Découverte de règles d'association pour l'aide à la prévision des accidents maritimes

Bilal Idiri*, Aldo Napoli*

*Mines ParisTech, CRC
Rue Claude Daunesse, 06904 Sophia Antipolis, France
Prénom.Nom@mines-paristech.fr

Résumé. Les systèmes de surveillance maritime permettent la récupération et la fusion des informations sur les navires (position, vitesse, etc.) à des fins de suivi du trafic maritime sur un dispositif d'affichage. Aujourd'hui, l'identification des risques à partir de ces systèmes est difficilement automatisable compte-tenu de l'expertise à formaliser, du nombre important de navires et de la multiplicité des risques (collision, échouement, etc). De plus, le remplacement périodique des opérateurs de surveillance complique la reconnaissance d'événements anormaux qui sont éparses et parcellaires dans le temps et l'espace. Dans l'objectif de faire évoluer ces systèmes de surveillance maritime, nous proposons dans cet article, une approche originale fondée sur le data mining pour l'extraction de motifs fréquents. Cette approche se focalise sur des règles de prévision et de ciblage pour l'identification automatique des situations induisant ou constituant le cadre des accidents maritimes.

1 Introduction

L'activité maritime est un secteur important alliant intérêts publics et privés. Elle compte à elle seule 90% des échanges internationaux avec 80% du transport d'énergie (CNUCED, 2009). Pour protéger ce secteur, plusieurs dispositifs de sécurité ont été mis en place comme le développement de systèmes de surveillance maritime : SpatioNav en France, SIVE en Espagne, MEVAT en Finlande (Morel, 2009). Ces systèmes de surveillance affichent les pistes de navires additionnées à d'autres informations complémentaires (cargaison, vitesse, cap, port de départ, etc.) sur une carte numérique pour permettre aux agents d'état la surveillance du trafic maritime. Vu le nombre important de navires à surveiller, l'immensité des territoires maritimes, la multiplicité des risques et l'organisation de la criminalité en mer, les systèmes de surveillance maritime sont insuffisamment adaptés pour l'aide à l'identification des risques maritimes et doivent évoluer pour faire face à ces nouveaux défis. Parmi les travaux traitant de ces questionnements de surveillance dite "*intelligente*" ou de "*nouvelle génération*", nous pouvons citer le projet PANDA (Darpa, 2005) du ministère de la défense américain, considéré comme le projet initiateur qui a inspiré de nombreux travaux. Nous pouvons également citer, le projet SCANMARIS (Morel et al., 2008) qui a évalué les algorithmes de détection automatique des comportements anormaux et TAMARIS (Morel et al., 2011) qui propose une couche fonc-

tionnelle pour l'authentification des comportements suspects à partir d'un ensemble d'alertes générées par SCANMARIS.

Dans le but d'améliorer la surveillance maritime, nous proposons la mise en place d'un système d'aide à l'identification des comportements anormaux de navires et la découverte a priori des situations à risques en utilisant une approche originale basée sur la fouille de données spatiale (Agrawal et al., 1993) (Srikant et Agrawal, 1995) (Zhenhui et al., 2010) (Lee et al., 2008) (Cao et al., 2007) (Marven et al., 2007). L'idée est d'explorer les historiques de données de déplacement de navires et d'accidentologie dans le but de découvrir les connaissances régissant la survenue des risques et identifiant les comportements anormaux des navires. Ces connaissances vont servir à l'identification automatique des risques maritimes à partir de bases de faits (flux de déplacement, météorologie, océanographie, carte de navigation).

Dans cet article, nous nous intéressons à l'extraction de motifs fréquents à partir d'une base de données d'accidents de navires dans le but de découvrir des règles [1] d'expertise pour l'aide à la détection de situations à risque (collision, échouement, avarie, etc.). Un exemple de règles d'association est "*Les accidents de navires britanniques de type Roll-on/Roll-off* [2] *sont dans 76% des cas localisés dans les ports ou à proximité des ports*"

2 Extraction des règles d'association pour la prévention des risques maritimes

L'extraction de règles d'association est un problème non supervisé de data mining qui permet à partir des itemsets [3] apparaissant fréquemment ensemble dans une base de données d'extraire des règles de connaissance. Ce problème a été proposé pour la première fois par Agrawal (Agrawal et al., 1993) pour l'analyse du panier de la ménagère dans le but d'améliorer les ventes. La découverte de règles d'associations dans une base de données de transactions (les paniers) consiste à chercher les produits (itemset fréquemment achetés ensemble). Nous allons voir par la suite, comment extraire ces règles dans un historique d'accidents de navires dans le but de construire une base de connaissances utile à l'identification automatique des risques maritimes.

