

Table des matières

1	On peut répartir les différentes méthodologies en plusieurs catégories.	2
1.1	Méthodes utilisant un test statistique	2
1.1.1	Concept drift detection based on Fisher's Exact test	2
1.1.2	Detecting Concept Drift Using Statistical Testing	2
1.1.3	Learning with Drift Detection	3
1.1.4	Handling adversarial concept drift in streaming data	3
1.2	Méthodes à base de modèles	4
1.2.1	Tracking recurring contexts using ensemble classifiers : an application to email filtering	4
1.2.2	Learning classification rules for telecom customer call data under concept drift	4
1.2.3	An Ensemble Approach for Incremental Learning in Nonstationary Environments	4
1.2.4	Data Stream Classification Guided by Clustering on Non Stationary Environments and Extreme Verification Latency	5
1.2.5	Titre	5
1.3	Méthodes de détection non supervisées	6
1.3.1	An adaptive algorithm for anomaly and novelty detection in evolving data streams	6
1.3.2	Learning from Time-Changing Data with Adaptive Windowing	6
1.4	Méthodes annexes	7
1.4.1	Efficient Data Stream Classification via Probabilistic Adaptive Windows	7

Chapitre 1

On peut répartir les différentes méthodologies en plusieurs catégories.

1.1 Méthodes utilisant un test statistique

1.1.1 Concept drift detection based on Fisher's Exact test

Danilo Rafael de Lima Cabral, Roberto Souto Maior de Barros, 2018

Mots clefs : Tests Statistiques, Détection sur score de modèle.

Résumé : L'article propose trois méthodes de détection de dérive à partir du test de Fischer fait sur les prédictions du modèle, les variantes sont en partie faites pour réduire le coût calculatoire. Les différences entre les méthodes sont les tests statistiques et la manière dont ils sont utilisés. Dans les trois cas, on suppose que l'on dispose des prédictions du modèle et de la vérité. On sépare les prédictions en 2 ensembles, les plus récentes et les anciennes, on va ensuite, à l'aide de tests statistiques regarder si une différence de distribution apparaît conséquence d'une dérive.

FPDD utilise le test de Fischer quand le nombre d'erreurs où de prédictions justes est faible (inférieur à 5) et utilise le test de l'hypothèse des proportions égales utilisé par la méthode STEPDD le cas échéant. $T(r_o, r_r, n_o, n_r) = \frac{|r_o/n_o - r_r/n_r| - 0.5 \times (1/n_o + 1/n_r)}{\sqrt{\hat{p} \times (1 - \hat{p}) \times (1/n_o + 1/n_r)}}$.

FSDD utilise le test de Fischer quand le nombre d'erreurs où de prédictions justes est inférieure à 5 et utilise le test du chi deux le cas échéant.

FTDD utilise le test de Fischer exclusivement.

Les trois méthodes sont testées avec plusieurs jeux de données contre DDM, ECDD, SEED, FHDDM, STEPDD et sortent avec en moyenne de meilleurs résultats.

À quelle(s) problématique(s) l'article répond ? Amélioration du temps de calcul et des performances de plusieurs méthodes

Quelles sont les pistes à explorer et ont-elles été explorées par d'autres articles ? Il reste à explorer plus de tests statistiques et étudier une possible combinaison, c'est-à-dire l'utilisation de plusieurs tests en simultané et un méta-modèle. Expérimenter l'impact de la taille des fenêtres d'observations prises en compte (elles sont brièvement étudiées).

1.1.2 Detecting Concept Drift Using Statistical Testing

Kyosuke Nishida and Koichiro Yamauchi, 2007

Mots clefs : Tests Statistiques, Détection sur score de modèle.

Résumé : L'article propose une méthode de détection de dérive à partir d'un test. On suppose que l'on dispose des prédictions du modèle et de la vérité. On sépare les prédictions en 2 ensembles, les plus récentes et les anciennes, on va ensuite, à l'aide de la statistique

$$T(r_0, r_r, n_o, n_r) = \frac{|r_0/n_o - r_r/n_r| - 0.5 \times (1/n_o + 1/n_r)}{\sqrt{\hat{p} \times (1 - \hat{p}) \times (1/n_o + 1/n_r)}}$$

rejeter ou accepter l'hypothèse de même distribution.

