Classification contextuelle pour les fouilles de données issues de machines-outils

Z. WANG[a,b], M. RITOU[a,b], C. da CUNHA[a,c], F. FURET[a,b]

a. Laboratoire des Sciences du Numérique de Nantes (LS2N, UMR CNRS 6004),
b. Université de Nantes, 2 av. J Rouxel, 44475 Carquefou – France
c. Centrale Nantes, 1 rue de la Noë, 44321 Nantes – France
zhiqiang.wang@univ-nantes.fr mathieu.ritou@univ-nantes.fr
catherine.da-cunha@ec-nantes.fr benoit.furet@univ-nantes.fr

Résumé :

Dans le contexte général de l'Industrie 4.0, une entreprise de fabrication moderne dispose de nombreuses données numériques qui pourraient être utilisées pour rendre les machines-outils plus intelligentes et faciliter la prise de décision en matière de gestion opérationnelle. L'une des premières étapes de l'approche d'exploration de données est la sélection précise de données pertinentes. Pour ce faire, les données brutes doivent être classées dans différents groupes de contexte. Cet article présente un algorithme d'apprentissage automatique non-supervisé, par mélanges gaussiens (GMM), pour la classification contextuelle. Le nombre de clusters optimal est déterminé par le critère BIC (Bayesian Information Criterion), à partir de données venant de l'industrie aéronautique. Les vérifications par fouilles manuelles et la validation par k-fold montrent ensuite que la méthode GMM permet d'obtenir de bons résultats de classification contextuelle.

Mots clefs : Industrie 4.0, machines-outils, apprentissage non-supervisé, classification contextuelle

1 Introduction

L'Usinage à Grande Vitesse (UGV) a fortement augmenté les vitesses de coupe par rapport à l'usinage conventionnel. Les systèmes de production sont très flexibles, surtout dans l'aéronautique, où plusieurs milliers de références de pièce sont usinées par une centaine d'outils sur un même site. Ce n'est pas facile pour le Bureau des Méthodes d'identifier les programmes ou les outils qui posent problèmes. Les principaux sont le broutement (instabilité en usinage), le bris d'outil, les sur-vibrations. Par conséquent, un système de fouille des données en usinage est nécessaire pour protéger la machine-outil et les pièces à usiner (surtout pour les pièces aéronautiques qui sont à forte valeur ajoutée). Des articles d'état de l'art sur la surveillance de l'usinage listent l'ensemble des défauts mesurables et étudiés [1]. Quintana et al. [2] présentent un état de l'art sur le broutement et les méthodes existantes pour le détecter et l'éviter. Zhou et al. [3] présentent une synthèse des méthodes utilisées pour la surveillance de l'usure d'outil dans les processus de fraisage, y compris les capteurs, l'extraction de caractéristiques et les modèles de surveillance. Godreau et al. [4] présentent un critère vibratoire pour la détection de broutement, source de non-qualité.

Dans le contexte général de l'Industrie 4.0, une entreprise de fabrication moderne dispose de nombreuses données numériques. Il est intéressant de les exploiter par fouilles de données, rendre les machines-outils plus intelligentes et faciliter la prise de décision en matière de gestion opérationnelle. Lenz et al. [5] proposent une approche holistique de l'analyse des données des machines-outils afin de combiner les tâches et de regrouper les objectifs d'analyse entre différents départements dans l'entreprise. Morgan et al. [6] présentent la conception et le développement d'un système de surveillance de processus cyber-physiques pour les machines-outils. Liu et al. [7] proposent une méthode de développement systématique pour Machines-outils 4.0. Afin d'extraire les données des bases de données de manière plus efficace et plus pertinente, la classification contextuelle est nécessaire pour bien connaitre les états de notre machine-outil ainsi que pour un calcul plus pertinent d'indicateurs (KPI). Pour la classification contextuelle, des algorithmes d'apprentissage automatiques sont développés.

L'apprentissage automatique (Machine Learning) est un champ d'étude de l'Intelligence Artificielle qui se base sur des approches statistiques et qui peut notamment être appliqué aux signaux de machines-outils. Kim et al. [8] listent et résument les contributions en usinage à l'aide d'algorithmes d'apprentissage automatique. Il est catégorisé par deux modes: apprentissage supervisé et apprentissage non-supervisé :

- L'apprentissage supervisé apprend à classer à partir d'échantillons de sortie déjà étiquetés ; par exemple, les réseaux de neurones, les machines à vecteurs de support, la méthode des plus proches voisins, l'arbre de décision, etc. Il a pour but de faire des prédictions correctes sur des données non présentes dans l'ensemble d'apprentissage.
- L'apprentissage non-supervisé apprend à classer des données non étiquetées. Il existe différents types d'apprentissage non-supervisé : K-means, classification hiérarchique (ascendante ou descendante), des estimations de densité de distribution (ex. modèle de mélange gaussien - GMM). Bhinge et al. [9] construisent un modèle de prédiction d'énergie pour usiner une pièce en appliquant un algorithme de régression gaussienne. Godreau et al. [10] calculent les seuils de critères en appliquant la méthode de l'estimateur du maximum de vraisemblance (Maximum Likelihood Estimation - MLE).

