Résistance au bruit et à la rareté de la détection d'anomalies par arbre de décision de systèmes physiques simulés

Nesrine Bannour*, Anne Jeannin-Girardon*
Nicolas Lachiche*, Etienne Schneider*

* ICube, Université de Strasbourg
300 Bd Sébastien Brant
67400 Illkirch-Graffenstaden
bannour.nesrine@gmail.com,
{anne.jeannin, nicolas.lachiche, etienneschneider}@unistra.fr

Résumé. La détection d'anomalie est une tâche d'apprentissage dans laquelle les anomalies sont beaucoup plus rares que les comportements normaux. Notre objectif est de détecter une anomalie, en l'occurrence une fuite de fluide, le plus tôt possible, avant l'arrêt préventif de la machine. Dans cet article, nous étudions la résistance au bruit et à la rareté des anomalies d'une technique d'apprentissage supervisée, les arbres de décision. Nous considérons des données artificielles représentatives d'anomalies de systèmes physiques comme la crevaison d'un pneumatique ou la fuite de fluide réfrigérant d'une pompe à chaleur. Nos tests montrent qu'un arbre de décision est capable d'apprendre un seuil sur la pression observée, en présence de bruit, qui s'adapte à des fréquences très faibles d'anomalies, jusqu'à 1 pour 100 000.

1 Introduction

La détection d'anomalies est définie comme la recherche de structures dans un jeu de données qui ne correspondent pas au comportement attendu (Chandola et al., 2009).

Dans cet article, nous étudions la résistance au bruit et à la rareté de la détection d'anomalies par arbre de décision dans le cas de systèmes physiques simulés. En fait, la question est de savoir si l'apprentissage supervisé est adapté à des problèmes de détection d'anomalies, où la classe positive (anomalie) est extrêmement moins fréquente que la classe négative (normale), en présence de bruit bien sûr. La fonction cachée étant simple (un hyperplan, sur une seule variable), toutes les techniques d'apprentissage supervisées pourraient être utilisées (perceptron, SVM, plus proche voisin, etc.). Nous avons choisi les arbres de décision car ils fournissent un modèle explicite et peuvent gérer un grand nombre de données.

La suite de notre article s'organise comme suit. La Section 2 expose le contexte et la problématique de la détection d'anomalies des systèmes physiques que nous considérons : le pneumatique et la pompe à chaleur. Ensuite, nous détaillons dans la Section 3 notre générateur de données artificielles. La Section 4 présente notre modèle de détection d'anomalies et les différents résultats obtenus suite à son évaluation. Enfin, une conclusion est établie dans la Section 5 et propose quelques perspectives.

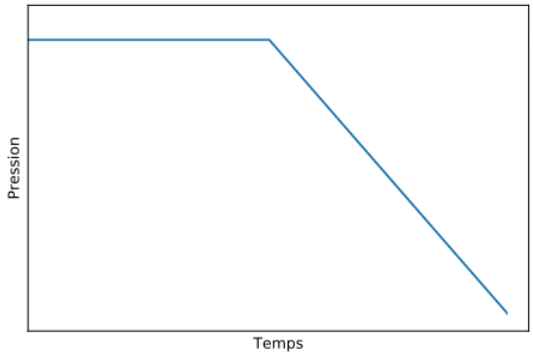

FIG. 1 – *Crevaison du pneumatique en l'absence de bruit.*

2 Problématique

Dans cet article, nous nous intéressons à deux systèmes physiques, le pneumatique et la pompe à chaleur. Cette section présente deux anomalies typiques de ces systèmes et comment les représenter sous la forme d'une même tâche d'apprentissage automatique.

Le pneumatique est un composant essentiel dans un véhicule qui assure sa sécurité et son confort. Des statistiques récentes montrent que la perte de pression du pneumatique est la principale cause de sa défaillance et par conséquent, de plusieurs accidents. Ainsi, des textes de réglementation européenne imposent la présence d'un système de surveillance de la pression des pneumatiques (SSPP) dans tous les nouveaux véhicules dès 2012. Ce dispositif doit être capable de détecter une perte de pression d'au moins 20% de la pression à chaud (El Tannoury, 2012). La crevaison d'un pneumatique est un problème simple à modéliser et à comprendre, cf. Figure 1. Ceci nous permet de mieux évaluer notre modèle d'apprentissage en nous focalisant sur la problématique du déséquilibre des données et de la robustesse de ce modèle.

