Labellisation semi-supervisée de données : Vers une approche experte étendue

Laure Crochepierre*, Antoine Marot*, Vincent Barbesant*,
Benjamin Donnot*,**, Lydia Boudjeloud-Assala ***

*RTE R&D {prénom.nom@rte-france.com}, ** INRIA,
*** Université de Lorraine, CNRS, LORIA, F-57000 Metz
lydia.boudjeloud@loria.fr

Résumé. Dans cet article, nous proposons une nouvelle approche semi supervisée de labellisation des événements du réseau électrique français. Après une première labellisation partielle par un système expert, nous utilisons un réseau de neurones siamois pour explorer et étendre les labels sur des données non-labellisés. En appliquant notre approche aux données du système électrique de la région de Lyon sur l'année 2017, les résultats de la métrique créée par le réseau approchent ceux obtenus sur la DTW et nous ouvrent la possibilité d'extension à de plus gros volumes de données à labelliser, tout en intégrant une expertise opérationnelle.

1 Introduction

L'apprentissage automatique a indéniablement progressé au cours de ces dernières années, surpassant dans de nombreux cas les systèmes experts reposant sur une logique explicite. Ceci, grâce à des systèmes entraînés sur de nombreuses observations à reproduire des décisions d'experts humains. De nos jours, c'est majoritairement par des méthodes d'apprentissage supervisé que l'on cherche à répondre à ces problématiques. Les entreprises comme RTE ("Réseau de Transport d'Electricité") reposant de plus en plus sur le numérique, se trouvent surchargées de colossaux volumes de données au format hétérogènes issues de sources variées, telles que les capteurs chargés de surveiller l'état du réseau. Le cadre décrit semble idéal pour l'utilisation d'algorithmes d'apprentissage automatique. Cependant, la majorité des informations captées proviennent des systèmes de contrôle en temps réel et ne sont pas labellisées lors de leur acquisition. En particulier, lorsqu'une action est menée sur le réseau (par exemple pour ré-aiguiller les flux électriques), cette intention n'est pas renseignée dans les bases de données, d'où l'absence de label sur les prises de décision. Cet historique détaillé constitue alors un potentiel inexploité dont l'utilisation directe est impossible pour des méthodes de machine learning supervisé dans le but d'apprendre de ces décisions. Le goulot d'étranglement ne réside donc plus aujourd'hui dans l'ingénierie des algorithmes de machine learning, mais dans la création d'ensembles de données labellisées suffisamment grands pour permettre un apprentissage performant. Dans le cadre de l'optimisation de l'exploitation de ses réseaux électriques, RTE souhaiterait en effet pouvoir assister les opérateurs de la conduite des réseaux (appelés *dispatchers*) en imitant leurs actions par des algorithmes de machine learning supervisé puis par

renforcement. Dans cette optique, elle a besoin au préalable de labelliser son historique d'actions a posteriori par des informations relatives au contexte des événements, avant de pouvoir l'utiliser comme ensemble d'apprentissage. Pour ce faire, une première approche *contrefactuelle* (Donnot et al., 2017) a été proposée, combinant simulation et logique. Une portion des actions prises a ainsi pu être labellisée avec un bon degré de certitude. Cependant, pour pouvoir envisager de les utiliser pour de l'apprentissage supervisé, une plus grande quantité de labels, portant une description plus fine des événements du réseau, est nécessaire.

En complément de ces premiers résultats, notre papier cherche à proposer une nouvelle méthode de labellisation plus systématique par une approche semi-supervisée intégrant un minimum de connaissance opérationnelle. Nous la comparerons dans un premier temps aux résultats d'une approche purement non-supervisée, puis nous chercherons à en montrer les performances en restreignant la quantité de données labellisées pour illustrer son applicabilité dans un cadre opérationnel avec peu de labels disponibles. L'article est structuré de la manière suivante : après une première présentation des données (section 2) ainsi qu'un état de l'art (section 3), nous décrirons les résultats des deux approches non supervisée et semi-supervisée. La section 4.1 fournit à la fois une description formelle et les résultats de plusieurs méthodes non-supervisées. Une nouvelle approche semi-supervisée sera ensuite présentée dans la section 4.2. La section 5 présente enfin les conclusions et perspectives de ces travaux.