Dans l'industrie, il est généralement difficile de collecter des données de sortie déjà étiquetées. En conséquence, des méthodes d'apprentissage non-supervisé doivent généralement être utilisées pour les applications industrielles.

Parmi toutes les méthodes d'apprentissage non-supervisé, GMM permet de modéliser la fonction de densité de probabilité par la somme pondérée de densités de composantes gaussiennes [11]. Les paramètres de GMM sont estimés à partir des données d'apprentissage à l'aide de l'algorithme espérance-maximisation (EM). Son avantage est que GMM peut correctement classifier les données avec des classes déséquilibrées. Alors que son inconvénient est que les résultats de GMM dépendent directement du choix des paramètres (par ex., le nombre de composantes). C'est pour cela qu'un algorithme complémentaire est nécessaire afin de déterminer automatiquement le nombre de classes nécessaires pour modéliser une variable aléatoire [12]. L'indicateur Bayesian Information Criterion (BIC) est particulièrement recommandé, en association avec les modèles de mélanges gaussiens (GMM). Des travaux de recherches proposent l'utilisation d'algorithmes d'apprentissage directement sur les données brutes pour rendre les machines-outils plus intelligentes. Cependant, peu de recherches se concentrent sur les prétraitements de données pour identifier le contexte, qui permettent une sélection très fine des données d'usinage.

Dans cet article, un algorithme d'apprentissage automatique non-supervisé, par mélange gaussien, est proposé pour la classification contextuelle en usinage. La méthode est appliquée à une base de données d'usinage collectée dans l'industrie aéronautique. Le nombre de classes du modèle gaussien

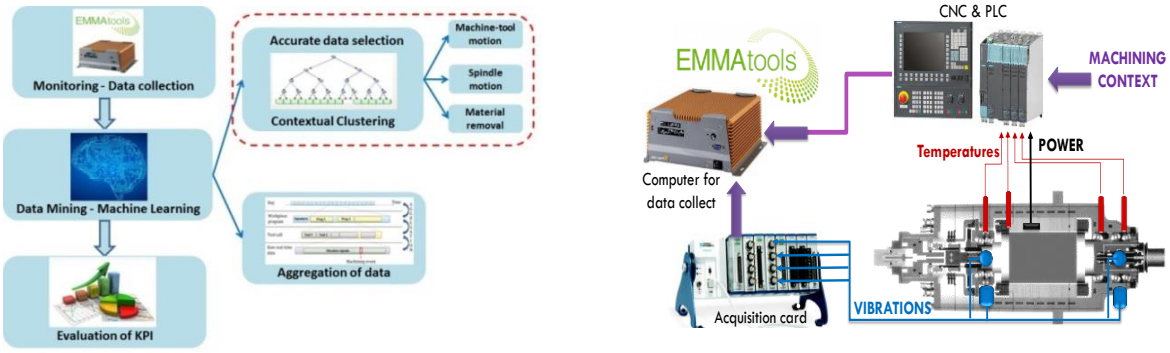
est déterminé par la méthode BIC. Les résultats des classifications sont ensuite évalués selon la méthode k-fold.

2 Démarche d'analyse des données

2.1 Cadre de l'étude

Les travaux présentés dans cet article participent au projet ANR SmartEmma qui vise à développer des machines-outils intelligentes et connectées. La Figure 1 (a) présente l'idée globale de ce projet. Un dispositif, appelé Emmatools, collecte les données mesurées pendant l'usinage et les stocke dans une base de données [13]. Il est installé sur des machines-outils dans des usines de constructeurs aéronautiques. La Figure 1 (b) présente le schéma de principe de l'Emmatools. Il collecte les données chaque dixième de seconde. Les données sont issues de deux sources d'information : la commande numérique et les capteurs ajoutés. La commande numérique fournit des informations de contexte ainsi que certaines données de capteurs déjà présents. Les données de contexte correspondent aux informations telles que l'outil en broche, le programme en cours… Les données de capteurs déjà présents comprennent la vitesse de rotation de la broche, les différentes vitesses d'axe, la puissance instantanée de la broche… Afin de compléter ces mesures, 4 accéléromètres sont intégrés à la broche (radialement à chaque palier).

