons qui valident ou au contraire invalident la présence d'un comportement anormal. Cet aspect est difficile à reproduire avec nombre d'approches telles que les approches statistiques ou s'appuyant sur des réseaux de neurones, par exemple.

3 Manipulation et caractérisation des données ferroviaires

Dans cette section, nous abordons le problème du traitement des données issues du suivi des trains via des capteurs embarqués sur les wagons, l'objectif étant de détecter les anomalies. Nous décrivons *(i)* les données issues de capteurs, et *(ii)* la caractérisation des comportements normaux à partir de données historisées.

3.1 Description et manipulation des données

Le système de surveillance des trains utilisé dans nos travaux s'appuie sur une grande quantité de capteurs répartis sur les trains étudiés. Chaque train est composé de 8 bogies[2], à raison de 2 par wagon, et sur chacun d'eux, 32 capteurs mesurent des informations diverses telles que des températures, des accélérations et la vitesse. Toutes les 5 minutes lorsqu'un train est en marche, chaque capteur collecte une valeur qui est alors stockée dans une base de données centralisée. Une description plus complète de ces données est disponible dans Carrascal et al. (2009).

[2]Un bogie est un chariot situé sous un train, sur lequel sont fixés les essieux.

TIME	A	B	C	...
2008/03/27 06: 36: 39	0	16	16	...
2008/03/27 06: 41: 39	82.5	16	16	...
2008/03/27 06: 46: 38	135.6	19	21	...

FIG. 1: Extrait de données brutes.

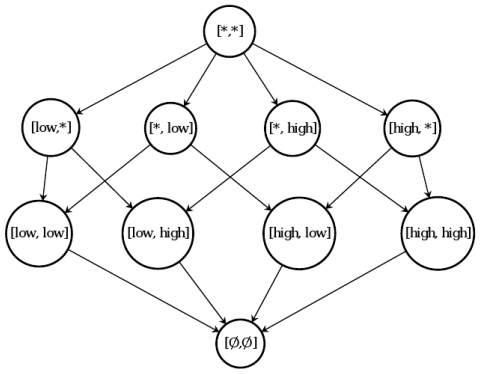

FIG. 2: Treillis de classes.

Les données issues de capteurs sont complexes pour deux raisons : *(i)* des erreurs diverses dégradent fréquemment les données (e.g., du bruit, des valeurs manquantes) *(ii)* des informations de types différents doivent être traitées simultanément (e.g., températures, vitesses, etc.).

Dans les données ferroviaires, les éléments suivants doivent être considérés. Chaque **capteur** décrit une propriété du comportement global d'un train. Les **mesures** sont des valeurs numériques collectées par les capteurs. Notons que ces valeurs sont discrétisées afin d'obtenir un jeu de données mieux adapté à l'étape de fouille de données décrite par la suite. Les **relevés** sont définis comme l'ensemble des valeurs mesurées par tous les capteurs à un moment donné. Un relevé décrit l'état global du train à ce moment. Les données manipulées sont décrites dans la figure 1, où le **relevé** à une date donnée (première colonne du tableau) est décrit par des **mesures** (cellules des autres colonnes) de **capteurs** (première ligne).

Classes de données. Les données que nous manipulons sont associées à des informations contextuelles. Dans le cas de la maintenance ferroviaire, il faut tenir compte de l'influence de paramètres environnants sur le comportement d'un train. Par exemple, un train se comportera différement en fonction des conditions climatiques, ou du type de trajet (e.g., en côte ou en descente). Voyons maintenant comment les paramètres contextuels sont manipulés pour diviser les données en classes selon les différents contextes.

Chaque trajet est décrit dans un ensemble D_C de n dimensions contextuelles (e.g., la température extérieure ou le taux d'humidité ambiant). Considérons la classe c, définie dans D_C. Une classe c est notée $[c_{D_1}, ..., c_{D_i}, ..., c_{D_k}]$, où c_{D_i} est la valeur de c pour la dimension D_i, $D_i \in D_C$. De plus, nous utilisons une *valeur joker*, notée $*$, qui peut substituer toute valeur sur chaque dimension de D_C.