L'extraction de règles d'association ou le data mining en général fait partie intégrante d'un processus d'extraction de connaissance de données (ECD). Un ECD regroupe l'ensemble des méthodes et des outils qui vont nous permettre de transformer les données volumineuses et hétérogènes des accidents maritimes en connaissances utiles à la prise de décision.

Nous allons détailler par la suite chaque étape du processus d'extraction de connaissances appliqué à l'historique de données d'accidents.

2.1 Sélection des données

Nous avons à notre disposition, une base de données [4] recensant les accidents qui ont affecté ou se sont produits à bord de navires entre *1991* et *2009*. Ces données concernent les

1. Une règle est l'unité composant la connaissance. Elle est de de la forme $A -> B$, tels que A est appelé *antécédent* de la règle et B est appelé *conséquent*. L'intersection entre A et B est vide.
2. Un navire roulier utilisé pour transporter des véhicules grâce à une ou plusieurs rampes d'accès
3. Un itemset est un ensemble d'items, et un item est une occurence d'un objet de la base
4. Fournie à titre gracieux par le bureau d'enquête britannique sur les accidents maritimes (MAIB).

navires britanniques se trouvant n'importe où dans le monde et les navires d'autres nationalités se trouvant dans les eaux territoriales britanniques au moment de leur accident. La base de données contient *14900* cas d'accidents et d'incidents qui concernent *16230* navires.

Nous avons sélectionné dans cette base, les données qui décrivent les accidents (type d'accident, position, temps, etc.), les caractéristiques des navires (identifiant IMO, type du navire, âge du navire, longueur, etc.) et la description de l'environnement (visibilité, état de la mer, force du vent, etc.). Cette sélection de données va constituer le contexte d'exploration sur lequel va porter l'extraction de règles d'association dans le but de trouver les relations entres les différents facteurs de situations. La sélection des attributs sur lesquels va porter notre analyse va réduire le nombre de variables à considérer, le nombre de règles générées et ainsi faciliter l'interprétation des résultats.

Dans la suite de cet article, nous allons voir comment mettre les données brutes en une forme exploitable par les algorithmes d'extraction de règles d'association.

2.2 Préparation des données

Avant toute exploration de données par les méthodes de data mining, une étape de préparation de ces données est nécessaire pour permettre leur exploitation. La préparation des données est difficile et demande plusieurs itérations compte-tenu de son lein fort avec la qualité des résultats. En effet, la quantité et la qualité des données ont un impact direct et significatif sur la qualité des règles obtenues. Nous nous proposons d'étudier dans cette section la distribution des variables pour identifier les anomalies (données manquantes, incohérence, imprécision, etc.), les corriger et préparer le contexte d'exploration.

1. **Population d'accidents non représentative**

   Nous avons remarqué que notre base de données n'était pas représentative de la population globale d'accidents mondiaux de navires car 82% des accidents concernent des navires du Royaume-Uni. Pour avoir une population d'accidents représentative, nous avons réduit notre étude aux accidents de navires britanniques.

2. **Données manquantes**

   Dans le but d'améliorer la qualité des résultats, nous avons envisagé plusieurs approches pour nettoyer ces données : la pondération par des moyennes, des médianes ; la prédiction des valeurs manquantes (Jami et al., 2005) ; etc..

3. **Variables continues et regroupement de classes**

   Les algorithmes d'extraction de règles d'association ne prennent pas en considération les variables (attributs) continues dans leur processus d'extraction. Pour ne pas perdre d'informations en entrée de ces algorithmes, nous avons discrétisé les variables continues en les séparant en classes (intervalles). Dans notre cas, nous avons choisi le critère d'effectifs égaux pour éviter de biaiser les résultats des algorithmes d'extraction de règles d'association. Ces algorithmes sont basés sur la découverte d'itemsets fréquents. Une classe contenant plus d'effectifs a donc plus de chance d'apparaitre dans les règles en sortie.