Les méthodes d'apprentissage automatique vont être développées pour d'une part effectuer des classifications contextuelles, et d'autre part agréger les données. Notons qu'1 Go de données est collecté par jour et par machine-outil. Avant l'exploration de ces données, il est souhaitable de faire une classification contextuelle. En effet, une meilleure sélection de données est préférable pour un calcul plus pertinent d'indicateurs de performance (KPI), réduisant ainsi le bruit. Enfin, des KPI et des méthodes d'analyse des données seront définis pour améliorer l'efficacité du processus UGV. Les KPI ainsi évalués deviennent un support d'aide à la décision pour le pilotage de l'entreprise.



a) b)

Fig.1 (a) Processus de fouille des données d'usinage, (b) Schéma de principe de l'Emmatools

2.2 Méthode de classification contextuelle

La classification contextuelle est nécessaire pour un calcul plus pertinent des indicateurs de performance (KPI). Il est intéressant de savoir :

- Si la broche est arrêtée ou si elle tourne (à vitesse constante ou variante)
- Si la machine-outil est arrêtée ou si elle avance (à vitesse constante ou variante)
- Si l'outil n'usine pas ou s'il usine (avec un engagement de l'outil constant ou variant)

			Vitesse d'avance de la machine-outil Vf					
			Vf=0	**Vf>0**				
				ΔVf > $T_{ΔVf}$		**ΔVf ≤ $T_{ΔVf}$**		
				Arms≤T_{Arms} & P≤P_{vide}	**Arms>T_{Arms} ou P>P_{vide}**		**Arms≤T_{Arms} & P≤P_{vide}**	**Arms>T_{Arms} ou P>P_{vide}**	
					ΔP≤$T_{ΔP}$	**ΔP>$T_{ΔP}$**		**ΔP≤$T_{ΔP}$**	**ΔP>$T_{ΔP}$**
Rotation de la broche N	**N=0**		Broche : A Machine : A Usinage : N	Broche : Arrêtée (A) Machine-outil : Vitesse Variante (VV) Usinage : Non (N)			Broche : Arrêtée (A) Machine-outil : Vitesse Constante (VC) Usinage : Non (N)		
	N>0	**ΔN > $T_{ΔN}$**	Broche : VV Machine : A Usinage : N	Broche : VV Machine : VV Usinage : Non (N)	Broche : VV Machine : VV Usinage : eng. Constant (UC)	Broche : VV Machine : VV Usinage : eng. Variant (UV)	Broche : VV Machine : VC Usinage : Non (N)	Broche : VV Machine : VC Usinage : eng. Constant (UC)	Broche : VV Machine : VC Usinage : eng. Variant (UV)
		ΔN ≤ $T_{ΔN}$	Broche : VC Machine : A Usinage : N	Broche : VC Machine : VV Usinage : N	Broche : VC Machine : VV Usinage : UC	Broche : VC Machine : VV Usinage : UV	Broche : VC Machine : VC Usinage : N	Broche : VC Machine : VC Usinage : UC	Broche : VC Machine : VC Usinage : UV

Tableau 1. Classification des différents états de la machine-outil

En combinant ces trois informations contextuelles, on obtient 17 états potentiels pour la machine-outil. Le Tableau 1 présente ces 17 états avec leurs conditions.

Pour effectuer ces classifications, prenons une variable X (par ex., Vf ou N). La classification va également être effectuée à partir des variations de X, notées ΔX (par ex., ΔVf ou ΔN) qui est défini par la dérivée simple :

$$\text{éq. (1) :}\quad \Delta X = (X_n - X_{n-1}) / \Delta t$$

où X_n est la mesure à l'enregistrement n, X_{n-1} la précédente et Δt est la période d'échantillonnage des données (0,1s). L'objectif est de calculer un seuil de classification $T_{\Delta X}$ par un apprentissage automatique non supervisé. La méthode par mélange gaussien GMM a été retenue dans une étude précédente [14].

Les classifications sont ensuite effectuées à partir de règles métier très simples, formalisées ci-après. Il y a *a priori* trois clusters pour les mouvements de la machine-outil : arrêtée ($Vf=0$), avance à vitesse constante ($Vf>0$ et $\Delta Vf \leq T_{\Delta Vf}$), ou variante ($Vf>0$ et $\Delta Vf > T_{\Delta Vf}$). Ainsi que trois clusters pour la rotation de la broche : arrêtée ($N=0$), tournant à vitesse constante ($N>0$ et $\Delta N \leq T_{\Delta N}$), ou variante ($N>0$ et $\Delta N > T_{\Delta N}$).

On souhaite également savoir si l'outil usine ou pas. Pour cela, il est proposé d'effectuer une classification suivant les vibrations en usinage, par le critère *Arms* (valeur efficace d'accélération, root mean square, en m/s²) et la puissance P (kW) quand la vitesse d'avance Vf n'est pas nulle. En usinage, il est admis que si l'outil usine, sa vibration (*Arms*) est plus grande que si l'outil n'usine pas, il est aussi admis que la puissance (P) va dépasser la puissance de rotation à vide de la broche (P_{vide} quand l'outil n'usine pas). C'est-à-dire, si $P > P_{vide}$ ou $Arms > T_{Arms}$, l'outil usine ; si $P \leq P_{vide}$ et $Arms \leq T_{Arms}$, l'outil n'usine pas. L'objectif est de trouver le seuil T_{Arms} (m/s²) et la puissance à vide P_{vide} qui permettent de déterminer les clusters « l'outil usine » et «l'outil n'usine pas» par GMM.