La pompe à chaleur est un dispositif thermodynamique qui transfère l'énergie du milieu le plus froid vers le milieu le plus chaud. Il s'agit d'un système physique compliqué qui fait intervenir plusieurs variables dans son fonctionnement. Il peut y avoir donc une grande diversité de pannes. Parmi ces pannes, nous considérons la perte de fluide frigorigène ou réfrigérant. Cette perte de fluide pourrait se représenter comme une perte de pression (ou de poids) de fluide et conduire à une courbe similaire à celle de la crevaison d'un pneumatique. Cependant, les pompes à chaleur ne sont pas forcément équipées de capteurs mesurant cette pression. Des experts indiquent que cette perte est détectable en observant la vitesse de dégivrage qui diminue.

Les anomalies que nous considérons pour le pneumatique et pour la pompe à chaleur se ramènent à observer une variable continue, respectivement la pression du pneumatique et la vitesse de dégivrage, qui est constante dans le cas normal, mais qui diminue en cas de fuite. La tâche d'apprentissage consiste donc simplement à apprendre un seuil sur cette unique variable d'entrée. Dans la suite de cet article, nous prendrons l'exemple de la pression du pneumatique. Il ne s'agit ici que d'une première étape, avant de considérer une représentation plus complexe de ces systèmes physiques ou d'autres anomalies. L'intérêt principal de cette première étape est d'étudier si une technique d'apprentissage supervisée peut être mise en oeuvre sur un tel

problème, où les anomalies sont extrêmement moins fréquentes que le fonctionnement normal, et avec quelle performance.

3 Génération de données artificielles bruitées

Un jeu de données réel a une taille donnée et une complexité fixe et inconnue (Barse et al., 2003). L'intérêt de générer des données artificielle est de pouvoir contrôler la complexité de la fonction cible et d'avoir autant de données que souhaité.

Nous fixons arbitrairement mais de façon réaliste la pression nominale à 2 bars, considérant que le pneumatique se dégonfle si la pression est inférieure, et la pente de dégonflement à 0.01, c'est-à-dire que le dégonflement se fait linéairement en 40 unités de temps, avant que le seuil d'alerte de 1.6 bar soit atteint. Les paramètres que nous faisons varier sont le nombre N d'observations et la durée T de la phase normale à l'issue de laquelle le dégonflement commence. La durée T nous permet de contrôler le déséquilibre de nos données en testant différentes fréquences de déséquilibre. Ces fréquences représentent la fréquence d'anomalies dans un jeu de données. Par exemple, une fréquence 1 pour 10 signifie que dans un jeu de données de 10 observations, une seule observation représente une anomalie. Les cinq fréquences que nous utilisons sont : 1 pour 10, 1 pour 100, 1 pour 1 000, 1 pour 10 000, 1 pour 100 000. Le nombre N d'observations est obtenu en répétant plusieurs cycles terminés chacun par un dégonflement.

Les données que nous obtenons à la sortie de notre générateur sont des fichiers d'enregistrements (fichiers logs) qui décrivent les mesures de la pression datées et classées par ordre chronologique. Cependant, comme notre objectif est de détecter une anomalie à un instant t précis, nous transformons nos fichiers logs en un format attribut-valeur et suite à cette transformation, chaque ligne de données représentera une observation de la valeur de la pression et la classe associée.

Dans notre simulation du comportement du pneumatique, nous ne modélisons pas explicitement d'autres paramètres qui peuvent influencer son comportement (Température extérieure, échauffement du roulement, etc). Afin de prendre en considération l'incertitude liée aux mesures de la pression et de nous rapprocher le plus possible du comportement réel du pneumatique, nous ajoutons un bruit de type Gaussien à la pression générée précédemment. La classe associée à chaque ligne ne change pas, bien sûr. Afin de mieux évaluer la résistance au bruit de la détection d'anomalies par arbre de décision, nous avons choisi de tester deux bruits gaussiens :

— **Premier niveau de bruit :** Un bruit gaussien avec une espérance μ égale à 0 et un écart type σ égale à 0.01. Cet écart-type correspond volontairement à la pente du dégonflement du pneumatique. Sans bruit, le dégonflement serait détecté dès que la pression devient strictement inférieure à 2 bars.

— **Deuxième niveau de bruit :** Un bruit gaussien avec une espérance μ égale à 0 et un écart type σ égale à 0.03, donc 3 fois plus "fort".