À quelle(s) problématique(s) l'article répond ? L'article propose une méthodologie de détection de dérive sur les scores des modèles.

Quelles sont les pistes à explorer et ont-elles été explorées par d'autres articles ? La technique laisse à désirer en présence de dérive graduelle. Taille de la fenêtre.

1.1.3 Learning with Drift Detection

João Gama, Pedro Medas, Gladys Castillo, Pedro Rodrigues, 2004

Mots clefs

Résumé : Le but de la méthode développée ici est d'utiliser des intervalles de confiance de la loi normale appliquées aux erreurs faite par le modèle pour détecter une dérive. L'erreur devant logiquement diminuer au fur et à mesure que le modèle dispose de donnée d'entraînement faisant l'hypothèse que la distribution sous-jacente est la même, quand celle-ci augmente au delà d'un certain seuil, on est en présence d'une dérive.

À quelle(s) problématique(s) l'article répond ? L'article propose une méthodologie de détection de dérive sur les scores des modèles.

Quelles sont les pistes à explorer et ont-elles été explorées par d'autres articles ? L'article ne compare pas les autres travaux, mais compare les différences de performances avec et sans la méthodologie.

1.1.4 Handling adversarial concept drift in streaming data

Tegjyot Singh Sethi , Mehmed Kantardzic, 2017

Mots clefs adversarial concept drift

Résumé : A lire.

À quelle(s) problématique(s) l'article répond ?

Quelles sont les pistes à explorer et ont-elles été explorées par d'autres articles ?

1.2 Méthodes à base de modèles

1.2.1 Tracking recurring contexts using ensemble classifiers : an application to email filtering

Ioannis Katakis, Grigorios Tsoumakas, Ioannis Vlahavas, 2009

Mots clefs

Résumé : On suppose que les données arrivent par batch où packets, on va les projeter les données dans un espace puis, un algorithme non supervisé va les grouper par concepts, l'espace de projection permettant de comparer les instances avec la distance euclidienne. Chaque concept différents dispose de son modèle prédictif. Quand une nouvelle série d'observations arrive, elle est projeté dans l'espace, si elle ne correspond à aucun concept, on en créer un nouveau, sinon, on score.

La projection des variables se fait comme : $z_i = \begin{cases} \{P_{i,j}^v : j = 1, \dots, m, v \in V_i\}, & \text{si } f_i \text{ est nominale} \\ \{\mu_{i,j}, \sigma_{i,j} : j = 1, \dots, m\}, & \text{si } f_i \text{ est numérique} \end{cases}$

Et où : $P_{i,j}^v = P(f_i = v \mid c_j), i \in [1, n], j \in [1, m], v \in V_i$. Où V_i est l'ensemble des valeurs possible que peut prendre la variable f_i .

À quelle(s) problématique(s) l'article répond ? À la projection des espaces des données dans un espace euclidien.

Quelles sont les pistes à explorer et ont-elles été explorées par d'autres articles ? L'article fait l'hypothèse que des données d'un même batch appartiennent aux mêmes concepts.

1.2.2 Learning classification rules for telecom customer call data under concept drift

Black, M., Hickey, R, 2003

Mots clefs :

Résumé : L'article développe plusieurs versions d'une méthodologie basée sur les arbres de décisions. On suppose les données arrivant par batchs. Un attribut de temporalité est introduit dans le jeu de donnée. Un arbre de décision est fait sur le dataset. Si l'attribut de temps est présent dans les décisions, alors on a un drift. Cela permet également de voir quelles variables ont été impactées par le drift. Seul les règles émanant des branches qui se trouve dans la branche de l'arbre de décision où un attribut temporel est présent sont supprimés. Les règles qui n'ont pas été impactées restent.

