Découverte de motifs intelligibles et caractéristiques d'anomalies dans les traces unitaires

Olivier Cavadenti, Victor Codocedo, Mehdi Kaytoue, Jean-François Boulicaut

Découverte de motifs intelligibles et caractéristiques d'anomalies dans les traces unitaires

Olivier Cavadenti[*,**], Victor Codocedo[*], Mehdi Kaytoue[*], Jean-François Boulicaut[*]

[*]Université de Lyon. CNRS, INSA-Lyon, LIRIS. UMR5205, F-69621, France.
[**]Actemium Saint-Etienne
contact : prenom.nom@insa-lyon.fr

Résumé. De nombreuses industries manufacturières s'intéressent aujourd'hui à l'exploitation des grandes collections de traces unitaires. Les applications sont multiples et vont du simple "reporting" à la détection de fraudes en passant par la gestion de retours ou encore la mise en évidence d'incohérences dans les circuits de distribution. Une étape importante consiste à détecter des anomalies dans des collections de traces. Si les travaux concernant la détection d'anomalies sont assez nombreux, peu permettent de caractériser les anomalies détectées par une description intelligible. Étant donné un ensemble de traces unitaires, nous développons une méthode d'extraction de motifs pour détecter et contextualiser des comportements non conformes à un modèle expert (fourni ou construit à partir des données). Le degré d'anomalie est alors quantifié grâce à la proportion du nombre de mouvements des objets qui ne sont pas prévus dans le modèle expert. Cette recherche est financée partiellement par un programme industriel qui ne permet ni de dévoiler le contexte concret ni de parler des données réelles. Ainsi, nous validons empiriquement la valeur ajoutée de la méthode proposée par l'étude de traces de mobilité dans un jeu vidéo : nous pouvons alors discuter d'un motif qui explicite les raisons de l'inexpérience de certains joueurs.

1 Introduction

Avec la dissémination de nombreux systèmes de capteurs, de très grands volumes de données sont accessibles sous la forme de collections de traces. Ces traces correspondent à la séquence des évènements *captés* dans un système qui définit ce qui est *captable*. Les traces modélisent alors la vie des objets dans ce système. Le type de traces qui motive cette recherche est celui des traces unitaires de produits manufacturés qui peuvent être tracés depuis leur fabrication jusqu'à leur vente en transitant via un réseau logistique éventuellement complexe. Les traces unitaires codent les comportements des objets, dont certains sont prévisibles, qui permettent d'avoir accès à de nombreuses informations spatio-temporelles ou sémantiques sur les processus appliqués à ces objets. Parallèlement, ces systèmes de traçage possèdent eux-mêmes de nombreuses caractéristiques comme les distances et positions entre les capteurs ainsi que leur type, leur propriétaire, les données qu'ils fournissent, leur état, etc. On peut donc se demander si les comportements des objets vérifient bien les processus attendus par ceux qui ont

mis en place le système : c'est un enjeu majeur pour la découverte d'anomalies dans des collections de traces unitaires. Ce cas est observable dans de nombreux scénarios d'applications : on peut rechercher des agents qui dévient de leurs déplacements habituels (taxis, joueurs, personnes) ou voir si des objets manufacturés suivent un circuit de distribution attendu. Ainsi, cette recherche est motivée par la détection de produits contrefaits ou vendus en dehors du marché pour lequel ils ont été achetés. Nous considérons dans cet article la recherche et la caractérisation d'anomalies dans un environnement de traces unitaires et/ou de comportements en définissant la notion d'anormalité par rapport à un modèle expert, un modèle, appelé modèle de filière dans le cas des applications aux produits manufacturés, qui détermine pour partie l'attendu. La recherche d'anomalies dans un ensemble de données est un problème de fouille de données bien connu (Aggarwal (2013),Chandola et al. (2009)). Il consiste à découvrir des ensembles d'objets dont la valeur des attributs dévient suffisamment de l'ensemble des objets de la base de données. Or dans notre problématique, les anomalies peuvent être fréquentes mais non conformes à un comportement décrit par une connaissance experte a priori et non au sein de l'ensemble des données captées. Notre première contribution est de formaliser cette tâche de détection d'anomalies. Bien qu'assez peu étudié dans la communauté de la fouille de données, quelques travaux ont souligné que les modèles experts pouvaient être exploités dans des processus de découverte de collection de motifs pertinents (par exemple, utilisation d'un réseau bayésien comme connaissance a priori pour la découverte d'ensembles fréquents par Jaroszewicz et al. (2009) ou exploitation d'un modèle expert mathématique par Flouvat et al. (2014)). De plus, si la détection d'objets anormaux a été bien étudiée, la description et la caractérisation de ces anomalies restent un champ restreint à quelques études récentes comme celles de Tang et al. (2013) ou de Duan et al. (2015). Notre seconde contribution est de progresser sur la description des anomalies (anomalies contextuelles) via une méthode de génération de contexte de fouille et une découverte d'anomalies, qui sont des ensembles de propriétés qui caractérisent davantage les traces considérées comme anormales que les traces normales, en exploitant des motifs émergents (Dong et Li (1999)). Nous voulons en effet établir leur capacité à décrire en premier lieu les anomalies et non les comportements attendus.