2 Présentation des données

2.1 Création d'un format de données

Afin de reconstituer la chronique des événements ayant eu lieu sur le réseau électrique, nous avons choisi de travailler sur l'historique des changements d'état des composants commandables du réseau. Chaque changement d'état est mesuré sous la forme d'un signal binaire discret, par des capteurs situés dans les postes électriques. Ils décrivent l'évolution temporelle de la connectivité entre les lignes du réseau sous la forme d'un ensemble d'événements, assimilables à des séquences d'actions. Un signal d'action unitaire est séparé du suivant par un intervalle de temps non-constant et correspond aux mesures topologiques de connexions et déconnexions entre les lignes électriques. Ce sont ces signaux qui permettent aux gestionnaires de la conduite de connaître le maillage du réseau en temps réel. En donnant une modélisation plus mathématique à ces données, à chaque instant i a lieu une action binaire a_i, associée à une connexion (1) ou déconnexion (0) entre deux points du réseau reliables électriquement par des lignes électriques. Chaque action a_i est caractérisée par 22 variables séparables en 4 grandes catégories tel que $a_i = (t_i, \delta p_i, \delta t_i, c_i)$ avec :

- t_i l'horodate de la mesure du changement de position,
- $\delta p_i \in \{0, 1\}$ la valeur du changement de position binaire,
- δt_i la durée de persistence du changement de position,
- c_i les caractéristiques du capteur d'où est issue la mesure (nature de l'objet mesuré et du groupement de capteurs auquel il appartient).

Afin de pouvoir exploiter ces données hétérogènes, nous avons été amenés à créer un format de données permettant de capter la cohérence spatio-temporelle entre les actions. Cependant, du fait de l'irrégularité de l'écart temporel entre deux instants successifs i et $i + 1$ (deux mesures successives sur un même capteur pouvant avoir lieu à la même seconde comme à

plusieurs jours d'intervalle), il n'était pas judicieux d'exploiter les séries temporelles à pas constant à partir de ces données, mais plutôt de considérer les actions comme des évènements ponctuels. Nous avons ainsi choisi de regrouper les actions a_i en un ensemble de séquences $S = (a_i)_{i \in N}$ d'évènements, de telle manière à ce que chaque séquence corresponde à une journée et un groupement de capteurs. Les regroupements spatiaux choisis ont été guidés et validés par des experts du fonctionnement des réseaux électriques et sont de l'échelle d'un ouvrage électrique, tel qu'une ligne électrique. Le choix de la maille temporelle d'un jour par séquence est quant à lui la conséquence de l'échelle des informations fournies par le personnel opérationnel. Ces informations, relatives aux évènements du réseau, ont en effet vocation à être utilisées par la suite pour la validation des résultats obtenus avec les différentes approches. Il est donc nécessaire qu'elles aient une échelle temporelle comparable à celle de nos données. Cependant, celles-ci ont été enregistrées avec une résolution journalière et nous empêchent de valider des résultats obtenus avec une granularité inférieure à un jour. Cette première étape de séquencement nous a finalement permis de regrouper 103075 actions enregistrées sur l'année 2017 dans la région de Lyon en 40486 séquences, chacune correspondant à une zone spatio-temporelle distincte.

En s'inspirant de la représentation de Ordóñez et Roggen (2016), nous pouvons alors représenter les séquences temporelles multivariées étudiées comme des images avec une colonne par pas de temps et une variable par ligne, comme présenté dans la figure 1.

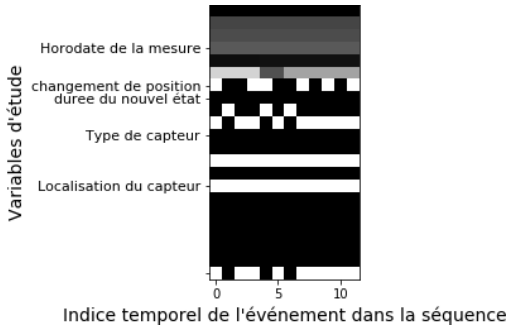


FIG. 1: Séquence temporelle multivariée. Chaque pas de temps correspond à la description détaillée du changement d'état d'un capteur dans un groupement de capteurs.

2.2 Labellisation partielle des données

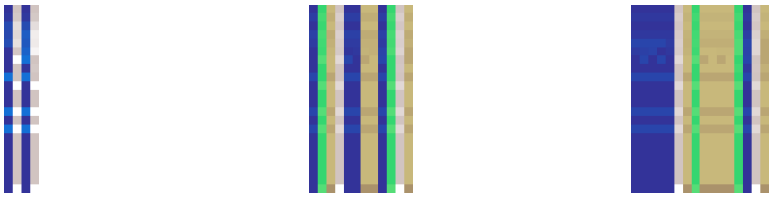
En parallèle de la création des séquences, nous avons également cherché à découvrir une partie des labels des actions opérateurs grâce à un système expert, sans avoir recours à une labellisation manuelle. Les missions de RTE ayant trait à l'exploitation efficace et sûre du réseau électrique, à sa maintenance et son développement, différentes catégories peuvent ainsi être attribués aux actions des opérateurs en lien avec ces missions. Dans la suite de cet article, nous avons choisi de nous focaliser sur trois classes d'opérations récurrentes du réseau électrique (notées par la suite A, B et C) :

— La classe A contient les consignations qui correspondent aux évènements de maintenance d'ouvrages du réseau électrique.