Il est possible de définir une relation de spécialisation ou de généralisation entre les classes. Considérons les dimensions *Température Extérieure* et *Humidité*. La classe $[low, high]$ désigne les données enregistrées avec une température extérieure basse et un taux d'humidité élevé. La classe notée $[high, *]$, plus générale, contient l'ensemble des données enregistrées avec une température extérieure élevée, pour n'importe quel taux d'humidité. La classe $[*, *]$ est la classe plus générale. Elle ne tient pas compte du contexte. La notion d'ordre entre les classes est utilisée pour construire un treillis (cf. figure 2).

3.2 Caractérisation des comportements normaux

L'objectif de la caractérisation des comportements normaux est, à partir d'une base de données de capteurs, de fournir une liste de motifs décrivant au mieux les comportements normaux en répondant à la question : "*Quels motifs apparaissent fréquemment dans les données ?*"

Dans le cadre de nos travaux, nous considérons que les motifs peuvent être représentés sous la forme de motifs séquentiels. Introduits dans Agrawal et Srikant (1995), ils peuvent être considérés comme une extension du concept de règles d'association Agrawal et al. (1993) tenant compte des étiquettes temporelles (ici, les dates de relevé) associées aux items. Le but de la recherche de motifs séquentiels est d'extraire les ensembles d'items qui sont fréquemment associés au cours du temps. Par exemple, dans le cas de données liées à l'achat de produits dans un supermarché, un motif séquentiel est : "*40% des clients achètent une télévision, puis achètent plus tard un lecteur DVD*". Nous adaptons ci-dessous ce problème aux données ferroviaires. Dans les définitions suivants, nous considérerons un ensemble de capteurs, noté $\mathbf{\Omega}$. Un **item** A_v est une paire $\{A, v\}$, où $A \in \mathbf{\Omega}$ et $v \in dom(A)$. Il représente la valeur v collectée par un capteur à un temps donné. L'item A_v est appelé un **A-item**. Soit $\boldsymbol{\omega} = \{A^1, A^2, ...A^i...A^n\}$ un ensemble de capteurs tel que $\boldsymbol{\omega} \subseteq \mathbf{\Omega}$. Un **itemset** I est un relevé ou une partie d'un relevé à un moment donné, i.e., un ensemble non ordonné d'items, noté $I = (A^1_{v_1} A^2_{v_2} ... A^i_{v_i} ... A^n_{v_n})$. Un **séquence** s est une liste d'itemsets notée $s = \langle I_1 I_2 ... I_i ... I_n \rangle$, où l'itemset I_i désigne le $i^{ème}$ itemset de s. Par exemple, les données brutes décrites dans la figure 1 sont traduites par la séquence $\langle (A_0 B_{16} C_{16})(A_{82.5} B_{16} C_{16})(A_{135.6} B_{19} C_{21}) \rangle$, et par $\langle (A_{low} B_{low} C_{low})(A_{avg} B_{low} C_{low})(A_{high} B_{avg} C_{avg}) \rangle$ une fois ces données discrétisées.

Soient deux séquences $s = \langle I_1 I_2 ... I_m \rangle$ et $s' = \langle I'_1 I'_2 ... I'_n \rangle$. S'il existe des entiers $1 \leq i1 < i2 < ... < im \leq n$ tels que $I_1 \subseteq I'_{i_1}$, $I_2 \subseteq I'_{i_2}$, ..., $I_m \subseteq I'_{i_m}$, alors la séquence s est une **sous-séquence** de la séquence s', notée $s \sqsubseteq s'$. On dit que s' *supporte* s. Si une séquence s n'est pas une sous-séquence d'une autre séquence, alors s est dite **maximale**.

La **longueur** de la séquence s, notée $|s|$, est le nombre d'itemsets dans s. De plus, nous notons $s + s'$ la concaténation de séquences.