L'article est organisé comme suit. Nous exposons notre méthode de découverte de descriptions d'anomalies en décrivant les principales notions utiles et en définissant le problème dans la Section 2. La Section 3 donne lieu à des expérimentations avec deux cas de découverte d'anomalies et leurs descriptions dans le jeu de stratégie Dota 2. Nous exposons les travaux similaires dans la Section 4 avant de conclure.

2 Méthode de découverte de descriptions d'anomalies

2.1 Notations et définition du problème

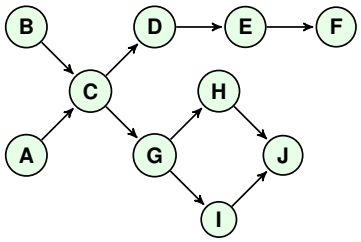
Nous appelons *trace unitaire* la séquence des enregistrements captés par un système donné lors du déplacement d'un objet. A chaque objet correspond une et une seule trace.

Définition 1 (Enregistrement)**.** *Soit un ensemble d'attributs $A = \{A_1, ..., A_n\}$ numériques ou catégoriels. Un enregistrement $r \in R$ est un n-uplet $r = (a_1, ..., a_n)$ avec $a_i \in dom(A_i)$.*

Définition 2 (Trace unitaire)**.** *On note $t = \langle r_1, ..., r_k \rangle$ une trace unitaire avec $r_i \in R$. Une trace unitaire indique la suite d'évènements captés lors du cycle de vie d'un produit. Une collection de traces est notée $\mathcal{T}$.*

Définition 3 (Modèle expert de filière)**.** *Un modèle de filière correspond à la connaissance globale que l'on a des sites et des transitions potentielles entre eux. Il se peut que des sites de captation ne soient pas présents dans le modèle. Chaque enregistrement est réalisé sur un site, c'est un attribut d'enregistrement r particulier (noté $site(r)$). L'ensemble des sites forme un graphe $G = (V, E)$ avec V les nœuds et $E \subseteq V \times V$ les arêtes, que l'on appelle modèle de filière. Un enregistrement r correspond à une action effectuée à un instant t et à un site $v \in V$.*

Exemple 1. *Soit le modèle de filière décrit dans la Figure 1 avec un ensemble de sites V. Deux séquences d'enregistrements appartenant à deux traces unitaires sont présentées dans le tableau à droite de la Figure 1. L'attribut 'loc' permet de préciser le site de l'enregistrement. On observe que la trace numéro 2 possède une séquence de sites : $< A, A, C, C, D >$. Nous voyons sur le modèle de filière que ces sites sont connectés par le chemin $A \rightarrow C \rightarrow D$.*



tid	event_id	timestamp	loc	action
1	1	101	A	COMMISSING
1	2	102	A	STORING
1	3	105	A	SHIPPING
1	4	251	C	RECEIVING
2	5	150	A	COMMISSING
2	6	152	A	SHIPPING
2	7	172	C	RECEIVING
2	8	174	C	SHIPPING
2	9	263	D	RECEIVING

FIG. 1 – *Un modèle expert (à gauche) ainsi que des traces captées qui sont données sous forme d'enregistrements (à droite).*

La trace unitaire $t \in \mathcal{T}$ d'un objet dénote son comportement tout au long de son déplacement dans un réseau de sites. Il est important de remarquer que le modèle de filière correspond aux comportements *captables* et que la trace unitaire correspond aux comportements *captés*. Ainsi, il est possible que les déplacements captables ne sont pas captés. Pour chaque trace unitaire, on génère des *descriptions* qui peuvent concerner des propriétés/attributs des nœuds du graphe visités par la trace, ou d'autres propriétés numériques, booléennes, etc.

Définition 4 (Description d'une trace unitaire)**.** *La fonction $description : \mathcal{T} \rightarrow D$ associe à chaque trace $t \in \mathcal{T}$ une description sous forme vectorielle classique formée de paires d'attribut/valeur, ou dans un langage de description où l'on peut construire un infimum demi-treillis.*

Problématique. Dans de nombreux scénarios, les comportements des objets, captés sous forme de traces, sont différents de ceux décrits par le modèle de filière. Deux enregistrements d'un même objet peuvent être captés consécutivement sur deux localisations sans que la transition n'apparaisse ($e \notin E$). Ces comportements correspondent à des comportements déviants et rares qui peuvent être dû à des erreurs ponctuelles ou à des actions intentionnelles (fraudes). Par exemple, si un groupe de produits manufacturés est acheminé par erreur (ou dans le cadre d'un détournement) à une destination non prévue, ils seront enregistrés à deux localisations non reliés par le modèle de la filière qui fait transiter les produits. S'il est possible de séparer les traces unitaires en deux bases –celles dont un grand pourcentage de transitions sont absentes

du modèle (anormales) et les autres (normales)–, il n'est pas possible d'expliquer les raisons de ces détournements. C'est le problème que nous voulons résoudre. Pour ce faire, on se tourne vers une méthode de découverte supervisée de motifs (Novak et al. (2009)), où l'objectif est d'extraire des descriptions qui caractérisent des classes. Dans notre cas, une description est intéressante si elle apparaît dans de nombreuses traces anormales, et très peu dans les classes normales, ce que l'on appelle des motifs émergents (Dong et Li (1999)).