— La classe B représente les évènements de fiabilisation du matériel en exploitation, appelés manoeuvres périodiques.
— La classe C représente les évènements lors desquels une manoeuvre périodique est mise en oeuvre simultanément avec une consignation.

L'intérêt d'étudier prioritairement ces trois catégories est double. Tout d'abord, ces séquences d'actions binaires possèdent une régularité apparente, à la fois liée à leur nature opérationnelle et aux règles qui régissent la conduite du réseau, ce qui rend facilement envisageable leur labellisation quasi-automatique. De plus, ces opérations-ci sont tracées quotidiennement par les opérateurs, ce qui nous permettra par la suite une validation de notre labellisation.

La figure 2 représente des prototypes d'événements selon la même représentation que celle de la figure 1. Chaque couleur correspond aux modifications relatives à un même capteur dans la séquence. En repérant l'ordre caractéristique d'apparition de chaque couleur (c'est-à-dire de chaque capteur) dans la figure 2b, il semble à priori possible reconnaître les d'événements de type B par la valeur et la durée du changement de position.



(a) Événement A (b) Événement B (c) Événement C

FIG. 2: Séquence caractéristique d'événement de type A, B et C. Chaque couleur représente un capteur différent présent dans la séquence.

Afin d'approfondir cette intuition, nous avons cherché un ensemble de règles expertes discriminantes qui permettraient de découvrir les trois catégories d'événements. Ces règles ont été choisies de manière itérative en concertation avec des dispatchers et sont définies sous forme de contraintes logiques utilisant, à chaque pas de temps, un sous-ensemble des 22 variables d'étude. Une fois les règles appliquées, nous avons ensuite validé les labels obtenus grâce aux listes fournies par le personnel opérationnel relatives aux événements A, B et C en 2017. Les résultats obtenus sont synthétisés dans le tableau 1.

		validation métier			
		A (%)	B (%)	C (%)	inconnu (%)
labels issus du système expert	A	$7,5$	$0,5$	$2,4$	$3,4$
	B	$0,5$	$8,4$	$1,9$	$0,01$
	C	$0,2$	$0,1$	$1,2$	$0,09$
	inconnu	$13,8$	13	4	43

TAB. 1: Matrices de confusion de l'approche par système experte, calculée en pourcentages sur l'ensemble des séquences à labelliser soit 40486 séquences.

Nous avons ainsi pu valider 6907 séquences réparties respectivement en 3044, 3389 et 474 séquences de classe A, B et C. Cependant 80% restent non-labellisées et ceci pour deux causes.

Tout d'abord, l'application de règles logiques a rapidement montré ses limites car il s'avère difficile d'intégrer des règles expertes supplémentaires. Trop restrictives, celles-ci diminuent en effet le nombre de séquences détectées. Ensuite, les classes possibles d'événements n'ont pas encore toutes été identifiées et il reste ainsi 60% des séquences de classe non connue. Pour parvenir à labelliser l'intégralité de l'historique français, il parait nécessaire d'envisager d'autres méthodes, en s'appuyant sur le sous ensemble de données labellisées par le système expert pour développer dans un premier temps un processus validable sur les données labellisées.

3 Etat de l'art

L'apprentissage automatique ou *machine learning* a vocation à construire des algorithmes qui s'améliorent automatiquement avec l'expérience (Jordan et Mitchell, 2015). En apprentissage supervisé, cette expérience s'acquiert à partir de données labellisées, sous réserve d'en posséder suffisamment. Aujourd'hui, le manque de jeux de données labellisées est devenu le principal frein au développement de l'apprentissage automatique supervisé dans des domaines tels que l'imagerie médicale (Shie et al., 2015), où l'obtention de labels est un processus manuel coûteux et chronophage nécessitant une expertise métier. Dans les paragraphes suivants nous présenterons deux solutions possibles pour palier le manque de données labellisées.