Le **support** d'une séquence est la fraction de séquences dans une base de séquence DB qui supportent cette séquence. Une séquence est **fréquente** si son support est plus grand ou égal à un seuil $minSupp$, spécifié par l'utilisateur.

Séquences agrégées. Les données manipulées sont redondantes : le comportement décrit par les mesures de capteurs sont généralement stables d'un relevé à un autre. Cette spécificité est perceptible pour les types de données qui changent peu (e.g., la température d'une roue). Aussi, les séquences obtenues sont très grandes avec des itemsets consécutifs contenant des informations généralement identiques. De telles séquences posent deux problèmes : *(i)* la fouille de ces séquences est difficile (la répétition d'itemsets augmente considérablement l'espace de recherche) et *(ii)* elles n'apportent pas d'informations supplémentaires. Nous proposons par conséquent le concept de *séquence agrégée*.

Définition 1 *Soit $s = \langle I_1 I_2 ... I_n \rangle$ une séquence. La **séquence agrégée** correspondante, notée s^*, est la sous-séquence maximale de s respectant la condition suivante :*

$$s^* = \langle I^*_1 I^*_2 ... I^*_i ... I^*_m \rangle, \text{ telle que } \forall I^*_i \in s^*, I^*_i \neq I^*_{i+1}.$$

Le problème de la recherche de motifs séquentiels pour la maintenance ferroviaire est finalement, pour un seuil $minSupp$ et une base de séquences agrégées DB, de trouver toutes les séquences fréquentes agrégées dans DB. De plus, nous souhaitons prendre en compte les différentes informations contextuelles afin de caractériser au mieux les comportements normaux. Pour ce faire, nous recherchons les motifs dans les données associées à chaque classe. Voyons ci-dessous comment sont définis les motifs séquentiels de chaque classe contextuelle.

Définition 2 *Une séquence s est une **séquence c-générale** si s est fréquent dans toutes les classes filles de c. Si c est une classe spécialisée, i.e., dont la seule classe fille est $[\emptyset, \emptyset]$, alors l'ensemble des séquences c-générales dans c est l'ensemble des séquences fréquentes de c.*

Une classe c est associée à l'ensemble des données enregistrées dans le contexte de c, ainsi qu'à l'ensemble des séquences c-générales. Par conséquent, une classe c est maintenant décrite par un triplet $c = \langle \mathcal{D}, \mathcal{E}, \mathcal{S} \rangle$, où :

- $\mathcal{D} = desc(c)$, est la **description** de c dans D_C,
- $\mathcal{E} = data(c)$, est l'**ensemble de données** décrivant les trajets dans le contexte de c[3],
- $\mathcal{S} = seq(c)$, est l'**ensemble des séquences c-générales**.

En exploitant ces définitions, l'ensemble des séquences générales de chaque classe dans le treillis contextuel est construit de la manière suivante :

1. Soit c une classe spécialisée. $seq(c)$ est obtenu par l'extraction de toutes les séquences fréquentes dans $data(c)$.
2. Soit c une classe non-spécialisée et $\check{c}$ l'ensemble de ses classes mères. $seq(c)$ est défini par $seq(c) = \bigcap_{c_i \in \check{c}} seq(c_i)$.

Ainsi, en recherchant les séquences fréquentes dans les classes spécialisées uniquement, nous pouvons construire tous les ensembles de séquences générales.

4 Aide à la décision

A partir des séquences décrites dans la section précédente, nous sommes capables d'extraire des motifs caractéristiques de comportements normaux de trains en tenant compte de la dimension environnement. Dans cette section, nous nous intéressons à l'analyse de nouveaux trains par rapport à la représentation de comportements normaux. Pour cela, nous décrivons un outil permettant d'aider le décideur lorsque des comportements inattendus interviennent. Son objectif est donc *(i)* de restituer les données pour l'utilisateur et *(ii)* d'aider l'utilisateur à prendre les décisions pertinentes de maintenance.