Méthode. En pratique, nous proposons donc une méthode qui permet de transformer les traces unitaires dans un contexte de fouille d'itemsets via un codage de propriétés décrivant les traces. Ces propriétés intègrent des connaissances de haut niveau sur les objets qui peuvent être de nature spatio-temporelle ou symbolique. Nous cherchons, parmi les *descriptions* des traces unitaires classées en anormales, les ensembles de propriétés qui décrivent ces traces et non les traces normales. Cette méthodologie permet d'une part d'exploiter la connaissance experte du domaine pour définir une notion d'anormalité plus proche et plus compréhensible pour les experts ; et d'autre part d'introduire la description des anomalies dans le codage des propriétés même et non dans le processus algorithmique. Il est possible en codant des propriétés intéressantes pour la caractérisation de causes possibles d'anomalies, de produire une infinité de scénarios à partir des mêmes traces unitaires et obtenir des explications des anomalies.

Exemple 2 (suite)**.** *Pour définir l'anormalité d'une trace unitaire, nous utilisons un encodage simple qui consiste à générer la séquence des sites visités. A partir des traces de la Figure 1, on a les séquences de sites* $< A, C >$ *et* $< A, C, D >$ *qui indiquent que l'objet de la trace 1 est passé par A puis C, celui de la trace 2 est passé par A,C puis D. Une trace ayant la séquence de sites* $< A, C, H, J >$ *est anormale : la transition entre C puis H n'existe pas dans le modèle de filière. Une trace peut contenir une proportion variable de transitions non présentes dans le modèle de filière. C'est ce qui définira son appartenance à l'ensemble des traces normales (proportion faible) ou à l'ensemble des traces anormales (proportion forte). Cependant, nous n'avons aucune information pour définir ce qui caractérise exclusivement la base de traces anormales (les causes des anomalies). Avec une description assignée à chaque trace nous pouvons fournir une explication plus riche de ces anomalies en étudiant les descriptions communes aux traces anormales qui ne décrivent pas les traces normales.*

2.2 Modèle expert ou modèle de filière

Le modèle de filière est la donnée de $G = (V, E)$. Il peut s'agir d'un modèle expert, c'est-à-dire les connaissances disponibles des experts sur les flux d'objets prévus lors du fonctionnement du système qui peuvent être partielles. Le modèle de filière peut aussi être produit de manière automatique quand la connaissance experte n'est pas disponible ou non formalisée. On peut imaginer diverses manières d'agréger la totalité des traces pour générer le modèle de filière, en partant du principe classique que la normalité est définie par un agrégat. Par exemple, on calcule la moyenne μ du nombre de passages entre les paires de nœuds du graphe. Deux localisations ont une arête dans le graphe si le nombre de passages est contenu dans l'intervalle $[\mu - 2\sigma, \mu + 2\sigma]$ où μ est la moyenne du nombre de passages pour toutes les arêtes et σ est l'écart type. Ce point sera traité tout particulièrement dans la partie expérimentale (Section 3).

Exemple 3 (suite)**.** *On considère le graphe donné en Figure 1 (gauche) comme un modèle expert, décrivant la connaissance experte, où chaque nœud et arête peuvent être enrichis d'in-*

formations, comme l'ensemble des types d'objets qui transitent par les sites et arêtes, la durée moyenne de passages des objets ou encore des informations spatiales selon différentes échelles (région, ville, pays du site).

2.3 Codage des propriétés

Pour chaque trace d'une base de traces unitaires $\mathcal{T}$ on peut construire une description, dont l'ensemble forme un contexte de fouille. Ces descriptions peuvent être de types variées et hétérogènes (numériques, symboliques, séquentiels, etc.) à partir du moment où ces descriptions peuvent être partiellement ordonnées et qu'il existe un solveur pour les extraire. Dans ce qui suit, on se limite à des propriétés booléennes. Celles-ci permettent de créer des conditions sur la présence de valeurs ou non parmi les attributs des traces unitaires ou des tests sur les valeurs de ces attributs. De manière plus générale, une propriété booléenne correspond à l'évaluation d'une expression sur la description de la trace $description(t) \in D, t \in \mathcal{T}$ avec t_i la trace dont l'attribut tid a pour valeur i.

Définition 5 (Encodage expert des traces par des propriétés.)**.** *Soit P un ensemble de propriétés booléennes. Chaque propriété $p \in P$ est vue comme une fonction $p : \mathcal{T} \rightarrow \{true, false\}$. Une propriété consiste à évaluer une expression logique $\mathcal{C}$ (règles expertes, expressions régulières, présence d'évènements,etc.) construite à partir de la description d'une trace $description(t)$.*

Exemple 4 (suite)**.** *Soit t_{A_i} l'ensemble des valeurs de l'attribut A_i dans les enregistrements d'une trace t, on peut définir les propriétés booléennes $\mathcal{C}_1(t_{action}) \equiv \exists\{'STORING'\} \in t_{action}$, notée $STORING$, qui indique que l'objet a été stocké au moins une fois, et $\mathcal{C}_2(t_{loc}) \equiv \exists\{'C','D'\} \in t_{loc}$, notée C_D, s'il est passé par les localisations C et D.*