La démarche pour vérifier si les classifications par GMM sont correctes est la suivante :

- Modélisation par GMM de différents descripteurs (ΔN, ΔVf, *Arms*, P et ΔP) pour apprendre les seuils de classification contextuelle, afin de déterminer les différents états de la machine-outil.

Chaque machine a des caractéristiques différentes, il faut donc effectuer un apprentissage initial pour chaque machine. Les données d'une journée d'usinage sont utilisées pour cela.

- Les classifications sont ensuite effectuées à partir des règles métiers décrites dans le paragraphe précédent.
- Constitution par fouille manuelle des vecteurs étiquetés avec des classes 100% vraies d'états de la machine-outil.
- Evaluation de la qualité des classifications avec des matrices de confusion.

3 Modèle de Mélange Gaussien (GMM)

Le Modèle de Mélange Gaussien (GMM), par définition est un modèle statistique utilisé pour estimer paramétriquement la distribution de variables aléatoires en les modélisant comme une somme de plusieurs distributions gaussiennes. Il a été utilisé dans plusieurs domaines comme la reconnaissance audio, la modélisation des erreurs de mesure, etc.

Le problème d'estimation d'un GMM, consiste à trouver une approximation appropriée de la densité de probabilité *f* à partir d'un échantillon de *n* réalisations du vecteur aléatoire $X=\{x_1, x_2, x_3 \ldots x_n\}$ en utilisant une combinaison linéaire de plusieurs composantes gaussiennes sous la forme suivante :

$$\text{éq. (2) :}\quad f(x_i\,;\theta) = \sum_{j=1}^{k} a_j \mathcal{N}_j(x_i\,;\mu_j,\sigma_j^2)$$

Où a_j représente l'amplitude de la j ème composante Elle vérifie bien les conditions de probabilités tels que $\sum_j a_j = 1$, et $0 \le a_j \le 1$. $\mathcal{N}_j(x_i\,;\mu_j,\sigma_j^2)$ est la j ème composante avec la moyenne μ_j et la variance σ_j^2.

Dans ce contexte, on définit l'ensemble des amplitudes $a = \{a_j\}$, l'ensemble des moyennes $\mu = \{\mu_j\}$, et celle des écarts-types $\sigma = \{\sigma_j\}$, On définit aussi $\theta = \{\theta_j\}$ avec $\theta_j = \{a_j, \mu_j, \sigma_j\}$. Dans la littérature, il existe plusieurs méthodes qui permettent d'estimer ces paramètres de GMM. Dans ce papier, on va utiliser la méthode de maximum de vraisemblance afin de déterminer les paramètres de GMM (a et θ).

L'idée principale de l'estimation par maximum de vraisemblance est de trouver un ensemble d'estimations des paramètres, pour que la vraisemblance de l'échantillon utilisé soit maximum. La vraisemblance *L* de θ au vu des observations *X* est définie comme suit [15]:

$$\text{éq. (3) :}\quad L(X;\theta) = f(x_1;\theta) * f(x_2;\theta) * \ldots * f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

Puis en important l'équation (2) dans l'équation (3), on obtient:

$$\text{éq. (4) :}\quad L(X;\theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} a_j \mathcal{N}_j(x_i;\mu_j,\sigma_j^2)$$

Il est plus facile de maximiser la log-vraisemblance au lieu de la fonction de vraisemblance elle-même. La fonction log-vraisemblance s'écrit :

$$\text{éq. (5) :}\quad ln(L(X;\theta)) = \sum_{i=1}^{n} \sum_{j=1}^{k} ln(a_j \mathcal{N}_j(x_i;\mu_j,\sigma_j^2))$$

Le problème d'estimation par la méthode du maximum de vraisemblance, revient donc à trouver les racines de l'équation :

$$\text{éq. (6) :}\ \frac{\partial L(X;\theta)}{\partial \theta} = 0$$

En général, la maximisation de la fonction de vraisemblance ne possède pas de solution analytique. C'est pour cela qu'il est nécessaire de recourir à des méthodes itératives. La plus courante est d'utiliser l'algorithme EM (espérance-maximisation) [16]. Cependant, l'algorithme EM a besoin des paramètres initiaux ; le plus important étant le nombre de cluster *k*. Ce dernier peut être déterminé par la méthode BIC en comparant différents modèles de mélange gaussien.
Le critère BIC est défini comme suit [13]:

$$\text{éq. (7) :} BIC(k) = 2\ ln_m\big(L(X;\theta)\big) - N_m ln(n)$$

Où $ln_m\big(L(X;\theta)\big)$est la fonction log-vraisemblance maximisée, N_m est le nombre de paramètres à estimer dans ce modèle (3 par gaussienne : amplitude, moyenne et écart-type).