À quelle(s) problématique(s) l'article répond ? Utiliser la temporalité afin de trouver les attributs en dépendant.

Quelles sont les pistes à explorer et ont-elles été explorées par d'autres articles ? Utiliser cette méthode afin de vérifier si

1.2.3 An Ensemble Approach for Incremental Learning in Nonstationary Environments

Michael D. Muhlbaier and Robi Polikar, 2007

Mots clefs :

Résumé : L'article propose Learn⁺⁺.NSE NSE comme Non Stationnary Environment. On suppose que les données arrivent en par paquets, ou en batch. On suppose ici que les dernières données labélisées correspondent au concept le plus récent. Ainsi, à chaque fois qu'un batch de donnée labélisées arrive, on entraîne un nouveau modèle. On va ensuite regarder les taux d'erreur respectifs de chaque ancien modèle sur l'ensemble des paquets de données. On va ensuite attribué une pondération à chaque modèle qui va dépendre des performances de ceux-ci sur les paquets de données. L'erreur des modèles sur les paquets récents va prendre plus de poids que les erreurs anciennes. Dans le cas d'une dérive récurrente, on va observer le poids d'un modèle osciller.

À quelle(s) problématique(s) l'article répond ? L'article répond aux problèmes des dérives récurrentes.

Quelles sont les pistes à explorer et ont-elles été explorées par d'autres articles ? L'article ne prends pas en compte la séparation des paquets en concepts. En effet, deux concepts peuvent arriver au sein d'un même paquets si il sont trop long. S'il sont trop courts, alors les performances du modèle seront insatisfaisantes. La méthodologie utilise seulement les performances dans son discernement des concepts, un ajout de méthodes statistiques augmenterait surement la pertinence de la pondération. La méthodologie entraîne un modèle unique sans se soucier des hyperparamètres.

1.2.4 Data Stream Classification Guided by Clustering on Non Stationary Environments and Extreme Verification Latency

VMA Souza, DF Silva, J Gama, GE Batista , 2015

Mots clefs : Semi-supervisé

Résumé : L'article présente une méthodologie semi-supervisé afin de prédire dans un contexte de dérive. Le contexte est un contexte de dérive incrémentale, le nombre de classes est supposé connu et ne change pas, on dispose initialement d'un nombre d'exemples labélisés.

À l'initialisation, on va attribuer chaque exemple à un cluster selon sa classe. On va ensuite récupérer des exemples non labélisés. On va utiliser k-means pour les répartir en clusters. On utilise les centroïdes des clusters précédemment définis comme centroïdes d'initialisations. Une fois les nouveaux points répartis en clusters, on va attribuer aux instances non labélisés composants les clusters le label du cluster de l'itération précédente qui est le plus proche. Les nouveaux exemples labellisés, on peut réentraîner le modèle sur les exemples les plus récents. On oscille ainsi entre classification supervisé et clustering.

À quelle(s) problématique(s) l'article répond ?

Quelles sont les pistes à explorer et ont-elles été explorées par d'autres articles ?

1.2.5 Titre

Auteurs, date de parution

Mots clefs :

Résumé :

À quelle(s) problématique(s) l'article répond ?

Quelles sont les pistes à explorer et ont-elles été explorées par d'autres articles ?

1.3 Méthodes de détection non supervisées

1.3.1 An adaptive algorithm for anomaly and novelty detection in evolving data streams

Mohamed-Rafik Bouguelia, Slawomir Nowaczyk, Amir H. Payberah ; 2018

Mots clefs : Gas Network Generator, Self Organising Maps

Résumé : L'article propose une amélioration du modèle GNG, en apportant les modifications nécessaires pour une meilleure adaptabilité à un changement de distribution.