4.1 Restitution des données

En termes de restitution, l'outil donne à l'expert une vue d'ensemble des données puis lui permet d'accéder à des informations plus fines selon ses besoins. Or, fournir une vue synthétique du comportement d'un train s'avère difficile. Aussi, l'outil proposé permet notamment de sélectionner via un formulaire les données relatives à une période ou à un train, et de suivre

[3] Ces trajets sont traduits sous la forme de séquences agrégées.

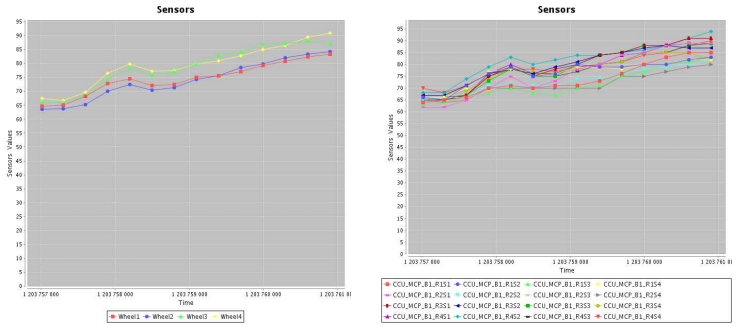

FIG. 3: Visualisation de l'évolution des valeurs de capteurs.

l'évolution des valeurs de capteurs sous forme de graphiques. Pour cela, l'utilisateur sélectionne des groupes de capteurs similaires (e.g., les capteurs de température d'une même roue) et visualise des valeurs agrégées de ces capteurs. Sur la figure 3, l'utilisateur a choisi de visualiser 4 courbes correspondant aux températures moyennes des 4 roues (à gauche), plutôt que les courbes des 16 capteurs (à droite).

4.2 Aide à la décision

Grâce aux courbes présentées dans la section 4.1, un expert peut consulter les données enregistrées. Nous allons maintenant décrire comment notre outil permet de prospecter et d' analyser les données afin de motiver les opérations de maintenance. L'outil alerte les preneurs de décision en cas d'anomalie et associe des informations à ces alertes pour en faciliter l'interprétation. Ainsi, ils pourront répondre à des questions comme : *Quel composant est en cause ? Quand l'anomalie est elle apparue ? S'agit-il réellement d'une anomalie ?*

4.2.1 Détection d'anomalies

Afin d'émettre des alertes lorsqu'un comportement anormal apparaît, nous proposons une méthode de détection d'anomalies adaptée aux motifs séquentiels. Il s'agit d'évaluer, par le biais d'un *score de conformité*, si l'état d'un capteur à un moment donné est conforme ou non au comportement attendu.

Définition 3 *Soit un itemset I et une séquence $s = \langle I_1...I_i...I_n \rangle$. I* ***recouvre*** *s si $\forall I_i \in s$, $I \subseteq I_i$. Soit un motif agrégé $p = \langle I_1^*...I_i^*...I_l^* \rangle$ et une séquence $s = \langle I_1...I_j...I_m \rangle$. p* ***recouvre*** *s, noté $p \prec s$, si il existe un ensemble de séquences $\{s_1, s_2, ..., s_m\}$ tel que : $s = s_1 + s_2 + ... + s_m$, et $\forall i | 1 \leq i \leq m$, I_i^* recouvre I_i. I_i^* est appelé l'itemset correspondant de I_i dans p.*

Définition 4 *Soit $A \in \mathbf{\Omega}$, $s = \langle I_1 I_2 ... I_n \rangle$ une séquence, et $p = \langle I_1^* I_2^* ... I_m^* \rangle$ un motif agrégé. p est un* ***motif*** *(A, i)-**concordant** dans s si (i) il existe des entiers h, j tels que $1 \leq h \leq i \leq j \leq n$, et $p \prec \langle I_h ... I_i ... I_j \rangle$, et (ii) soit I^* l'itemset correspondant de I_i dans p, il existe un item $A_v \in I^*$.*