En général, plus le nombre de paramètres à entrer dans le modèle est grand, plus le modèle est approché sur la densité des données (ainsi plus la valeur de fonction log-vraisemblance est maximisée). Mais la complexité de calcul va aussi augmenter quand il y a davantage de paramètres à estimer. Donc, un compromis doit être trouvé entre le nombre de paramètres à estimer et la fonction log-vraisemblance maximisée, d'où vient la conception du critère BIC. Dans le BIC, le terme N_m est ajouté afin de pénaliser la complexité du modèle. Par conséquent, BIC est maximisé pour des paramétrages plus parcimonieux, i.e., avec N_m petit, et la fonction log-vraisemblance plus grande. BIC est un indicateur permettant d'évaluer le nombre de classes nécessaires (ainsi que la performance du modèle GMM alors obtenu).

4 Validation du modèle gaussien par BIC

Nous appliquons ici la méthode d'apprentissage non-supervisé par GMM proposée précédemment, sur des données venant d'une usine de production de pièces de structure aéronautique. Les résultats sont illustrés sur les classifications visant à déterminer si l'outil usine ou non. Il est admis que, dans l'industrie aéronautique, l'outil ne va usiner qu'après que « la broche tourne à vitesse constante ». Ainsi selon le Tableau 1, l'outil usine si $Arms>T_{Arms}$ ou $P>P_{vide}$ (avec $Vf>0$ et N constante). Il a deux variables (*Arms* et *P*) à classifier. Les étapes de validation du modèle gaussien par BIC seront suivantes :

- Appliquer GMM pour *Arms* et pour *P* en différents nombres de clusters *(1, 2,…, k.*, Pour chaque nombre de clusters, une valeur de BIC peut être déterminée selon l'équation (8).
- Tracer les graphes de BIC en fonction du nombre de clusters pour *Arms* et pour *P* (cf. Fig.4). Dans notre cas ici, la valeur du BIC converge vers un maximum.
- Nous considérons que le nombre de clusters optimal se situe à moins de 5 % de ce BIC maximal.
- En tenant compte de la complexité de l'apprentissage, on choisit le nombre de clusters minimal nécessaire.

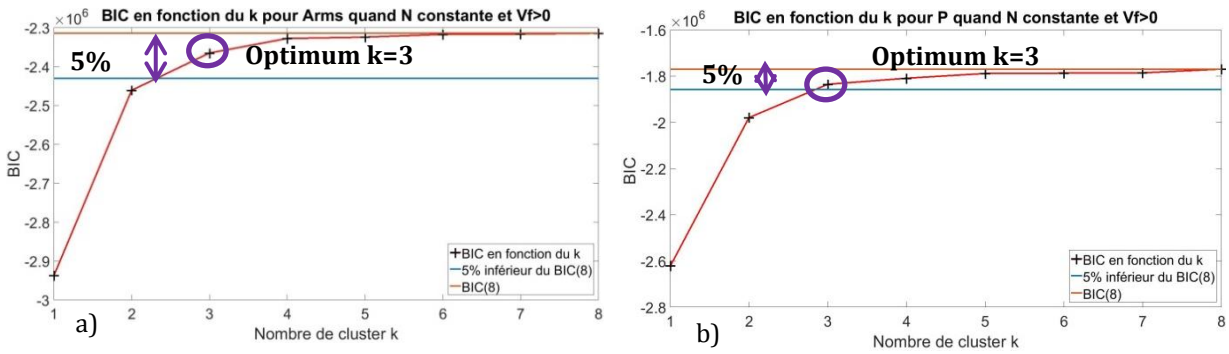


Fig.4 Evolution du critère BIC en fonction du nombre de clusters *k* pour les vibrations *Arms* (a), et pour la Puissance *P* (b)

Pour savoir si l'outil usine ou pas, il faut donc classifier en 3 clusters à partir des vibrations *Arms* et de la puissance *P*. Les données expérimentales confirment ce qui avait été présupposé dans le Tableau 1.

5 Classification contextuelle pour savoir l'outil usine ou pas

Les résultats de classification, basée sur la méthode définie précédemment, sont présentés dans cette section. Un apprentissage par GMM est appliqué sur les enregistrements, pour lesquels « la broche tourne à vitesse constante » et *Vf>0* (cf. Tableau 1). La distribution de *Arms* dans ce cluster peut être modélisée par 3 gaussiennes selon l'analyse selon le critère BIC (cf. section 4). La Figure 5a présente les résultats de modélisation, par une gaussienne qui représente le cluster « l'outil n'usine pas » (Y1 vert), une autre pour celui où « l'outil usine » (Y3 bleu), une 3ème gaussienne (Y2 rouge) représentant de fortes vibrations d'usinage, et enfin la distribution de Y4 (en cyan) est la somme des 3 gaussiennes (Y4=Y1+Y2+Y3). Nous définissons le seuil T_{Arms} comme l'intersection entre les deux gaussiennes Y1 et Y3, afin de minimiser les erreurs de classification (faux positifs ou faux négatifs). On obtient ainsi T_{Arms} =*5,93* m/s².