3.1 Classification non supervisée

Lorsqu'aucun label n'est disponible, la classification non-supervisée (ou *clustering*) analyse la structure des données pour organiser les données par regroupements, ou *clusters*, de telle manière à ce que chaque élément soit plus similaire aux éléments du même cluster qu'ils ne le sont vis-à-vis des éléments des autres clusters (Grira et al., 2005). Un label majoritaire peut alors être attribué à chaque cluster pour labelliser l'ensemble de ses exemples. Ce type de clustering nécessite à la fois de définir une mesure de similarité entre les objets et un algorithme de classification. Pour analyser des séquences temporelles, l'une des mesures de dissimilarité les plus utilisées à ce jour est la mesure de Dynamic Time Warping (DTW) (Sakoe et Chiba, 1978). Elle permet notamment de prendre en compte des déformations temporelles, ce qui n'est pas possible avec une distance euclidienne classique. Cependant, bien que de nombreuses optimisations aient été développées (Lemire, 2009; Cuturi et Blondel, 2017), elle reste très coûteuse en temps de calcul. En parallèle du choix de la distance vient le choix de l'algorithme de classification non-supervisé. Ceux-ci appartiennent principalement aux deux catégories suivantes. Les méthodes hiérarchiques (Ward Jr, 1963; Balcan et al., 2014) et les méthodes par partitions (MacQueen et al., 1967) (Kaufman et Rousseeuw, 2009). Le nombre de partitions doit être définie au préalable par des critères tels que la silhouette (Rousseeuw, 1987) ou l'indice de Calinski-Harabasz (Caliński et Harabasz, 1974).

3.2 Classification semi-supervisée

A mi-chemin entre les approches supervisée et non-supervisée, l'apprentissage semi supervisé a vu le jour. Cette nouvelle typologie de méthodes est principalement utilisée lorsqu'une faible quantité de labels est disponible, par exemple sous forme de contraintes sur des

paires d'objets (must-link, cannot-link) dans COP-KMeans (Wagstaff et al., 2001) et PCK-means (Basu et al., 2004). Au lieu de simplement utiliser ces informations pour valider les résultats du clustering, celles-ci servent ici à guider la classification. L'approche semi-supervisée est relativement récente. Elle se décompose principalement selon deux axes définis par Grira et al. (2005). Le premier axe correspond aux méthodes d'adaptation de la mesure de similarité (Karasuyama et Mamitsuka, 2013; Reddy et al., 2016), alors que le second modifie l'algorithme de clustering en prenant en compte des labels ou des contraintes entre les entrées, lors de l'initialisation (Basu et al., 2002) ou en vérifiant les contraintes lors de la mise à jour des clusters (Pelleg et Baras, 2007). De nouvelles méthodes semi-supervisées explorent également l'utilisation de réseaux de neurones profonds en s'inspirant des deux axes présentés précédemment. Des réseaux tels que SEVEN (Noroozi et al., 2017) ou DIRECT (Bahaadini et al., 2018) permettent à la fois de créer une métrique de similarité entre les entrées en sortie du réseau, mais également d'extraire des caractéristiques d'une structure complexe afin d'obtenir une nouvelle représentation des données. Pour ce faire, une architecture inspirée des réseaux siamois (Bromley et al., 1994) est utilisée avec une fonction objective particulière appelée *contrastive loss* (Hadsell et al., 2006) qui rapproche deux objets similaires et éloigne deux objets dissimilaires dans un *espace de projection discriminant* créé par le réseau.

4 Approches proposées

4.1 Approche non-supervisée

Protocole expérimental Comme nous l'avons identifié dans la partie 2.1, nous ne connaissons pas les labels de la majorité des séquences. Afin de minimiser l'effort de labellisation, une première possibilité serait de s'appuyer sur la structure des données pour les regrouper en clusters. Le label attribué au cluster serait alors le label majoritaire parmi ceux connus dans le cluster. Cette approche étant possible sous réserve de la cohérence structurelle au sein des classes, nous allons l'éprouver dans un premier temps sur les séquences dont les labels ont été validés, afin de tenter de retrouver les 3 classes connues.

En s'inspirant du protocole décrit dans (Sardá-Espinosa, 2017), nous avons cherché dans un premier temps à comparer les performances des algorithmes de classification sur les 6907 séquences multivariées préalablement labellisées comme décrit dans la partie 2. A ces fins, nous avons utilisé la mesure de DTW comme mesure de similarité entre séquences. Les calculs ont été menés sur différents algorithmes de clustering tel que la Classification Ascendante Hiérarchique (CAH) (S. Michalski et E. Stepp, 1983) et les K-Medoids (MacQueen et al., 1967) en retenant deux critères pour la validation des clusters obtenus : l'indice de Calinski Harabaz (Caliński et Harabasz, 1974) et la silhouette (Rousseeuw, 1987).