Définition 6 (Contexte de fouille)**.** *Soit $\mathcal{T}$ l'ensemble de traces et P l'ensemble de propriétés proposées par l'expert. Le contexte de fouille est donné par la relation binaire $D_c \subseteq \mathcal{T} \times P$ où $(t, p) \in D_c$ si la trace t respecte la propriété p. D_c est souvent appelé base de transactions dans la littérature en fouille de motifs.*

Exemple 5 (suite)**.** *On peut construire un contexte de fouille D_c avec les propriétés décrites dans l'exemple précédent. Par exemple dans le cas de la base d'enregistrements de la Figure 1, on obtient le contexte de fouille suivant : $D_c = \{(t_1, \{STORING\}), (t_2, \{C\_D\})\}$.*

2.4 Motifs émergents de descriptions d'anomalies

Chaque trace $t \in \mathcal{T}$ est maintenant décrite par des propriétés booléennes. Nous proposons de fouiller ce contexte afin de trouver les descriptions de traces qui à la fois (i) apparaissent fréquemment et (ii) montrent un degré d'anomalie élevé. Considérons ici que la vie d'un objet est anormale si sa trace unitaire possède une forte proportion de transitions entre les sites qui ne sont pas présentes dans le modèle de filière.

Pour comprendre ces notions, nous rappelons les définitions de motifs fréquents et leur support. Nous restons dans le cadre de motifs binaires (itemsets), on notera qu'une généralisation directe a été proposée par Ganter et Kuznetsov (2001) pour traiter des données hétérogènes dont les descriptions possibles peuvent être ordonnées au sein d'un demi-treillis. Cette généralisation est suffisante pour traiter les données de type numérique, séquentiels, ordres partiels et graphes dans certaines conditions.

Définition 7 (Motifs, support, motifs fréquents et motifs fermés)**.** *Soit un ensemble de propriétés P et $\mathcal{D}_c$ une base de transactions où chaque transaction correspond à la description d'une trace unitaire. Un ensemble $X \subseteq P$ est appelé un itemset. Le support d'un itemset X dans $\mathcal{D}_c$ noté $supp(X)$ correspond au nombre de descriptions des traces qui contiennent X soit $supp(X) = |\{X|X' \in \mathcal{D}_c, X \subseteq X'\}|$. Un itemset fermé est tel qu'il n'existe pas d'itemsets de même support qui le contiennent. On appelle un motif fermé fréquent un itemset fermé ayant un support supérieur à un seuil min_sup.*

Pour définir la mesure d'anomalie d'un motif, on partitionne l'ensemble de traces en deux : $\mathcal{T} = \mathcal{T}^+ \cup \mathcal{T}^-$ avec $\mathcal{T}^+ \cap \mathcal{T}^- = \emptyset$. On fixe un seuil d'anomalie $\theta \in [0, 1]$: une trace appartient à l'ensemble $\mathcal{T}^-$ si au moins une proportion θ de ses transitions entre ses sites ne sont pas décrites dans le graphe ; sinon à l'ensemble $\mathcal{T}^+$.

Définition 8. *(Traces positives et négatives) Soit la fonction $transitions : \mathcal{T} \rightarrow E$ qui associe à chaque trace unitaire l'ensemble de ses transitions présentes dans $G = (V, E)$, la cardinalité d'un trace unitaire $|t|$ indiquant le nombre de sites qu'elle traverse, et $\theta \in [0, 1]$ un seuil d'anomalie, on a : $\mathcal{T}^+ = \{t|t \in \mathcal{T}, \frac{|transitions(t)|}{|t|-1} < \theta\}$ et $T^- = \mathcal{T} \backslash \mathcal{T}^+$.*

Ayant introduit les classes positives et négatives de traces, on peut alors définir un mesure d'émergence qui est d'autant plus forte que le motif apparait dans une classe choisie. Une mesure générale a été introduite par Dong et Li (1999). Nous préférons une mesure normalisée entre $[-1, +1]$ dans nos expériences.

Définition 9 (Mesure d'émergence normalisée)**.** *Soit un motif X, la base $\mathcal{T}^+$ et la base $\mathcal{T}^-$, on a la mesure d'émergence normalisée suivante :*

$$\phi(X) = \frac{|supp(X, \mathcal{T}^+)| - |supp(X, \mathcal{T}^-)|}{|supp(X, \mathcal{T}^+)| + |supp(X, \mathcal{T}^-)|}$$

Si la valeur d'émergence du motif est strictement inférieure à 0, alors on a un motif émergent de description d'anomalie. Si la mesure vaut 0, alors il y a autant de traces dans $\mathcal{T}^-$ que dans $\mathcal{T}^+$ qui respectent ce motif.

