Noémie Varjabedian
VARN06600100

TP individuel 1

1. Présenter la méthode : principe, objectifs, algorithme, avantages et inconvénients.

La méthode de l'EFDT est une technique d'apprentissage automatique supervisé utilisée pour la classification de données en temps réel.

Son principe utilise une approche de type arbre de décision pour classifier les données d'entrée. Elle se distingue des autres méthodes d'arbre de décision par sa rapidité et son efficacité en temps réel. L'algorithme EFDT utilise une approche incrémentale, c'est-à-dire qu'il construit un arbre de décision en temps réel, à mesure que les nouvelles données d'entrée sont reçues. Cette méthode permet une mise à jour constante de l'arbre de décision et une meilleure adaptation aux changements des données d'entrée.

L'objectif de l'EFDT est de fournir une classification en temps réel des données d'entrée. Cela permet de prendre des décisions rapides et précises en se basant sur les données disponibles.

L'algorithme EFDT peut être divisé en trois étapes principales :

1. Construction de l'arbre de décision initial : Lors de cette étape, un nœud racine est créé, qui contient toutes les données d'entrée. Ensuite, l'algorithme sélectionne la variable qui permet de diviser les données en deux sous-ensembles les plus homogènes possible en termes de classe. Il utilise pour cela une mesure de qualité basée sur l'entropie de Shannon avec le Höffding Bound. Les sous-ensembles sont ensuite divisés en deux sous-ensembles plus petits jusqu'à ce que chaque sous-ensemble contienne un petit nombre de données (appelé seuil de fragmentation).

2. Mise à jour de l'arbre de décision : Lorsqu'une nouvelle donnée d'entrée est reçue, l'algorithme EFDT la traverse de la racine jusqu'à la feuille correspondante. Ensuite, il met à jour le nœud feuille en ajoutant la nouvelle donnée. Si le nombre de données dans le nœud feuille dépasse le seuil de fragmentation, alors le nœud est divisé en deux sous-ensembles en utilisant la même mesure de qualité que celle utilisée lors de la construction initiale de l'arbre. Le processus de fragmentation est répété jusqu'à ce que chaque nœud atteigne le seuil de fragmentation.

3. Élagage de l'arbre de décision : L'algorithme EFDT utilise un mécanisme d'élagage pour supprimer les nœuds de l'arbre qui ne sont plus utiles ou qui ont une faible précision de classification. Les nœuds sont supprimés en fonction de leur précision de classification, qui est mesurée à l'aide d'une méthode basée sur l'entropie de Shannon.

Avantages	L'EFDT présente plusieurs avantages par rapport aux autres méthodes d'arbre de décision. Tout d'abord, elle est très rapide et efficace en temps réel grâce à son approche incrémentale. Elle est également très robuste face aux changements de données d'entrée, car elle peut mettre à jour son arbre de décision à chaque nouvelle donnée. Enfin, elle peut être utilisée avec des données de grande dimension sans perte de performance.
Inconvénients	L'inconvénient principal de l'EFDT est sa sensibilité aux données bruitées. En effet, la fragmentation des données peut générer des nœuds qui ne sont pas représentatifs de la classe, ce qui peut affecter la précision de la classification. Par conséquent, il est recommandé d'utiliser des méthodes de prétraitement des données pour éliminer les données bruitées avant d'appliquer l'EFDT.

2. Avec un data set de votre choix, créer un modèle de prédiction incrémental EFDT en utilisant la classe `ExtremelyFastDecisionTreeClassifier` du package `skmuliflow.trees`.

Le système de surveillance des facteurs de risque comportementaux (BRFSS) est une enquête téléphonique liée à la santé qui est collectée chaque année par le CDC. Chaque année, l'enquête recueille les réponses de plus de 400 000 Américains sur les comportements à risque liés à la santé, les problèmes de santé chroniques et l'utilisation des services de prévention. Elle est menée chaque année depuis 1984.

Le dataset en question est un ensemble de données épurées de 70 692 réponses à l'enquête BRFSS 2015 du CDC. Il a une répartition égale 50-50 de répondants sans diabète et avec prédiabète ou diabète. La variable cible Diabetes_binary a 2 classes. 0 correspond à l'absence de diabète et 1 au prédiabète ou au diabète. Cet ensemble de données à 21 variables de caractéristiques et est équilibré.

Source :
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv

3. Comparer la performance du modèle trouvé avec les algorithmes DT (Decision Tree), VFDT et CVFDT en termes de :

   a. Temps d'exécution total

Extremely Fast Decision Tree	318.33 s
Very Fast Decision Tree	2.57 s
Concept-adaptating Very Fast Decision Tree	8.32 s
Decision Tree	0.48 s

   b. Taille de l'arbre généré (profondeur, nombre de nœuds, nombre de feuilles)

Extremely Fast Decision Tree	• Tree depth : 3 • Number of nodes : 9 • Number of leaves : 5
Very Fast Decision Tree	• Tree depth : 5 • Number of nodes : 17 • Number of leaves : 9
Concept-adaptating Very Fast Decision Tree	• Tree depth : 7 • Number of nodes : 25 • Number of leaves : 13
Decision Tree	• Tree depth : 38 • Number of nodes : 19759 • Number of leaves : 39517

   c. Performance de classification en précisant les métriques d'évaluation utilisées (plus qu'une)

Extremely Fast Decision Tree

	Precision	Recall	F1-score	Support
diabetes	0.70	0.70	0.70	999
Non diabetes	0.70	0.70	0.70	1000
accuracy			0.70	1999
macro avg	0.70	0.70	0.70	1999
weighted avg	0.70	0.70	0.70	1999

Very Fast Decision Tree

	Precision	Recall	F1-score	Support
diabetes	0.72	0.71	0.71	999
Non diabetes	0.71	0.72	0.72	1000
accuracy			0.71	1999
macro avg	0.71	0.71	0.71	1999
weighted avg	0.71	0.71	0.71	1999

Concept-adaptating Very Fast Decision Tree

	Precision	Recall	F1-score	Support
diabetes	0.77	0.68	0.72	999
Non diabetes	0.71	0.79	0.75	1000
accuracy			0.74	1999
macro avg	0.74	0.74	0.74	1999
weighted avg	0.74	0.74	0.74	1999

Decision Tree

	Precision	Recall	F1-score	Support
diabetes	0.66	0.66	0.66	1000
Non diabetes	0.66	0.66	0.66	1000
accuracy			0.66	1000
macro avg	0.66	0.66	0.66	2000
weighted avg	0.66	0.66	0.66	2000

d. Interpréter les résultats

En termes de temps, l'extremely fast decision tree est largement plus lent que les autres algorithmes. L'arbre construit est cependant plus optimisé car il possède moins de nœuds, moins de feuilles et a une profondeur plus petite que tous les autres algorithmes. Les métriques d'évaluation montrent que ses performances de classification sont néanmoins similaires aux algorithmes VFDT et CVFDT voire moins performantes.

Sources :

https://scikit-multiflow.readthedocs.io/en/stable/api/generated/skmultiflow.trees.ExtremelyFastDecisionTreeClassifier.html

https://scikit-learn.org/stable/user_guide.html

https://medium.com/@kohlishivendra/precision-recall-f1-score-accuracy-ccc85736b647