   Nous avons comptabilisé des variables discrètes ayant une distribution hétérogène de leur effectif. Cette hétérogénéité révèle les classes ayant les plus grandes fréquences d'apparition dans les règles d'association et peuvent empêcher l'apparition des fameuses

pépites d'or. Les navires de pêche par exemple apparaissaient presque systématiquement dans les règles d'association car ils représentent 66% de l'effectif total. Pour faire apparaitre les autres catégories de navires, nous avons regroupé toutes les catégories de navires en trois grandes classes :

- Classe Transport : Avec un effectif de 32%, elle regroupe tous les navires de transports de personnes, d'hydrocarbures (Tanker) et de marchandises,
- Classe Plaisance : La classe des navires de plaisance représente un effectif de 1.2%,
- Classe Pêche : Les navires de pêche représentent un effectif de 66%.

4. **Données aberrantes**

   Les données aberrantes sont des données erronées. Leur identification demande d'avoir une bonne connaissance du domaine étudié. La répartition des accidents sur une carte numérique nous a permis d'identifier et d'écarter les positions aberrantes localisées sur terre, loin des zones de navigation.

   Ces erreurs de positionnement des accidents sont peut être dues aux dysfonctionnements de GPS ou à une mauvaise saisie des coordonnées. L'analyse des valeurs extrêmes (les valeurs maximales et minimales) nous a permis aussi de détecter un ensemble de valeurs erronées comme le cas de valeurs négatives de la variable Age-of-vessel et d'autres encore.

2.3 Extraction des règles d'associations

Nous avons appliqué l'algorithme Apriori, implémenté par Christian Borgelt et se trouvant dans le package Rattle 2.6.4 de R, sur les données préparées dans l'étape précédente (voir section 2.2). Nous avons fait varier les seuils du support [5] "*minsupp*", de la confiance [6] "*minconf*" et nous avons extrait plusieurs fichiers de règles d'association. En plus des mesures support-confiance, nous utilisons aussi, une autre mesure appelée Lift pour vérifier que les résultats obtenus ne sont pas le fruit du hasard. Si la mesure est supérieure à un, la règle est considérée comme intéressante.

Nous avons défini trois grandes catégories à partir des règles découvertes pour les regrouper et faciliter leur exploitation :

- *Règles de prédiction* : Nous appelons règle de prédiction toute règle ayant son antécédent connu a priori, son conséquent non connu et la confiance de la règle est supérieure à 50%. Une règle de prédiction peut être du genre "*Si nous avons un contexte Ci alors à c% il implique un accident de type Ti* ",
- *Règles de ciblage* : Ce sont les règles de connaissances générales qui identifient les relations entres les différentes dimensions (type de navire, type d'accident, zone maritime, etc.). L'antécédent et le conséquent de la règle sont connus mais pas la relation d'implication entre les deux parties. Les règles sont par exemple du genre "*Les accidents de navires de type Ti, concernent à c% les navires de type Ni*" et "*Les accidents de navires de type Ti sont localisés dans c% des cas dans la zone Zi*".,
- *Règles banales* : Ce sont les règles qui n'apportent pas d'informations nouvelles.

5. C'est un idicateur de fiabilité, il est égal au nombre d'occurences de la règle dans la base de données.
6. C'est un indicateur de précision de la règle qui est égal à la fréquence de la règle par rapport à la fréquence de l'antécédent.

2.4 Interprétation et validation des résultats

Nous avons découvert plus de deux cents règles intéressantes, que nous ne pouvons pas toutes exposer. Quelques unes de ces règles sont présentées et discutées ci-après. Ces règles ont été présentées à un sous capitaine de la marine marchande qui nous a aidé à les interpréter.

- *Règle 1 (Règle de prédiction)* : Vessel-Category=Fish catching –> Incident-Type= Machinery Failure (supp=0.39 ; conf= 0.60 ; lift=1.23)
  La première règle, nous informe que les incidents de navires de pêche sont causés dans 60% des cas par une panne mécanique.
- *Règle 2 (Règle de ciblage)* : Vessel-Category=Fish catching –> Vessel-Type= Trawler (supp= 0.14 ; conf= 0.43 ; Lift= 3)
  La deuxième règle, nous informe que si un accident concerne un navire de pêche alors dans 43% des cas c'est un chalutier. Selon le sous officier, les chalutiers sont les plus exposés au risque de naufrage car ils tirent un chalut qui peut s'accrocher et entrainer vers le fond le chalutier sans que les pêcheurs n'aient le temps de l'abandonner.
- *Règle 3 (Règle Banale)* : Vessel-Category=Passenger –> Pollution-Caused=No (supp= 0.15 ; conf= 0.73 ; lift= 1.2)
  Enfin, la dernière règle, présente une règle banale (inutile). La règle signifie que les accidents de navires transportant des voyageurs ne causent pas de pollution ce qui est normal car ils transportent des passagers et non des substances polluantes.