Résultats Une analyse croisée des scores des deux critères de validation à l'optimum, calculés pour différents algorithmes, nous a permis d'identifier 3 clusters. Du fait du temps de calcul élevé de la mesure de DTW (en 17h en parallélisant sur un processeur i7 2,8 GHz x 8 coeurs avec 16Go de mémoire), les résultats présentés par la suite ont été obtenus en calculant au préalable la matrice de similarité sur l'ensemble des séquences.

Pour approfondir plus en détail la compréhension des clusters obtenus, nous avons procédé à une vérification de la répartition des classes A, B et C dans les différents clusters trou-

vés par les deux algorithmes, en attribuant le label majoritaire au cluster. Sur la classification correspondante, nous obtenons les scores les plus élevés pour l'algorithme de la CAH avec respectivement un score de 0.53 pour le critère de Ward et 0.49 avec le lien simple.

		validation métier		
		A	B	C
prédiction	A	$39,8$	$35,8$	$3,9$
	B	$4,2$	$13,4$	$2,9$
	C	0	0	0

TAB. 2: Matrices de confusion de l'approche non supervisée par méthode de Ward. Les résultats ont été calculés en pourcentages sur 6907 séquences.

Dans le tableau 2, la matrice de confusion du meilleur score, obtenu avec le clustering de Ward, nous indique que ces algorithmes de clustering ne permettent pas de dissocier les classes d'événements A/B et que la classe C n'est jamais prédite. Ces éléments semblent indiquer que les méthodes de clustering non supervisé ne sont pas adaptées à la découverte de labels pour le type de données dont nous disposons. L'approche non-supervisée a ainsi montré ses limites à la fois par la difficulté de composer avec la structure complexe des données, par le temps élevé du calcul de la DTW ainsi que par la faiblesse des résultats obtenus. De plus, bien que nous l'ayons appliquée à un sous ensemble de nos données composé de 6907 séquences, l'objectif final est de pouvoir l'utiliser sur l'ensemble de données soit 40486 séquences. Or, le passage à l'échelle pour un aussi gros volume de données sera difficile. En somme, ne pas tirer parti de l'expertise métier décrite dans la partie 2.2 parait être une contrainte trop forte et nous allons donc explorer par la suite un protocole semi-supervisé offrant une plus grande flexibilité et mieux adaptée à notre problème.

4.2 Approche semi-supervisée ou "experte étendue"

4.2.1 Méthode semi-supervisée proposée

Nos données présentant un format hétérogène, aussi bien dans la nature des variables que dans la longueur des séquences associées, la définition explicite d'une mesure de similarité adaptée et calculable avec un budget computationnel limité s'avère être une tâche complexe. L'approche que nous allons maintenant présenter s'appuie sur un processus d'apprentissage : plutôt que de définir une mesure, il est possible de l'apprendre spécifiquement pour un problème donné à partir d'exemples sélectionnés. Cet apprentissage permet en outre d'intégrer une expertise opérationnelle sur des labels existants et d'exploiter les similarités dans des séquences. Inspirés par les récentes performances atteintes par les réseaux de neurones siamois (Bromley et al., 1994) dans les tâches de clustering, notamment pour l'exploration et la découverte de classes inconnues (Bahaadini et al., 2018), et lorsque peu d'exemples de chaque classe sont disponibles grâce à un apprentissage par paires (Koch, 2015), nous avons choisi d'explorer leur utilisabilité sur nos données. Ceux-ci permettent d'exploiter les caractéristiques de la structure initialement complexe des données d'entrée, pour conjointement créer leur représentation vectorielle et une métrique de similarité entre les objets présentés. La méthode de labellisation faiblement supervisée que nous proposons se décompose alors en deux étapes : l'entraînement d'un réseau siamois puis l'exploitation de la projection réalisée par celui-ci.

Construction d'un réseau siamois Tout d'abord, un réseau de neurones siamois est entraîné à partir de paires de séquences S_1 et S_2 labellisées, chaque séquence pouvant appartenir à la classe A, B ou C : le label de la paire, utilisé en sortie du réseau, valant 0 pour deux séquences de même classe et 1 sinon. Après entraînement, une fonction non linéaire de projection P est alors créée. Elle permet de comparer les deux séquences projetées $P(S_1)$ et $P(S_2)$ directement via une norme choisie avant à l'entraînement (L_2 dans notre cas). La mesure en sortie est alors minimale pour deux séquences de même classe et maximale sinon.