Définition 5 *Soit $A \in \mathbf{\Omega}$, $s = \langle I_1 ... I_i ... I_n \rangle$ une séquence telle que I_i contient un A-item, noté i_s, et $p = \langle I_1^* I_2^* ... I_j^* ... I_m^* \rangle$ un motif agrégé tel que I_j^* contient un A-item, noté i_p.*
p' est la séquence p où i_p a été remplacé par i_s dans I_j.
*p est un **motif** (A, i)**-discordant** dans s si (1) p n'est pas (A, i)-concordant, et (2) p' est (A, i)-concordant.*

Exemple 1 *Soit $s = \langle (A_{avg} B_{low} C_{low})(A_{high} B_{avg} C_{avg})(A_{high}) \rangle$.*

Le motif agrégé $p = \langle (A_{low} B_{low} C_{low})(B_{avg}) \rangle$ n'est pas (A,1)-concordant. Cependant, le motif $p' = \langle (A_{avg} B_{low} C_{low})(B_{avg}) \rangle$, où A_{low} a été remplacé par A_{avg} dans le premier itemset de p, est (A,1)-concordant. Par conséquent, p est (A,1)-discordant.

Score de concordance. Le calcul d'un score de concordance pour un capteur donné A dans le $i^{ème}$ itemset d'une séquence données s consiste à évaluer à quel point l'état d'un capteur à un moment donné (i.e., l'estampille temporelle associée à l'itemset) est conforme aux motifs qui caractérisent le comportement normal.

Définition 6 *Soit $\mathcal{P}^c$ l'ensemble de tous les motifs (A,i)-concordants. Le **score de concordance** d'un capteur A dans le $i^{ème}$ itemset d'une séquence s est :*

$$score_{conc}(A, i) = \sum_{p \in \mathcal{P}^c} ||p|| \times support(p).$$

Score de discordance. Le score de discordance permet d'estimer à quel point le comportement d'un capteur à un moment donné est déviant par rapport aux motifs séquentiels décrivant le comportement normal.

Définition 7 *Soit $\mathcal{P}^d$ l'ensemble de tous les motifs (A,i)-discordants. Le **score de discordance** d'un capteur A dans le $i^{ème}$ itemset d'une séquence s est :*

$$score_{disc}(A, i) = \sum_{p \in \mathcal{P}^d} (||p|| - 1) \times support(p).$$

Score de conformité. Le score de conformité global basé sur les deux scores précédents et défini entre -1 et 1, doit répondre aux besoins suivants :

- si $score(A, i)$ est proche de 1, l'état de A dans i est considéré comme normal,
- si $score(A, i)$ est proche de -1, l'état de A dans i est considéré comme anormal,
- si $score(A, i)$ est proche de 0, l'état de A dans i est considéré comme incertain.

Définition 8 *Soit $A \in \mathbf{\Omega}$ un capteur, et s une séquence. Le **score de conformité** de A dans le $i^{ème}$ itemset de s, noté $score(A, i)$, est défini par :*

$$score(A, i) = \frac{score_{conc}(A, i) - score_{disc}(A, i)}{max(score_{conc}(A, i), score_{disc}(A, i))}.$$

4.2.2 Visualisation des anomalies

Le score de conformité permet de détecter des anomalies liées à un capteur. Or, l'état d'un composant peut être décrit par plusieurs capteurs. Par exemple, une roue possède 4 capteurs de température. Ces capteurs sont physiquement très proches et relèvent des informations similaires. Le score peut donc être généralisé au niveau du composant. Il faut toutefois distinguer

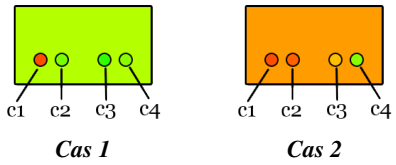

FIG. 4: Anomalie de capteur et anomalie de composant.