4 Détection d'anomalies par arbre de décision

Peu de travaux sur la détection d'anomalies s'appliquent à une fonction cible aussi simple en présence de données très nombreuses et de classes très déséquilibrées. Par exemple nos tests

d'une approche typique pour traiter des données déséquilibrées, one-class SVM (Scholkopf et al., 1999), montrent qu'elle ne résiste pas au déséquilibre de nos données et génère un grand nombre de fausses alertes. Les forêts d'arbres d'isolation, d'après l'algorithme et les tests publiés (Liu et al., 2008), risquent d'être moins précises car elles sous-échantillonnent les données alors que nous cherchons des anomalies jusqu'à 10 fois plus rares que les tests publiés, et plus lentes puisqu'elles construisent -aléatoirement qui plus est- un arbre séparant les données alors que nous ne cherchons qu'un seuil/noeud et considérons plus de 100 fois plus de données.

Détecter le dégonflement du pneumatique consiste à trouver la pression qui représente le seuil de séparation entre le comportement normal et le comportement anormal. Ainsi, nous imposons que la profondeur de notre arbre de décision ait une valeur de 1 en conservant les autres paramètres par défaut. Nous utilisons dans ce travail l'implémentation CART de l'arbre de décision de la librairie Scikit-learn (Pedregosa et al., 2011) de Python dont le critère de segmentation est l'indice de diversité de Gini. Les évaluations sont réalisées avec 80 millions de données pour chaque expérience. Elles sont divisées en 75% pour la phase d'entraînement et 25% pour la phase de test.

La performance de notre modèle est évalué par la précision, le rappel et la F-mesure. Rappelons que l'exactitude (*accuracy* en anglais) n'est pas une mesure de performance adaptée à notre problème. En effet, en travaillant avec un jeu de données déséquilibré, un programme d'apprentissage a tendance à construire des modèles qui prédisent correctement la classe majoritaire et ont donc une exactitude élevée. Mais cette exactitude ne sera pas utile parce qu'elle ne reflète pas l'aptitude à détecter les anomalies et donc la classe minoritaire.

Nous interprétons les résultats de notre modèle d'apprentissage supervisé en fonction en des deux niveaux de bruits afin d'établir un bilan de sa résistance au bruit et au déséquilibre des données.

— **Premier niveau de bruit :** Avec le premier niveau de bruit, la performance de l'arbre de décision est assez impressionnante même avec des données très déséquilibrées. En effet, quand la classe majoritaire (classe normale) devient plus importante, l'arbre de décision devient plus exigeant en terme du seuil de décision qui baisse. L'arbre de décision "préfère" rater quelques exemples positifs en les attribuant à la classe normale (et donc le rappel diminue) plutôt que de déclencher de nombreuses fausses alertes en classant mal les exemples négatifs de plus en plus nombreux. Il maintient ainsi une précision élevée. Vu que la F-mesure est une moyenne de la précision et du rappel, elle diminue lentement.

— **Deuxième niveau de bruit :** Avec le deuxième niveau de bruit, nous observons un comportement similaire : le seuil de décision diminue en augmentant le déséquilibre. Néanmoins, l'arbre de décision est moins performant qu'avec le premier niveau de bruit : il n'arrive pas à maintenir une précision parfaite (égale à 1) même pour un déséquilibre minimal. Ceci est dû à l'augmentation du bruit.

Pourtant pour la cinquième fréquence, nous observons une précision égale à 1 avec les deux niveaux de bruits. C'est étonnant vu qu'il s'agit d'un très grand déséquilibre de données. Nous supposons qu'il s'agit d'une aberration probablement liée aux limites d'implémentation de notre générateur. En effet, vu que nous ajoutons un bruit Gaussien à nos données, les exemples négatifs (dont la pression est nominale, 2 bars) suivent une loi normale également. Si nous reprenons l'exemple de la cinquième fréquence avec le premier niveau de bruit, nous aurons

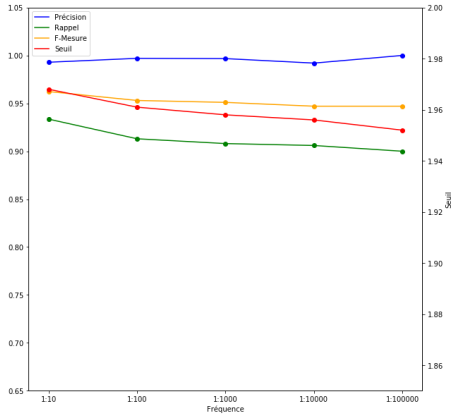

FIG. 2 – *Performance de l'arbre de décision avec le premier niveau de bruit.*

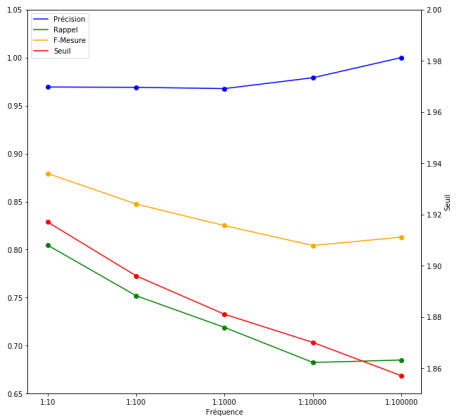

FIG. 3 – *Performance de l'arbre de décision avec le deuxième niveau de bruit.*

z	4.4	4.5	4.6	4.7	4.8	4.9
F(z)	0.999995	0.999997	0.999998	0.999999	0.999999	1.000000