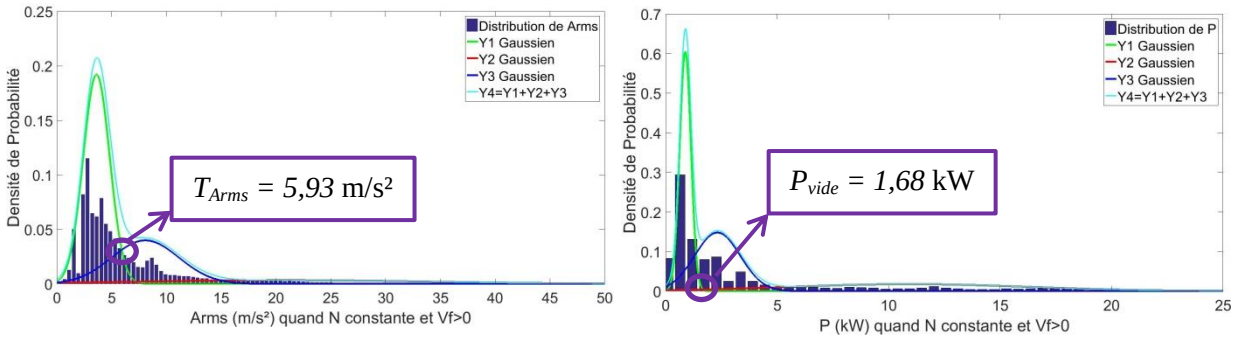


Figure.5 (a) Densité de Probabilité de Arms modélisé par GMM (3 gaussiennes), (b) Densité de Probabilité de P (kW) modélisé par GMM (3 gaussiennes)

La distribution de puissance *P* peut aussi être modélisée par 3 gaussiennes, cf. Fig. 5b : une gaussienne représente le cluster où « l'outil n'usine pas » (Y1 vert), une autre représente celui où « l'outil usine avec un faible enlèvement matière » (Y3 bleu), la 3ème gaussienne (Y2 rouge) modélise lorsque « l'outil usine avec un fort enlèvement matière», la distribution de Y4 (cyan) est la somme des 3 gaussiennes (Y4=Y1+Y2+Y3). Nous définissons le seuil P_{vide} à l'intersection entre les deux gaussiennes Y1 et Y2, obtenant ainsi P_{vide} =*1,68* kW. Le seuil P_{vide} est donc choisi à l'intersection entre la gaussienne « l'outil n'usine pas » et l'autre gaussienne « l'outil usine avec un fort enlèvement matière». En effet, seul *Arms* permet de détecter correctement les phases où « l'outil usine avec un faible enlèvement matière ». Une combinaison des deux mesures est donc effectuée

Ainsi, pour les données pour lesquelles « la broche tourne à vitesse constante » et *Vf>0* (cf. Tableau 1), on définit que, si *Arms* > T_{Arms} ou *P* > P_{vide}, la classification est attribuée au cluster « l'outil usine ». Pour vérifier la classification par des fouilles manuelles, *N* (tr/min), *Vf* (m/min), *Arms* (m/s²) et *P* (kW) vont être tracés sur une même figure. Le résultat de classification en 3 clusters de la puissance *P* seront représentés en différents couleurs sur la même figure. Une journée de données est tracée et vérifiée, et une zone est illustrée comme un exemple dans la Figure 6.