Problématique 1 (Découverte des motifs émergents de descriptions d'anomalie)**.** *Soit une base de traces unitaires $\mathcal{T}$ et un contexte de fouille D_c produit à partir des propriétés P, un modèle de filière $G = (E, V)$, un seuil d'anormalité θ, et un support minimum min_sup, nous cherchons l'ensemble de tous les motifs émergents de description d'anomalie X de D_c tel que $supp(X) > min\_sup$ et $\phi(X) < 0$.*

2.5 Notes algorithmiques

Tout d'abord, il faut souligner que l'on peut se restreindre à l'extraction de motifs fermés. En effet, Plantevit et Crémilleux (2009) ont montré que les motifs fermés sont ceux qui maximisent les mesures basées sur le support, ce qui est le cas de la mesure d'émergence. De fait, les solveurs classiques d'extraction de motifs fermés peuvent être utilisés. Dans nos expérimentations, nous utiliserons l'algorithme CHARM proposé par Zaki et Hsiao (2005) car nous encodons les traces avec des propriétés booléennes. Un autre solveur pourra être utilisé pour un autre domaine de motifs. Une technique classique est d'inclure la classe (positive ou négative)

directement dans la description de chaque objet pour connaitre directement le support d'un motif dans une base (positive ou négative) sans avoir à la scanner. Par exemple, cette technique a été utilisée par Bosc et al. (2014) dans un contexte similaire.

3 Expérimentations

Nous expérimentons notre approche afin de valider de manière empirique son utilité. Bien que cet article considère le problème de caractérisation d'anomalies à partir de traces unitaires, comme nous avons des problèmes de confidentialité concernant les traces de produits manufacturés, les expérimentations sont réalisées sur des traces comportementales de joueurs dans un jeu vidéo. Nous présentons le principe de ce jeu avant de développer notre méthodologie.

3.1 DOTA2 : un jeu de stratégie en temps réel

Une partie est jouée sur un terrain de jeu où deux équipes de cinq joueurs s'affrontent en temps réel. Chaque équipe doit défendre son *château* et détruire celui de l'opposant pour gagner. Chaque joueur contrôle un héros qu'il peut déplacer sur le terrain (contrôle souris/clavier) et doit entrainer en collectant de l'argent, de nouveaux objets, des compétences et en se battant contre les forces ennemies. La Figure 2 présente les zones d'influence initiales des deux équipes. L'équipe rouge appelée *the dire* (resp. verte pour *the radiant*) défend son château situé au coin bas-gauche (resp. coin haut-droit). Trois "chemins" principaux (*top*, *mid*, *bot*) séparent les équipes et contiennent des tours défensives. Chaque joueur a un rôle bien défini, qui dépend du héros qu'il a choisi (parmi 110 héros). Par exemple, un rôle consiste à défendre et étendre la zone d'influence sur un chemin spécifique ; un autre est de changer souvent de chemin pour rejoindre un allié et attaquer un ennemi par surprise. Sachant qu'une équipe ne peut voir que les zones qu'elle contrôle et donc estimer la position des ennemis, le fait de déclencher des attaques synchronisées est la clef du succès, tout en sachant garder son rôle initial sans quoi la progression du héros (amélioration de ses capacités) est beaucoup plus lente.

FIG. 2 – *Terrain de Dota2*

3.2 Découverte d'anomalies et leurs descriptions à DOTA2

De manière similaire à un jeu de rugby, le positionnement d'un héros sur le terrain est crucial pour tenir la ligne de défense. Rappelons qu'à chaque héros correspond un rôle : ce rôle n'est cependant pas indiqué aux nouveaux joueurs, et l'apprentissage du jeu est assez long. A chaque rôle correspondent aussi des zones optimales à contrôler pour avoir de meilleures chances de gagner. D'après les joueurs experts, chaque héros peut prétendre à au moins trois rôles parmi une quinzaine [1]. Nous modélisons donc le terrain de jeu par un graphe dont les sommets sont des points d'intérêt connus (château, tours, magasins, ... un sur-ensemble des points rouges et verts sur la Figure 2). Les arêtes correspondent aux transitions entre deux

1. http://www.dota2.fr/apprendre/guides/guide-des-differents-roles

points d'intérêt, c'est-à-dire quand un héros est détecté à un point d'intérêt différent du point d'intérêt précédent dans la trace de comportement.

On part du principe qu'ayant une base de traces de mouvements assez conséquente pour un héros/rôle particulier, le comportement moyen devrait présenter la normalité, i.e. les zones qu'il doit contrôler. En décrivant les traces par un ensemble de propriétés qui dénotent des éléments de stratégies, on suppose que les descriptions d'anomalies seront des stratégies qui sont inefficaces, générées par les joueurs en phase d'apprentissage. Un tel exemple de graphe moyen (construit comme expliqué en Section 2) est donné par la Figure 3 pour le héros *Dark Seer* à partir de 500 traces de jeux issues du site `http://dotabank.com/`. Le motif décrira donc les erreurs stratégiques, erreurs d'autant plus fortes que l'est la mesure d'anomalie, et le support indiquera classiquement la fréquence d'apparition de cette anomalie.