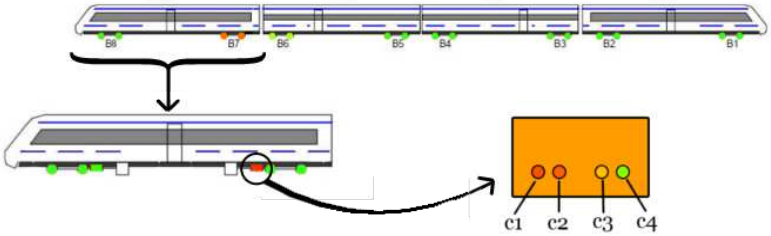

FIG. 5: Différents niveaux de granularité pour la visualisation des anomalies.

deux cas : *(i)* le cas où l'anomalie décelée correspond à un problème de capteur (i.e., le comportement du composant est normal, mais les mesures d'un seul capteur sont corrompues), et *(ii)* le cas plus sérieux où le comportement du capteur est réellement impliqué dans l'anomalie. Finalement, si seulement 1 capteur parmi un groupe de capteurs a un score faible, alors l'anomalie est liée à ce capteur seulement, et non au composant[4].

La figure 4 propose une vue d'un composant muni de 4 capteurs dans deux cas différents. Dans le cas 1, 3 des 4 capteurs (cercles) ont un score élevé (traduit par la couleur verte claire) tandis que le dernier (le capteur *c1*) a un score bas (rouge foncé), témoin d'une anomalie détectée. Cependant, le fait que le score du composant (i.e., la moyenne des scores de tous ses capteurs) est élevée (proche du vert clair) montre qu'il s'agit d'un problème de capteur uniquement, mais que le comportement du composant lui-même n'est pas en cause. En revanche, le cas 2 est différent : le score moyen du composant est bas. Par conséquent, l'anomalie est liée au composant, et non à un capteur défectueux. L'outil rend donc possible l'analyse des données suivant différents niveaux de granularité : au *niveau capteur* en considérant uniquement le score de chaque capteur, au *niveau composant* en interprétant la valeur moyenne des scores de tous les capteurs installés sur celui-ci, et de même au niveau d'un *bogie*, d'un *wagon*, ou d'un *train*.

Nous exploitons ces niveaux de granularité lors de la recherche des anomalies, en proposant à l'utilisateur de naviguer du niveau le plus général (vue globale du train et de ses différents bogies) jusqu'au niveau le plus fin (vue du score de chaque capteur). La figure 5 montre les différents niveaux de granularité et la navigation de l'un à l'autre : le premier schéma fournit les scores de conformité de chaque bogie d'un train. L'utilisateur peut alors choisir un wagon en particulier pour l'observer de manière plus précise et vérifier les scores de conformité de

[4]Cette conclusion est possible car dans l'ensemble de capteurs manipulé, les capteurs sont toujours installés en groupes.

chacun des principaux composants séparément. De même, il est possible de choisir un bogie pour observer précisément un de ses composants et analyser le score global de celui-ci, mais également le score de chacun des capteurs qui le décrivent.

4.2.3 Informations complémentaires : comment exploiter les informations contextuelles

Dans un deuxième temps, nous visons à donner plus d'informations sur la gravité de l'anomalie détectée en tirant parti de la caractérisation contextualisée des comportements (voir Section 3). Nous exploitons le principe suivant : plus le comportement d'un train est peu conforme aux comportements généraux (i.e., son score est bas dans les classes les plus générales) et plus son comportement est problématique. Ainsi, lorsqu'une anomalie est détectée, l'outil calcule le score dans les classes plus générales du treillis de classes. Plusieurs cas sont possibles (voir figure 6) :

- **Cas 1 :** le score de conformité est bas dans la classe spécialisée et irrégulier dans les classes générales. Par exemple, ici le score dans la classe $[low, low]$ est bas, mais les scores dans $[*, low]$ et $[high, low]$ sont élevés. C'est une nouvelle information pour l' expert. Le capteur se comporte comme si la température extérieure était haute, il peut s'agir d'une surchauffe.
- **Cas 2 :** le score de conformité est bas dans la classe spécialisée $[low, low]$ et élevé dans les classes plus générales. Le capteur étudié a un comportement qui ne s'accorde pas avec les comportements spécifiques de sa classe mais correspond aux comportements plus généraux. Le comportement analysé n'est alors pas aussi inattendu que son score dans la classe spécialisée le laisse paraître.
- **Cas 3 :** le score de conformité est bas dans la classe spécialisée $[low, low]$ et bas dans les classes générales. Le comportement du capteur n'est en accord ni avec les comportements spécifiques de sa classe, ni avec ceux plus généraux. Dès lors, l'utilisateur peut juger qu'il s'agit d'un comportement réellement déviant.