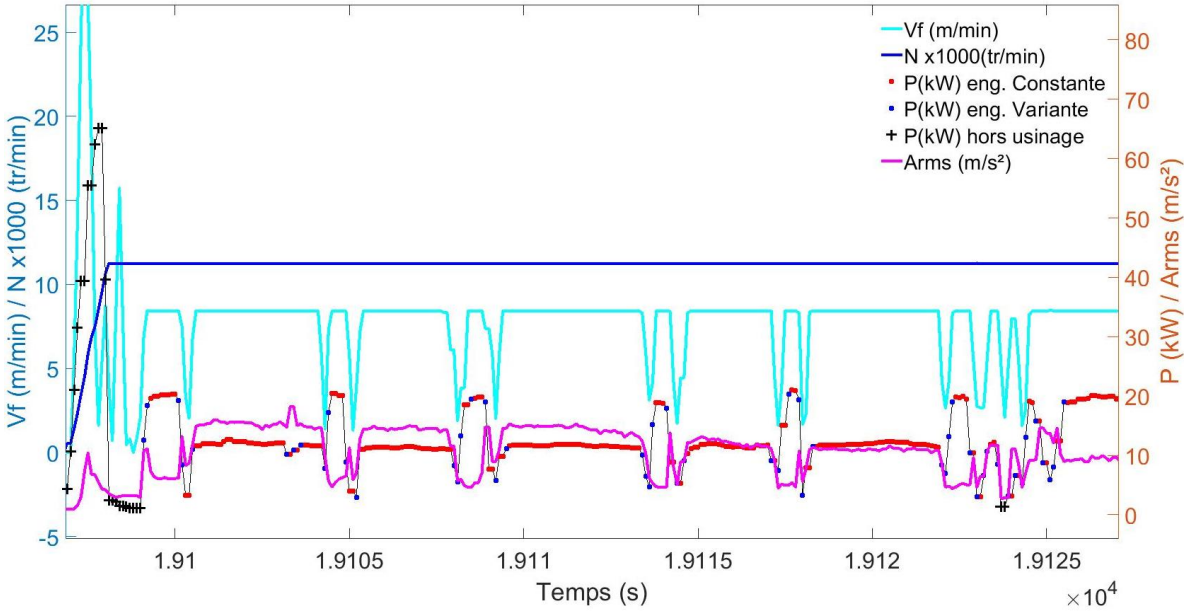


Figure. 6 *N*, *Vf*, *Arms* et *P* en 3 clusters

Vf est en ligne cyan, *N* est en ligne bleu, *Arms* est en ligne rose. Les classifications « l'outil n'usine pas » sont représentés par des signes « + » de couleur noir sur la courbe de puissance *P*, « l'outil usine » sont présentés par des points rouges et des points bleus. Le cluster « l'outil usine » peut être classifié en deux sous-clusters : « l'outil usine avec un engagement constant de l'outil » (en points rouges) et « l'outil usine avec un engagement variant de l'outil » (en points bleus) par la classification sur *ΔP* (cf. Tableau1).

6 Validation des classifications par k-fold

L'estimation de la précision d'une classification par des algorithmes d'apprentissage automatique est importante non seulement pour prévoir sa précision de prédiction, mais également pour choisir un échantillon dans un ensemble donné ou pour choisir un modèle de classification [18].

Pour estimer les précisions de la classification « l'outil usine » ou « l'outil n'usine pas » selon GMM, on peut utiliser la méthode de validation croisée. Ron et al. [19] ont présenté différentes méthodes de validation croisée comme 'Holdout', 'Leave-one-out', 'k-fold' et 'Stratification'. Dans cet article, on va appliquer la méthode de validation croisée 'k-fold'. L'échantillon original est divisé en *k* échantillons, puis un des *k* échantillons est sélectionné comme l'ensemble de validation et les *(k-1)* autres échantillons constitueront l'ensemble d'apprentissage. On calcule le score de performance pour le premier échantillon, puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les *(k-1)* échantillons qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète *k* fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La moyenne des scores de performance est enfin calculée pour estimer la précision de notre classification.

Dans notre exemple le modèle GMM permettra d'obtenir les classes estimées et la fouille manuelle les classe réelles. Les étapes principales sont les suivantes :

1. Appliquer GMM sur une journée de données et les seuils d'*Arms* et *P* sont obtenues. Appliquer les deux seuils sur les données pour obtenir des classes estimées : 1 - « l'outil usine » et 0 - « l'outil n'usine pas ».
2. Choisir 4 heures de données (144 000 points) et créer des classes estimées.
3. Tracer la figure de *N*, *Vf*, *Arms* et *P* en fonction de l'indice selon les classes estimées (comme la Fig. 6). Passer en fouilles manuelles pour corriger les erreurs de classifications et obtenir un autre vecteur. On suppose que ce sont les classes réelles.
4. Comparer les classes estimées et les classes réelles et calculer la matrice de confusion.
5. Calculer les occurrences de vrais positifs (*VP*) pour chaque matrice de confusion, ainsi que les occurrences de faux positifs (*FP*), faux négatifs (*FN*) et vrais négatifs (*VN*). Calculer la précision *ACC* selon *ACC= (VP+VN) / (VP+VN+FP+FN)*, comme la précision de la classification.

Voici la matrice de confusion pour 4h de données dans une journée :

Matrice de confusion			
4 h de la journée 20/09/18		**Classes réelles**	
Classes estimées		L'outil Usine	L'outil N'usine pas
	L'outil Usine	*VP=51 476*	*FP=18*
	L'outil N'usine pas	*FN=67*	*VN=92 470*

Tableau 2. Matrice de confusion pour 4h de donnée de validation

Finalement, on obtient la précision de la classification est : *ACC (*4h*)=99,94%.* La classification sur « l'outil usine » et « l'outil n'usine pas » selon GMM appliqué sur *Arms* et *P* en 3 gaussiens pendant ces 4 heures est bien validée.