L'algorithme des GNG est un algorithme non supervisé *online*. On initialise un nombre n de cluster où de sommets de même dimension que le jeu de donnée. À chaque apparition d'un exemple x , on calcul les deux sommets les plus proches. On va rajouter décaler le sommet le plus proche légèrement en direction du nouveau point, puis on trace une arrête entre les deux sommets les plus proches d'âge 0. On incrémente de 1 toutes les arrêtes, puis, on supprime les arrêtes d'âge supérieur à un seuil. Puis, on va supprimer tout les sommets isolés. Toutes les λ itérations on va créer un nouveau sommet à partir du sommet existant ayant la distance moyenne entre ses points x et son centre la plus élevées.

Les problèmes de ce modèle est qu'un grand nombre de paramètres est nécessaire à son fonctionnement et donc que l'on manque d'adaptabilité, que certains vieux neurones ne sont pas supprimés alors que peu utiles.

L'algorithme proposé le GNG-A comme adaptive, l'adaptabilité signifie que l'initialisation des paramètres perd son importance les hyperparamètres évoluant. Le processus de suppression et création des sommets n'est plus systématique, mais prends en compte la dérive.

À quelle(s) problématique(s) l'article répond ? La paramétrisation des hyperparamètres n'est plus capitale, les sommets ne sont plus créés et supprimés de façon systématique. L'algorithme est maintenant utilisable complètement en ligne.

Quelles sont les pistes à explorer et ont-elles été explorées par d'autres articles ? On a ici un algorithme non supervisé. Adapter ces méthodes à un problème supervisé. Adapter au concepts récurrents lors de la suppression en gardant en mémoire les anciens sommets. Ne traite pas les variables n'étant pas de type numériques.

1.3.2 Learning from Time-Changing Data with Adaptive Windowing

A Bifet, R Gavalda , 2007

Mots clefs : ADWIN

Résumé : Un algorithme de sélection de fenêtre d'apprentissage est présenté. L'algorithme conserve les n plus récentes observations dans une fenêtre W de taille n avant de la séparer en deux parties W_1 et W_2 de taille n_1 et n_2 tel que $n_1 + n_2 = n$.

Avec $m = \frac{1}{1/n_0 + 1/n_1}$ (moyenne harmonique de n_0 et n_1) ; $\delta' = \frac{\delta}{n}$ et $\epsilon_{cut} = \sqrt{\frac{1}{2m} \cdot \ln \frac{4}{\delta'}}$.

Tant que l'on observe pas $|\hat{\mu}_{W_0} - \hat{\mu}_{W_1}| \geq \epsilon_{cut}$ où $\hat{\mu}_{W_0}$ est la moyenne de la fenêtre W_0 et $\hat{\mu}_{W_1}$ celle de W_1 , on supprime les observations les plus anciennes

À quelle(s) problématique(s) l'article répond ? Détection d'une dérive, sélection des observations correspondant au concept le plus récent.

Quelles sont les pistes à explorer et ont-elles été explorées par d'autres articles ? Utiliser ADWIN pour séparer les données en concept différents C_1, C_2, \dots, C_i avec des statistiques tel que la moyenne, la variance les quantiles... Permettant, quand un nouveau concept apparaîtrait de vérifier si l'on ne dispose pas d'autres instances de ce concept dans le cas d'un concept récurrent.

1.4 Méthodes annexes

1.4.1 Efficient Data Stream Classification via Probabilistic Adaptive Windows

Albert Bifet, Jesse Read, Bernhard Pfahringer, Geoff Holmes ; 2013

Mots clefs : Memory optimisation, online learning

Résumé : Ici l'article détaille un procédé pour garder en mémoire un échantillon représentatif des données, gardant avec une probabilité forte les observations récentes et une probabilité décroissante en fonction de l'âge les données anciennes. Couplée avec d'autres outils, les auteurs constatent un gain de performance.

À quelle(s) problématique(s) l'article répond ? Réduction de la mémoire, implémentation élégante avec probabilité de garder les exemple récents fortes.

Quelles sont les pistes à explorer et ont-elles été explorées par d'autres articles ? Pour moi, il y a un petit problème d'implémentation car l'algorithme est très coûteux en terme de RAM pour des gains de performances faibles.