L'architecture du réseau siamois se décompose en deux réseaux "jumeaux" de poids identiques mais prenant en entrée une séquence distincte. Pour la partie jumelle, nous avons choisi un réseau convolutif à 3 couches suivi de 3 couches denses, afin d'obtenir une projection sous forme d'un vecteur de taille 32. La taille de l'espace de projection a été choisi pour être la taille minimale permettant la convergence de l'entraînement du réseau. Les filtres convolutifs sont de taille (nombre de variables x taille caractéristique d'une fenêtre temporelle) avec une fenêtre temporelle de 10, assimilable à une fenêtre glissante appliquée sur la dimension temporelle de la séquence. Le réseau est entraîné au moyen de la contrastive loss (Hadsell et al., 2006).

Exploration de la projection Une fois entraîné, le réseau permet d'obtenir une projection des données dans un espace facilement explorable par des algorithmes tels que la t-SNE (Maaten et Hinton, 2008) afin de déduire des regroupements sur les données (manuels ou algorithmiques) qui n'auraient pas été envisagés dans l'étude des données sous un format de séquences. Un exemple de projection obtenu est présenté dans la figure 3, illustrant une meilleure séparabilité apparente des données dans l'espace de projection.

(a) similarités avec DTW sur les données brutes (b) similarités apprises sur les données projetées

FIG. 3: Représentation des données réduite par t-SNE. Les couleurs vert, orange et violet représentent respectivement des séquences de type A, B et C.

4.2.2 Expérimentations et résultats

Avant d'induire de nouveaux regroupements à partir de la projection, nous avons souhaité vérifier la cohérence de l'espace obtenu sur des séquences dont nous connaissions la classe. Pour ce faire, nous avons construit différents réseaux, en faisant varier le pourcentage de séquences labellisées utilisées pendant l'entraînement entre 10 et 100%, ceci afin d'estimer les limites de sa capacité d'apprentissage sur peu d'exemples. Les projections issues de ces réseaux seront l'objet des comparaisons présentées dans les prochains paragraphes.

Non-supervisé Dans un premier temps, nous avons choisi d'appliquer les mêmes méthodes non supervisées que celles présentées dans la partie 4.1, cette fois-ci sur les données projetées. L'objectif est désormais de comparer les critères, ainsi que le nombre de clusters optimaux selon le pourcentage de labels, avec ceux trouvés dans la partie 4.1. Les résultats synthétisés dans le tableau 3 sont issus du clustering de Ward qui ici, comme lors du clustering de séquences, s'est révélé être la méthode atteignant les scores les plus élevés.

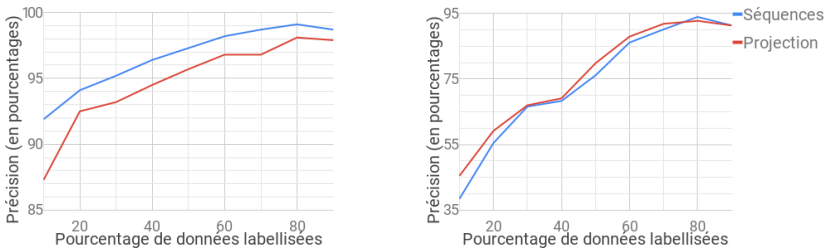
pourcentage de données labellisées	100	90	50	20	10
nombre de clusters	9	8	8	6	4
score de silhouette	$0,6$	$0,6$	$0,56$	$0,56$	$0,55$
score de Calinski	103029	106350	111251	88007	104244
score de l'extension de labels	0,65	0,64	0,63	0,65	0,63

TAB. 3: Clustering hiérarchique par méthode de Ward en fonction du pourcentage de données labellisées utilisées pendant l'entrainement du réseau siamois.

De ces premières expériences, nous mettons en évidence que les indices de silhouette sont à la fois supérieurs à ceux obtenus sur le format de séquences mais également supérieurs à $0,5$. Ceci indique que l'algorithme est capable de déduire des clusters plus "cohérents" de la projection que sur les séquences. De plus, quelque soit le pourcentage de labels utilisés pendant l'entraînement du réseau, le nombre de clusters identifiés sur les projections est systématiquement supérieur aux 3 clusters trouvés dans la partie précédente. Il semble donc que l'algorithme de clustering parvienne à déduire des sous-regroupements à partir des projections créées par le réseau siamois. Enfin, comme nous l'avions fait dans la partie 4.1, nous appliquons les labels majoritaires aux clusters. Les résultats sont ici significativement supérieurs à ceux trouvés précédemment et oscillent entre $0,63$ et $0,65$, indiquant du même fait que la projection soit une représentation des données rendant chaque classe plus homogène.

