Projet Industriel : Détection d'anomalies dans les systèmes industriels

**Contexte**

Les machines industrielles produisent une quantité importante de données en continu. L’analyse de ces données peut révéler des anomalies, permettant ainsi d’intervenir rapidement avant que des pannes ou des problèmes techniques n'entraînent des interruptions de production coûteuses. Les anomalies dans ce contexte industriel peuvent se manifester de plusieurs façons : des pics soudains de valeur (outliers), des changements de comportement (brouillages) ou encore des modifications de tendance. Développer une solution automatisée et efficace pour détecter ces anomalies sans nécessiter un historique étendu est essentiel pour une application industrielle, surtout pour un usage prédictif.

**Processus de Détection des Anomalies**

1. **Chargement des Données** : Les données industrielles sont chargées depuis un sonogramme. Un prétraitement des données est nécessaire pour éliminer les bruits parasites et normaliser les valeurs.
2. **Application des Algorithmes de Détection** : Des algorithmes sont appliqués pour détecter les anomalies, qui se regroupent en trois catégories :
   * **Pics (outliers)** : Surveiller des dépassements de seuils ou des valeurs extrêmes.
   * **Brouillages** : Identifier les changements de comportement (par ex., rupture d’une tendance linéaire).
   * **Changements de tendance** : Détecter des évolutions progressives ou soudaines des patterns dans les données.
3. **Détection des Anomalies** : Déterminer ce qui constitue un comportement "normal" est un défi initial. Dans un premier temps, il est conseillé de simuler des anomalies en forçant les défauts sur un jeu de données de référence.

**Méthodologies et Algorithmes de Détection**

Beaucoup de méthodes de detections d’anomalies existent. Nous cherchons des méthodes adaptées a notre situation, c’est dire, a appliquer a des données industrielles et surtout avec peu ou pas d’apprentissage.

1. **Méthodes Statistiques** :
   * **Bornes de Chernoff, inégalités de Hoeffding, distributions normales** : Utilisées pour fournir des bornes serrées sur les probabilités de queue pour la détection de fractions spécifiques d’éléments défectueux.
   * **Méthode des valeurs extrêmes** : Permet d’estimer la probabilité d'événements rares en relation avec certaines lois de probabilité.
2. **Méthodes de Machine Learning** :
   * **K-Nearest Neighbors (KNN)** : Approches exactes ou moyennes, pour les cas où la classe d'anomalies est sous-représentée (supervisé).
   * **SVM et One-Class SVM** : Appropriés pour les cas supervisés et non supervisés respectivement.
   * **Isolation Forest** : Spécifiquement adaptée pour la détection d’anomalies dans les données non supervisées.
3. **Transport Optimal (OT)** :
   * Le concept de transport optimal, initié par Gaspard Monge, pourrait être appliqué pour la détection d’anomalies. Le principe consiste à déplacer une masse de points (données normales) vers une autre tout en minimisant les coûts associés, formalisés dans une matrice de coûts. Cette matrice peut être analysée pour identifier les points de divergence, où une anomalie pourrait provoquer une augmentation significative de la distance parcourue.
   * **Wasserstein Distance** : La mesure de Wasserstein est utile pour évaluer la distance entre deux distributions (par exemple, données normales vs données anormales) et permet d’identifier les anomalies de manière optimisée.
4. **Analyse de la Densité et Méthode des Distances** :
   * **Densité k-distance** et **analyse de densité** : La densité autour d'un point est calculée pour identifier des anomalies basées sur l'espacement des points voisins.
   * **Distances diverses** : La **distance dynamique de Warping (DTW)**, pour des signaux comparés sur des longueurs inégales, et **la distance Euclidienne** pour des signaux similaires peuvent être explorées selon le type de signal.
5. **Analyse en Composantes Principales (ACP)** :
   * La projection dans un sous-espace via ACP peut aider à détecter les anomalies en mesurant l'erreur de reconstruction après retour dans l’espace initial, les anomalies ayant une erreur plus élevée.

On retient KNN, isolation forest et SVM qui donnent des résultats convaincants sur des échantillons de données sans apprentissage préalable.