FIG. 3 – *Modèle* $G_{darkseer}$

3.3 Analyse qualitative des motifs extraits

Anomalies expliquées par une erreur dans le choix des compétences. Nous détaillons le choix des propriétés P pour créer le contexte de fouille D_c à partir de l'ensemble des 500 traces de jeux. Durant la partie, chaque héros, en fonction d'actions précises, gagne de l'expérience et des niveaux. A chaque niveau, il choisit une compétence, parmi 3, sauf aux niveaux 6 et 11 où il peut en choisir une quatrième. Le choix d'un chemin dans cet arbre de compétences est stratégiquement très important et donne lieu à des guides (manière de choisir les compétences à un niveau précis pour maximiser ses possibilités de victoire). Nous codons alors une propriété booléenne de type '*La compétence* n *a été prise au niveau* x', qui indique qu'une compétence (1,2,3 ou 4) a été prise par le joueur à un niveau donné. De plus, nous introduisons une propriété qui indique s'il a acheté ou non un objet [2] très fréquemment pris par ce *héros*. Avec un seuil d'anomalie $\theta = 0.3$ (il faut que 30% des transitions du héros ne respectent pas le modèle pour que la trace soit considérée anormale) et un seuil de support minimum très bas $min\_sup = 1\% = 5,175849$ motifs fréquents sont extraits, mais seuls 6 ont une mesure d'anomalie négative, i.e. des descriptions de traces qui s'écartent fortement du modèle. En partie listés dans la Table 1, ces motifs ont la même mesure d'anomalie -0.199 et un support très faible : il y a 5 instances dans l'image de chaque motif, nous faisons bien face à une anomalie. Selon notre expert, ces motifs montrent des erreurs de stratégies évidentes. Par exemple, choisir la compétence 4 au niveau 8 (propriété $comp\_4\_at\_level\_8$) est une erreur majeure, malgré tout faite dans 5.4% des 500 parties dont nous disposons [3]. On observe aussi qu'ils n'ont pas acheté l'objet soul ring ce qui est rare également. Nous avons retrouvé le profil des 5 joueurs impliqués dans le motif $\#1$ au moment auquel les parties avaient été jouées : chacun n'avait réalisé qu'un nombre très faible de parties, ce sont bien des débutants.

Autres explications d'anomalies. En intégrant d'autres propriétés P, établies après discussions avec un expert du jeu, nous enrichissons le contexte de fouille et la manière d'expliquer une anomalie. Nous avons (i) un ensemble de propriétés qui indiquent les adversaires

2. Le *soul ring* `http://www.dotabuff.com/heroes/dark-seer/items`
3. http ://www.dotabuff.com/heroes/dark-seer/builds

#	Emergence	Support	Motif
1	-0.199	0.1	$\{dire, no\_soul\_ring\_item,$ $comp\_2\_at\_level\_1, comp\_2\_at\_level\_3, comp\_4\_at\_level\_8,$ $comp\_1\_at\_level\_10, comp\_4\_at\_level\_11, comp\_3\_at\_level\_14\}$
2			$\{dire, no\_soul\_ring\_item,$ $comp\_2\_at\_level\_3, comp\_2\_at\_level\_7, comp\_4\_at\_level\_8,$ $comp\_1\_at\_level\_9, comp\_1\_at\_level\_10, comp\_4\_at\_level\_11,$ $comp\_3\_at\_level\_14\}$
3			$\{dire, no\_soul\_ring\_item,$ $comp\_2\_at\_level\_1, comp\_3\_at\_level\_2, comp\_2\_at\_level\_3,$ $comp\_2\_at\_level\_5, comp\_1\_at\_level\_6, comp\_4\_at\_level\_8,$ $comp\_1\_at\_level\_9, comp\_4\_at\_level\_11, comp\_3\_at\_level\_14\}$

TAB. 1 – *3 motifs, avec* $\theta = 30\%$ *et* $min\_sup = 1\%$.

#	Emergence	Support	Motif
1	-0.11	0.018	$\{creeps\_killed\_between\_20\_and\_30, no\_comp\_4\_level\_6\}$
2	-0.14	0.014	$\{enemy\_lifestealer, enemy\_keeperofthelight, no\_comp\_4\_level\_6\}$
3	-0.16	0.024	$\{enemy\_riki, no\_comp\_4\_level\_6\}$
4	-0.19	0.01	$\{no\_comp\_4\_level\_6, > \_40\_dire\_fountain\}$
5			$\{no\_comp\_4\_level\_6, > \_40\_radiance\_fountain\}$
6			$\{enemy\_medusa, no\_comp\_4\_level\_6\}$
7			$\{enemy\_chen, enemy\_gyrocopter\}$
8			$\{enemy\_queenofpain, enemy\_gyrocopter\}$
9			$\{enemy\_bountyhunter, no\_comp\_4\_level\_6, no\_dire\_fort\}$

TAB. 2 – *9 motifs, avec* $\theta = 22\%$ *et* $min\_sup = 1\%$.