Semi-supervision Dans cette dernière étape, nous souhaitons estimer les performances de l'extension de labels sur la projection avec la mesure du réseau siamois. Pour cela, nous appliquons l'algorithme des K-plus proches voisins (KNN), aux données qui n'ont pas été utilisées pendant l'entraînement, parallèlement sur le format de séquence et celui de la projection.

Les résultats présentés dans la figure 4 comparent l'application du KNN à 10 voisins, sur les séquences et la projection. Le choix de la classe se fait alors par vote pondéré par la distance. Lorsque le pourcentage de données labellisées diminue, on remarque de manière globale une diminution de la précision des résultats. Cependant, le score moyen représenté dans la figure 4a reste élevé, n'allant pas en deçà de 87%. De même, l'accuracy moyenne de la classification avec la projection reste proche de celle sur le format de séquences. Les performances de la classification sont donc aussi acceptables dans l'espace de projection qu'en travaillant avec des données au format de séquences. Le dernier point intéressant de ces expériences est le résultat d'extension de labels sur la classe C. Il s'agit de la catégorie d'événements la plus difficile à classifier car disposant de moins d'exemples que les classes A et B. Cette complexité de labellisation se traduit par un score de classification globalement plus faible que pour les autres



(a) Score moyen (b) Score de la classe C

FIG. 4: Précision (Accuracy) de la classification en fonction du pourcentage de données labellisées. Les courbes bleue et rouge correspondent respectivement aux scores obtenus sur les séquences et les données projetées. Les mêmes labels sont utilisées entre les classifications des séquences et de la projection.

classes. Cependant, dans ce cadre où peu de labels sont disponibles, les performances sont supérieures en utilisant les données projetées. Ainsi, avec moins de 80% de données labellisées, la classification est systématiquement meilleure dans l'espace de projection, avec un écart plus important lorsque le volume de données labellisées diminue.

En conclusion, les expériences menées ont permis d'identifier un moyen de transformer un format de séquences initialement complexes en un vecteur, rendant possible un passage à l'échelle pour labelliser une plus grande quantité de séquences. Nous avons également pu mettre en évidence plusieurs atouts de cette méthode. Tout d'abord, comme l'indique les résultats de la CAH, la mesure de similarité créée par le réseau de neurones parvient mieux à comparer les événements qui lui sont présentés que la DTW sur le même problème. Enfin, du fait de la faible différence de ces résultats sur des problèmes de classification et les bonnes performances du clustering, il parait possible de poursuivre notre exploration de données en approfondissant des pistes dont nous ne disposions pas sur les séquences seules.

5 Conclusions et perspectives

Dans cet article, nous avons présenté l'enjeu de la labellisation sur les données des actions du réseau électrique français. Après avoir envisagé un système expert trop restrictif, nous avons éprouvé l'apprentissage non-supervisé qui s'est révélé insuffisant pour déduire des regroupements à partir des données. Du fait du temps de calcul très élevé de la DTW, nous avons mis en évidence l'impossibilité d'extension de ces approches pour labelliser l'ensemble des séquences de l'historique. Pour surmonter ces obstacles, nous avons proposé une méthode de projection de données permettant à la fois de définir une métrique personnalisée, mais également de créer une représentation standardisée des séquences grâce à un réseau de neurones siamois. Les résultats obtenus sur de la classification non supervisée et supervisée ont permis de mettre en évidence la capacité du réseau à extrapoler des regroupements à partir de paires de séquences, mais a également prouvé son intérêt lorsque peu de données labellisées sont disponibles.

Cette étape de validation de la projection constitue un premier pas vers la création d'un processus de labellisation appliqué à l'ensemble de l'historique du réseau français et facilitera

à l'avenir l'exploration des données. La nouvelle représentation créée par le réseau conservant les principales caractéristiques des séquences, il sera alors possible de travailler uniquement sur cette projection plutôt que sur le format initial de séquences. Il sera notamment intéressant d'examiner les sous-groupes identifiés par le clustering de la projection pour y découvrir de nouvelles catégories d'évènements. Afin d'identifier de nouveaux évènements et corriger les erreurs de labellisation de manière incrémentale, nous souhaiterions également intégrer les experts du métier dans un processus de labelisation itératif, grâce à une exploration experte de la projection. Enfin, bien que nous ayons choisi une architecture de réseaux de neurones siamois, d'autres architectures telles que les conditionnal variationnal autoencoders pourront être envisagés pour obtenir d'autres types de projections.