3 Conclusion et perspectives

Nous avons présenté dans cet article, une approche originale de découverte de règles d'association appliquée à des données spatiales statiques d'accidentologie de navires. Le résultat obtenu est un ensemble de règles de connaissances de prévision et de ciblage des situations à risque. Nous nous sommes focalisés dans cet article sur les risques liés à la sécurité (collision, échouement, etc.) et nous projetons de travailler par la suite sur les données spatiales dynamiques de déplacement de navires (données AIS). L'exploration automatique de ces données, peut révéler des modèles intéressants (les outliers de trajectoire, les périodiques, etc.) pour l'identification automatique des comportements anormaux de navires.

Dans la perspective d'améliorer les systèmes de surveillance maritime, les connaissances obtenues vont être intégrées dans un moteur de règles pour les exécuter d'une manière continue sur des flux de données de déplacements de navires, météorologique et de contexte. Le déploiement de ces règles va permettre d'identifier automatiquement dans ces flux de données, les situations vérifiant une ou plusieurs règles de connaissances, identifiant ainsi en quasi temps réel les situations à risque et les comportements anormaux de navires.

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining Association Rules between Sets of Items in Large Databases. In *the 1993 ACM SIGMOD International Conference on Management of Data, Washington*, Number May, Washington, D.C., pp. 207–216.

Cao, H., N. Mamoulis, et D. Cheung (2007). Discovery of Periodic Patterns in Spatiotemporal Sequences. *IEEE Transactions on Knowledge and Data Engineering 19*(4), 453–467.

CNUCED (2009). Etude sur les transports maritimes. Technical report, ConférenCe des nations Unies sUr le CommerCe et le développement, Geneva.

Darpa (2005). Predictive Analysis For Naval Deployment Activities (PANDA).

Jami, S., T.-Y. Jen, D. Laurent, G. Loizou, et O. Sy (2005). Extraction de règles d'association pour la prédiction de valeurs manquantes. *Revue africaine de la recherche en informatique et mathématiques appliquées (AREMA) 3*(numéro spécial CARI'04), 103–124.

Lee, J.-G., J. Han, et X. Li (2008). Trajectory Outlier Detection : A Partition-and-Detect Framework. In *Data Engineering, IEEE International Conference on Data Engineering (ICDE'2008)*, Cancun, Mexique, pp. 140–149.

Marven, C., R. Canessa, et P. Keller (2007). *Exploratory Spatial Data Analysis to Support Maritime Search and Rescue Planning*, pp. 271–288. New York : Springer Berlin Heidelberg.

Morel, M. (2009). SisMaris : Système d'Information et de Surveillance MARitime pour l'Identification des comportements Suspects de navire. In *première Conférence Méditerranéenne Côtière et Maritime CM 2*, Hammamet - Tunisie, pp. 261–264.

Morel, M., V. Flori, O. Poirel, A. Napoli, P. Salom, et G. Proutiere Maulion (2011). Trairtement et Authentification des MenAces et RISques en mer. In *Workshop Interdisciplinaire sur la Sécurité Globale (WISG08)*, Troyes, France.

Morel, M., A. Napoli, A. Littaye, J.-P. Georgé, et F. Jangal (2008). Surveillance et contrôle des activités des navires en mer. In *Workshop Interdisciplinaire sur la Sécurité Globale (WISG08)*, Troyes, France.

Srikant, R. et R. Agrawal (1995). Mining Sequential Patterns : Generalizations and Performance Improvements. In *Proceedings of the 5th International Conference on Extending Database Technology : Advances in Database Technology*, Volume 1057, pp. 3–14. Springer-Verlag.

Zhenhui, L., M. Ji, J.-G. Lee, L. Tang, et J. Han (2010). MoveMine : Mining Moving Object Databases. In *International Conference On Management of Data (SIGMOD)*, Indianapolis.

Summary

Maritime surveillance systems allow the recovery and the fusion of information on vessels (position, speed, etc.) for monitoring traffic on a display device. Today, the automatic identification of risks through these systems is difficult because of the complexity of formalizing the expertise, the large number of ships and the multiplicity of risks (collision, grounding, etc.). In addition, the periodic replacement of surveillance operators complicates the recognition of abnormal events which are scattered and fragmented in time and space. With the aim to upgrade the maritime surveillance systems, we propose in this paper, a novel approach based on data mining for the extraction of frequent patterns. This approach focuses on rules for forecasting and targeting for the automatic identification of situations inducing or constituting part of maritime accidents.