du héros dans la partie (chaque héros a des avantages/inconvénients face aux autres), (ii) le nombre de sbires (*creeps*) éliminés par le héros (les sbires apparaissent régulièrement sur le terrain, ne sont pas contrôlables, mais les éliminer rapporte de l'expérience au joueur), (iii) la réponse à *la compétence 4 est-elle prise au plus tôt ?* et enfin (iv) quand on observe que le nombre de passages aux bases est supérieur à un seuil anormalement haut (c'est-à-dire que le joueur est revenu souvent à la base pour se régénérer, la moyenne étant de 20 passages). Nous avons mis le seuil d'anomalie à 22% et le support minimum à 1% et nous obtenons 193026 motifs fréquents. Nous remarquons là encore qu'avec un seuil d'anormalie approprié le nombre de motifs émergents en sortie est très faible puisque nous obtenons 16 motifs émergents dont 9 sont présentés dans la table 2. On constate que parmi les héros adversaires du *Dark Seer* ceux-ci ont souvent gagné contre lui dans la plupart des matchs [4]. Le motif de la ligne 1 possède une propriété qui indique un très faible nombre de sbires tués (entre 20 et 30 alors que la moyenne est de 100 à 200). On observe alors plusieurs types d'anomalies : la première qui porte sur le nombre de $sbires$ tué en co-occurrence avec la compétence 4 qui n'est pas prise niveau 6. Les motifs de la ligne 4 et 5 concernent des joueurs étant souvent revenus à leur base et le motif 9 permet de voir que des joueurs n'ont pas réussi à pénétrer dans la base ennemie (no_dire_fort). Notre méthode permet de caractériser différents types d'anomalies à partir d'une même classification des traces.

3.4 Sur le choix des seuils

Nous utilisons deux paramètres : un support minimum min_sup ainsi qu'un seuil d'anomalie θ. Nous cherchons les motifs qui maximisent à la fois le support et le score d'anomalie, deux mesures antinomiques. Le choix du seuil d'anomalie θ impacte fortement le résultat. Il

4. Voir la page `www.dotabuff.com/heroes/dark-seer/matchups`

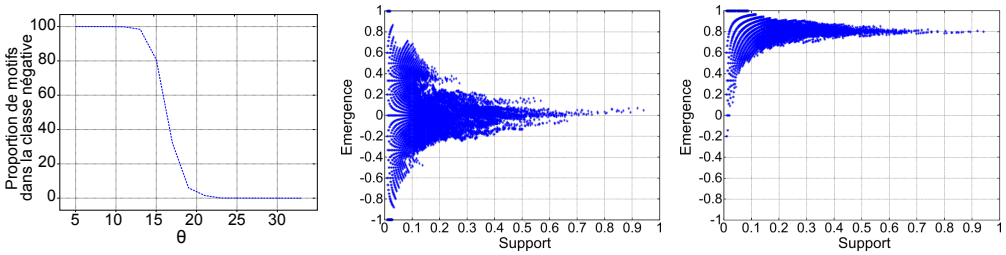


FIG. 4 – *Nombre de motifs émergents en fonction de θ à gauche, et distribution des motifs selon leur support/score d'anomalie pour $\theta = 0.16$ (milieu) et $\theta = 0.24$ (droite)*

s'agit du pourcentage de transitions qui ne respectent pas le modèle : avec $\theta = 0$ la moindre transition non licite rend la trace anormale. La Figure 4 montre le pourcentage de traces assignées à la classe anormale en fonction de θ : augmenter θ c'est être plus permissif, il en résulte moins de motifs. En faisant varier θ, on peut trouver les descriptions d'anomalies les plus flagrantes. On voit que l'équilibre est atteint vers $\theta = 0.16$. Si on affiche un nuage de points des motifs avec leur valeur de support et leur mesure d'anomalie normalisée dans $[-1, +1]$, avec -1 quand le motif n'apparait que dans des instances de la classe négative, l'influence du seuil est encore plus évidente. Sur la figure du milieu, avec $\theta = 0.16$ on voit que la distribution est centrée autour de 0. En augmentant le seuil θ la distribution est majoritairement dans la classe positive et très peu de motifs sont caractéristiques d'anomalies.

4 État de l'art

La détection d'anomalies dans les données a donné lieu a une myriade de travaux ces dernières années comme rappelé par Chandola et al. (2009) et Aggarwal (2013). La tâche réside principalement à détecter un petit groupe d'objets qui est significativement différent du reste de la base de données. De nombreuses mesures pour calculer cette différence ont été données, basées sur la distance (Ramaswamy et al. (2000) et He et al. (2003)), la densité (Breunig et al. (2000)) ou encore l'angle (Kriegel et al. (2008)). Cependant, très peu de travaux cherchent à expliquer les causes de ces anomalies. Des méthodes ont été proposées récemment comme par Keller et al. (2012) pour sélectionner les sous-espaces où il existe des anomalies avec de hautes déviations. La méthode de Duan et al. (2015) cherche à extraire les sous-espaces basés sur un objet unique donné en entrée. Tang et al. (2013) cherchent à découvrir, dans des données catégorielles, des motifs multidimensionnels anormaux dont un attribut et son support varient fortement par rapport un motif de référence similaire. Nous nous distinguons de ces méthodes car nous proposons de quantifier le degré d'anormalité par une mesure basée sur un modèle expert. De plus, nous intégrons la connaissance contextuelle des anomalies dans le codage même des propriétés ce qui permet de proposer une méthode générique pour décrire de plusieurs façons des ensembles d'objets et non un seul ensemble d'objets ou un unique objet choisis en entrée. La fouille de motifs grâce à des modèles experts a été peu étudié. Jaroszewicz et al. (2009) ont proposé une méthode efficace de fouille d'itemsets basée sur un réseau bayésien qui encodent les comportements attendus. Dans Flouvat et al. (2014), les auteurs utilisent un modèle mathématique pour définir une contrainte d'élagage qui réduit fortement le nombre de motifs tout en maximisant leur pertinence. La recherche d'anomalies à

l'aide de l'exploitation de la connaissance experte a été proposée par Angiulli et Fassetti (2014) en caractérisant exclusivement des instances négatives par un ensemble de règles en utilisant la programmation logique. La différence est que nous intégrons un degré de liberté dans la séparation des instances et que nous cherchons à détecter des tendances anormales fréquentes.