Calcul des scores de conformité. La restitution des scores dans les classes parentes d'une classe spécialisée c est rapide car automatiquement déduite des scores calculés au préalable. En effet, lors du calcul du score de conformité de la classe c (décrit dans la section 3), nous stockons le poids individuel de chaque motif concordant ou discordant. Or, par construction, toute séquence fréquente dans une classe parente de c est également fréquente dans c. Par extension, nous déduisons que tout motif concordant (respectivement discordant) d'une classe parente de c est également un motif concordant (respectivement discordant) de c. Ainsi, en stockant le poids accordé à chaque motif lors du calcul du score de c, nous pouvons aisément déduire celui de chacune de ses classes parentes. Le coût lié au calcul des scores dans les classes parentes est alors réduit car nous évitons tout test d'inclusion des séquences associées au score de conformité.

Le calcul des scores de conformité offre une vue à la fois succinte et globale du comportement mesuré par un capteur à un moment donné. De plus, pour répondre aux besoins des experts, nous devons fournir les raisons exactes qui expliquent un score bas. Pour cette raison, l'outil fournit également la liste des motifs discordants et concordants pour l'utilisateur. Ces motifs sont facilement compréhensibles et interprétables et donc particulièrement appréciables pour assister l'analyse de l'expert.

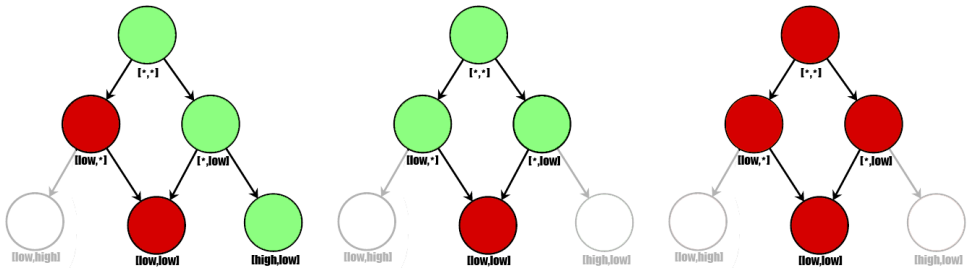

FIG. 6: Propagation des scores de conformité dans la hiérarchie contextuelle.

5 Conclusion

La maintenance de systèmes complexes comme les trains est un problème important pour diverses raisons à la fois financières, sécuritaires ou énergétiques. Une politique de maintenance visant à prévenir les problèmes les plus sérieux (pannes, immobilisation des trains, etc.) par la mise en œuvre d'opérations préventives de maintenance est nécessaire. Cependant, il s'agit d'un problème difficile. Tout d'abord, la maitenance préventive requiert un système de suivi du comportement des trains s'appuyant sur une grande quantité de capteurs rassemblant au cours du temps diverses informations : températures, vitesses, accélérations, etc. Afin de proposer une approche automatique d'émission d'alertes lorsque une anomalie comportementale est repérée (i.e., lorsqu'un comportement dévie du comportement attendu), ces données sont traitées en tenant compte de leurs spécificités (e.g., données temporelles, multi-sources, etc.). Nous proposons une méthode de détection d'anomalies dans les données comportementales, exploitant les motifs séquentiels préalablement extraits pour décrire le comportement normal d'un train. De plus, l'outil présenté ne se limite pas à l'émission d'alertes. Il offre aussi aux experts les informations utiles pour procéder au diagnostic des anomalies détectées, ainsi que pour assister la prise des décisions nécessaires en termes de maintenance. Les travaux présentés ouvrent de nombreuses perspectives. En premier lieu, compléter l'approche proposée en y ajoutant un aspect prédictif constitue une piste intéressante. De plus, nous pouvons en ajustant les méthodes utilisées, adapter l'approche à des données traitées en temps réel.