TAB. 1 – *Probabilité F(z) de trouver une valeur inférieure à z*

des données de moyenne μ égale à 2 avec presque le même écart type du bruit σ qui est égal à 0.01. Selon la figure 2, le seuil a été fixé par l'arbre à 1,952. La probabilité d'avoir un exemple inférieur ou égal à ce seuil revient à calculer la probabilité $P(X \leq 1.952)$. Nous pouvons ramener les calculs à une loi normale centrée réduite $(|-1,952-2|/0,01 = 4,8)$ et la probabilité à calculer sera donc $P(Z \geq 4.8) = 1/1000000$, cf. Table 1. Nous avons donc une chance sur un million, en théorie, de tirer une observation de valeur inférieure à ce seuil mais il est probable qu'en pratique notre générateur ne génère aucune donnée au delà de 5 σ. Ceci explique la valeur de la précision, égale à 1, lors du dernier test puisque notre modèle n'a trouvé aucune valeur de la classe normale inférieure au seuil de décision pour la classer comme étant dégonflement. De même pour le deuxième niveau bruit suite à l'augmentation du bruit et donc de σ, le seuil se rapproche de $2 - 0,03 * 4,8 = 1,856$. Sur 80 millions d'exemples négatifs générés, aucun ne descend en dessous de ce seuil. Même en générant un nombre encore plus grand d'exemples négatifs, il est probable qu'aucun ne serait inférieur à ce seuil.

5 Conclusion

Malgré l'aspect temporel de nos données, notre objectif est de faire une détection d'anomalie à chaque instant. Nous constatons que nous n'avons pas eu besoin de prendre en compte les valeurs précédentes de la pression et que la valeur courante de la pression a suffi.

Nous avons généré un très grand nombre de données (80 millions pour chaque fréquence et niveau de bruit) afin d'avoir au moins 200 exemples positifs dans le jeu de test dans le cas

des jeux de données les plus déséquilibrés. Nous avons observé que nous avons probablement atteint la limite pratique du notre générateur aléatoire de bruit gaussien utilisé.

Ces expériences permettent de constater qu'un arbre de décision est capable de traiter plusieurs millions d'exemples et d'apprendre un seuil permettant d'obtenir d'excellents précision, rappel et F-mesure alors que la fréquence de la classe positive descend jusqu'à 1 pour 100 000.

Les prochaines étapes sont de tester sur des données réelles. Nous espérons que ce sera l'occasion de considérer des fonctions cachées plus complexes (sur plusieurs attributs) et qui nécessiteront éventuellement de prendre en compte l'historique de la séquence de données.

Références

Barse, E. L., H. Kvarnström, et E. Jonsson (2003). Synthesizing test data for fraud detection systems. In *Proceedings of the 19th Annual Computer Security Applications Conference*, ACSAC '03, Washington, DC, USA, pp. 384–. IEEE Computer Society.

Chandola, V., A. Banerjee, et V. Kumar (2009). Anomaly detection : A survey. *ACM Comput. Surv. 41*(3), 15 :1–15 :58.

El Tannoury, C. (2012). *Development of vehicle tire pressure monitoring tools using methods based on spectral analysis and observers synthesis*. Theses, Ecole Centrale de Nantes (ECN).

Liu, F. T., K. M. Ting, et Z.-H. Zhou (2008). Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008)*, pp. 413–422.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, et E. Duchesnay (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Scholkopf, B., J. Platt, J. Shawe-Taylor, A. Smola, et R. Williamson (1999). Estimating the support of a high-dimensional distribution. Technical report, Microsoft Research. MSR-TR- 99-87.

Summary

Anomaly detection is a learning task in which anomalies are extremely less frequent than the normal behaviour. We aim at detecting anomaly, actually fluid leakage, as soon as possible, before a preventive shutdown of the machine. In this article, we study the resistance to noise and to rarity of anomalies of a supervised learning technique, decision trees. We consider artificial data representative of physical system anomalies such as a tire puncture or a refrigerant leak from a heat pump. Our tests show that a decision tree is able to learn a threshold on the pressure observed, in the presence of noise, which adapts to very low frequencies of anomalies, down to 1 per 100,000.