5 Conclusion

Dans de nombreuses applications, la détection et la caractérisation d'anomalie sont importantes. C'est par exemple le cas de l'analyse de fraudes au cours de la distribution de produits manufacturés traçables (contrefaçons par duplication d'objets, détournement de produits ou incohérence des circuits de distribution). Nous proposons une nouvelle méthode qui utilise une connaissance experte codée sous la forme d'un graphe pour caractériser des traces normales et anormales, et qui contextualise les anomalies grâce aux comportements modélisées par les traces. Nos expériences sur des données de jeux vidéo choisies du fait d'une confidentialité stricte sur le cas d'étude industriel qui motive cette recherche, ont démontré l'efficacité de notre méthode pour extraire des motifs pertinents grâce à une mesure d'émergence. Dans les deux cas présentés, nous montrons qu'en travaillant avec des propriétés diverses, nous trouvons des motifs pertinents décrivant plusieurs erreurs de stratégie. Ces premiers résultats sont encourageants concernant l'utilité des modèles experts et des propriétés expertes pour la découverte des anomalies. De nombreuses pistes restent à explorer comme l'utilisation de graphes attribués, la proposition de mesures adaptées, ou encore l'adaptation à des contextes non booléen.

Remerciements. Nous remercions Rob Jackson pour nous avoir fourni le jeu de donnée. Cette recherche a été en partie financée par le Projet FUI AAP 14 Tracaverre 2012-2016.

Références

Aggarwal, C. C. (2013). *Outlier Analysis*. Springer.

Angiulli, F. et F. Fassetti (2014). Exploiting domain knowledge to detect outliers. *Data Mining and Knowledge Discovery 28*(2), 519–568.

Bosc, G., M. Kaytoue, C. Raïssi, et J.-F. Boulicaut (2014). Fouille de motifs séquentiels pour l'élicitation de stratégies à partir de traces d'interactions entre agents en compétition. In *EGC'14*, pp. 359–370.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, et J. Sander (2000). Lof : Identifying density-based local outliers. *SIGMOD Record 29*(2), 93–104.

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

Dong, G. et J. Li (1999). Efficient mining of emerging patterns : Discovering trends and differences. In *SIGKDD'99*, pp. 43–52.

Duan, L., G. Tang, J. Pei, J. Bailey, A. Campbell, et C. Tang (2015). Mining outlying aspects on numeric data. *Data Mining and Knowledge Discovery 29*(5), 1116–1151.

Flouvat, F., J. Sanhes, C. Pasquier, N. Selmaoui, et J.-F. Boulicaut (2014). Improving pattern discovery relevancy by deriving constraints from expert models. In *ECAI'14*, pp. 327–332.

Ganter, B. et S. O. Kuznetsov (2001). Pattern structures and their projections. In *Conceptual Structures : Broadening the Base*, pp. 129–142.

He, Z., X. Xu, et S. Deng (2003). Discovering cluster based local outliers. *Pattern Recognition Letters 2003*, 9–10.

Jaroszewicz, S., T. Scheffer, et D. A. Simovici (2009). Scalable pattern mining with bayesian networks as background knowledge. *Data Mining and Knowledge Discovery 18*(1), 56–100.

Keller, F., E. Muller, et K. Bohm (2012). Hics : High contrast subspaces for density-based outlier ranking. In *ICDE'12*, pp. 1037–1048.

Kriegel, H.-P., M. S. Hubert, et A. Zimek (2008). Angle-based outlier detection in high-dimensional data. In *SIGKDD'08*, pp. 444–452.

Novak, P. K., N. Lavrač, et G. I. Webb (2009). Supervised descriptive rule discovery : A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research 10*, 377–403.

Plantevit, M. et B. Crémilleux (2009). Condensed representation of sequential patterns according to frequency-based measures. In *IDA'09*, pp. 155–166.

Ramaswamy, S., R. Rastogi, et K. Shim (2000). Efficient algorithms for mining outliers from large data sets. *SIGMOD Rec. 29*(2), 427–438.

Tang, G., J. Bailey, J. Pei, et G. Dong (2013). Mining multidimensional contextual outliers from categorical relational data. In *SSDBM'13*, pp. 43 :1–43 :4.

Zaki, M. et C.-J. Hsiao (2005). Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Transactions on Knowledge and Data Engineering 17*(4), 462–478.

Summary

The problem of anomaly detection has been deeply investigated over the last past years, however, a few method only enable to understand or contextualize the detected anomalies. In this article, we present a method rooted in pattern mining and supervised descriptive rule discovery that allows to jointly discover anomalies, the strenth of the anomaly and their explanation. Given a set of objects, Given a model of the data, either manually or automatically built from the data, the key idea is to find descriptions of objects that do not respect the global model. We experiment our approach with success on behavioral data where a contextualize anomaly explains the reason of why a player achieves badly.