Références

Agrawal, R., T. Imieliński, et A. Swami (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec. 22*(2).

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In P. S. Yu et A. S. P. Chen (Eds.), *Eleventh International Conference on Data Engineering*. IEEE Computer Society Press.

Barnett, V. et T. Lewis (1994). *Outliers in Statistical Data*. Wiley Series in Probability & Statistics.

Carrascal, A., A. Díez, et A. Azpeitia (2009). Unsupervised methods for anomalies detection through intelligent monitoring systems. In *HAIS '09 : Proceedings of the 4th International Conference on Hybrid Artificial Intelligence Systems*, Berlin, Heidelberg. Springer-Verlag.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics 11*.

Hawkins, D. M. (1980). *Identification of Outliers*. Chapman and Hall.

Heng, A., S. Zhang, A. C. Tan, et J. Mathew (2009). Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical Systems and Signal Processing 23*(3).

Jardine, A. K. S., D. Lin, et D. Banjevic (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing 20*.

Keogh, E., J. Lin, S.-H. Lee, et H. Van Herle (2006). Finding the most unusual time series subsequence: algorithms and applications. *Knowl. Inf. Syst. 11*(1).

Lee, J., R. Abujamra, A. Jardine, D. Lin, et D. Banjevic (2004). An integrated platform for diagnostics, prognostics and maintenance optimization. *Proceedings of the Intelligent Maintenance Systems*.

Li, Y. et P. Nilkitsaranont (2009). Gas turbine performance prognostic for condition-based maintenance. *Applied Energy 86*(10).

Lin, J., E. Keogh, A. Fu, et H. Van Herle (2005). Approximations to magic: Finding unusual medical time series. In *CBMS '05: Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, Washington, DC, USA. IEEE Computer Society.

Phua, C., D. Alahakoon, et V. Lee (2004). Minority report in fraud detection: classification of skewed data. *SIGKDD Explor. Newsl. 6*(1).

Ramaswamy, S., R. Rastogi, et K. Shim (2000). Efficient algorithms for mining outliers from large data sets. In *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, New York, NY, USA. ACM.

Samanta, B. et C. Nataraj (2008). Prognostics of machine condition using soft computing. *Robotics and Computer-Integrated Manufacturing 24*(6). FAIM 2007, 17th International Conference on Flexible Automation and Intelligent Manufacturing.

Shaikh, R. A., H. Jameel, B. J. d'Auriol, S. Lee, Y.-J. Song, et H. Lee (2008). Trusting anomaly and intrusion claims for cooperative distributed intrusion detection schemes of wireless sensor networks. In *ICYCS '08: Proceedings of the 2008 The 9th International Conference for Young Computer Scientists*, Washington, DC, USA. IEEE Computer Society.

Wang, W., X. Guan, et X. Zhang (2008). Processing of massive audit data streams for real-time anomaly intrusion detection. *Comput. Commun. 31*(1).

Summary

Train maintenance is a difficult problem associated with many security, financial and energy resources challenges. Curative maintenance, consisting in restoring the equipment after the occurrence of a failure, do not meet these three challenges. Similarly, planned maintenance, i.e., to periodically and systematically replace equipment before the emergence of a failure is extremely costly. For these reasons, it is necessary to propose new approaches to assist experts in the development of preventive maintenance operations by detecting abnormal behaviors, which can cause a major problem (failure, malfunction, immobilization of the train) in the near future. To meet these needs, we offer a decision support tool to *(i)* extract useful knowledge for users about past train journeys and *(ii)* detect anomalous behavior to alert the expert. Moreover, this tool provides numerous information to allow the expert to analyze the possible causes and the seriousness of detected anomalies, in order to make optimal decisions in terms of maintenance and resolve incidents.