7 Conclusion et perspective

L'une des premières étapes de fouille de données est la sélection précise de données pertinentes. Pour ce faire, les données brutes doivent être classées dans différents groupes de contextes. Cet article propose tout d'abord les 17 classes potentielles en usinage. Puis, l'algorithme d'apprentissage automatique non-supervisés (GMM) est testé sur des données venant d'une entreprise d'usinage aéronautique. Les résultats de classification sont comparés à des fouilles manuelles et conduisent aux conclusions suivantes :

1) BIC est un bon indicateur pour trouver le nombre de clusters optimal *k* pour le modèle de GMM.
2) Le critère BIC a montré que, pour la classification en 2 clusters « l'outil n'usine pas » et « l'outil usine », il convenait de modéliser par 3 gaussiens lors de la détermination par GMM des seuils de classification sur *Arms* et *P*, pour une machine-outil donnée.
3) La classification de l'outil usinant ou non, par GMM en combinant les mesures de vibration *Arms* et de puissance *P*, a été validée par fouilles manuelles sur 4 heures de données. La précision de classification atteint *99,94%.*

Ces valeurs sont spécifiques à la machine-outil étudiée, mais la méthodologie de Machine Learning proposée est réplicable.

Dans une prochaine étape, la méthode de GMM va être appliquée sur la classification de *N*, *Vf* et *ΔP* pour la classification des autres états de cette machine-outil dans le Tableau1. Et puis, sur la base de cette classification, les KPIs pourront être calculés.

Références

[1] Teti, R., Jemielniak, K., O'Donnell, G., & Dornfeld, D. (2010). Advanced monitoring of machining operations. CIRP Annals-Manufacturing Technology, 59(2), 717-739.
[2] Quintana, G., & Ciurana, J. (2011). Chatter in machining processes: A review. International Journal of Machine Tools and Manufacture, 51(5), 363-376.
[3] Zhou, Y., & Xue, W. (2018). Review of tool condition monitoring methods in milling processes. The International Journal of Advanced Manufacturing Technology, 1-15.
[4] Godreau, V. (2017). Extraction des connaissances à partir des données de la surveillance de l'usinage, thèse de l'Université de Nantes.
[5] Lenz, J., Wuest, T., & Westkämper, E. (2018). Holistic approach to machine tool data analytics. Journal of Manufacturing Systems.
[6] Morgan, J., & O'Donnell, G. E. (2018). Cyber physical process monitoring systems. Journal of Intelligent Manufacturing, 29(6), 1317-1328.
[7] Liu, C., Vengayil, H., Zhong, R. Y., & Xu, X. (2018). A systematic development method for cyber-physical machine tools. Journal of Manufacturing Systems, 48,13-24.
[8] Kim, D.-H., Kim, T. J., Wang, X., Kim, M., Quan, Y.-J., Oh, J. Min S.-H., Kim, H.., Bhandari, B., Yang, I. & Ahn, S.-H. (2018). Smart Machining Process Using Machine Learning: A Review and Perspective on Machining Industry. International Journal of Precision Engineering and Manufacturing-Green Technology, 5(4), 555-568.
[9] Bhinge, R., Biswas, N., Dornfeld, D., Park, J., Law, K. H., Helu, M., & Rachuri, S. (2014). An intelligent machine monitoring system for energy prediction using a Gaussian Process regression. 2014 IEEE International Conference on Big Data, 978-986.
[10] Godreau, V., Mathieu, R., Etienne, C. et al. (2018). Continuous improvement of HSM process by data mining. Journal of Intelligent Manufacturing, p. 1-8.
[11] Reynolds, D. (2015). Gaussian mixture models. Encyclopedia of biometrics, 827-832.
[12] Fraley, Chris, and Adrian E. Raftery. "How many clusters? Which clustering method? Answers via model-based cluster analysis." The computer journal 41.8 (1998): 578-588.
[13] De Castelbajac, C., Ritou, M., Laporte, S., & Furet, B. (2014). Monitoring of distributed defects on HSM spindle bearings. Applied Acoustics, 77, 159-168.
[14] Wang, Z., Da Cunha, C., Ritou, M., & Furet, B. (2019). Comparison of K-means and GMM methods for contextual clustering in HSM. Procedia Manufacturing, 28, 154-159.
[15] Aldrich J. RA Fisher and the making of maximum likelihood 1912-1922[J]. Statistical science, 1997, 12(3): 162-176.
[16] E. A. Ali, Estimation robuste des modèles de mélange sur des données distribuées, Thèse, Université de Nantes, 2012.
[17] Wang, Z., Ritou, M., Da Cunha, C., & Furet, B. Classification contextuelle pour système d'aide à la décision pour machines-outils. Colloque National S-mart/AIP-PRIMECA, Apr 2019, Les Karellis, France.
[18] Wolpert, David H. "Stacked generalization." Neural networks 5.2 (1992): 241-259.
[19] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai(Vol. 14, No. 2, pp. 1137-1145).