Références

Bahaadini, S., V. Noroozi, N. Rohani, S. Coughlin, M. Zevin, et A. K. Katsaggelos (2018). Direct : Deep discriminative embedding for clustering of ligo data. *arXiv preprint arXiv :1805.02296*.

Balcan, M.-F., Y. Liang, et P. Gupta (2014). Robust hierarchical clustering. *The Journal of Machine Learning Research 15*(1), 3831–3871.

Basu, S., A. Banerjee, et R. Mooney (2002). Semi-supervised clustering by seeding. In *In Proceedings of 19th International Conference on Machine Learning (ICML-2002)*. Citeseer.

Basu, S., A. Banerjee, et R. J. Mooney (2004). Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM international conference on data mining*, pp. 333–344. SIAM.

Bromley, J., I. Guyon, Y. LeCun, E. Säckinger, et R. Shah (1994). Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pp. 737–744.

Caliński, T. et J. Harabasz (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods 3*(1), 1–27.

Cuturi, M. et M. Blondel (2017). Soft-dtw : a differentiable loss function for time-series. *arXiv preprint arXiv :1703.01541*.

Donnot, B., I. Guyon, M. Schoenauer, P. Panciatici, et A. Marot (2017). Introducing machine learning for power system operation support. *arXiv preprint arXiv :1709.09527*.

Grira, N., M. Crucianu, et N. Boujemaa (2005). Unsupervised and semi-supervised clustering : a brief survey.

Hadsell, R., S. Chopra, et Y. LeCun (2006). Dimensionality reduction by learning an invariant mapping. In *null*, pp. 1735–1742. IEEE.

Jordan, M. I. et T. M. Mitchell (2015). Machine learning : Trends, perspectives, and prospects. *Science 349*(6245), 255–260.

Karasuyama, M. et H. Mamitsuka (2013). Manifold-based similarity adaptation for label propagation. In *Advances in Neural Information Processing Systems 26*, pp. 1547–1555.

Kaufman, L. et P. J. Rousseeuw (2009). *Finding groups in data : an introduction to cluster analysis*, Volume 344. John Wiley & Sons.

Koch, G. (2015). Siamese neural networks for one-shot image recognition. In *ICML*, Volume 2.

Lemire, D. (2009). Faster retrieval with a two-pass dynamic-time-warping lower bound. *Pattern recognition 42*(9), 2169–2180.

Maaten, L. v. d. et G. Hinton (2008). Visualizing data using t-sne. *Journal of machine learning research 9*(Nov), 2579–2605.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations.

Noroozi, V., L. Zheng, S. Bahaadini, S. Xie, et P. S. Yu (2017). Seven : deep semi-supervised verification networks. *arXiv preprint arXiv :1706.03692*.

Ordóñez, F. J. et D. Roggen (2016). Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors 16*(1), 115.

Pelleg, D. et D. Baras (2007). K-means with large and noisy constraint sets. In *European Conference on Machine Learning*, pp. 674–682. Springer.

Reddy, Y. C. A. P., P. Viswanath, et B. E. Reddy (2016). Semi-supervised single-link clustering method. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1–5.

Rousseeuw, P. J. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20*, 53–65.

S. Michalski, R. et R. E. Stepp (1983). Automated construction of classifications conceptual clustering versus numerical taxonomy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-5*, 396 – 410.

Sakoe, H. et S. Chiba (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing 26*(1), 43–49.

Sardá-Espinosa, A. (2017). Comparing time-series clustering algorithms in r using the dtwclust package. *R package vignette 2*.

Shie, C., C. Chuang, C. Chou, M. Wu, et E. Y. Chang (2015). Transfer representation learning for medical image analysis. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 711–714.

Wagstaff, K., C. Cardie, S. Rogers, S. Schrödl, et al. (2001). Constrained k-means clustering with background knowledge. In *ICML*, Volume 1, pp. 577–584.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association 58*(301), 236–244.

Summary

We propose a novel approach to label unknown events on the French power grid based on an extended expert mehod. After labelling events using logical rules, we trained a siamese neural network in a semi-supervised way to extend these labels. When applying our approach on the data of 2017 in Lyon region, the obtained results with the created metric approach those of the DTW. This method offers us a scalable method, namely applicable to more data, where the integration of expert knowledge is possible.