Fouille de Données Massives
Projet : Sujet 2

M2 Informatique - SISE

Guillaume Metzler
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Ce projet sera effectué par groupe de deux ou trois étudiants maximum et porte sur l'apprentissage dans un contexte de données déséquilibrées avec une application plus précise sur la détection de fraudes.

Le travail sera à rendre par mail à **guillaume.metzler@univ-lyon2.fr** **avant le 9 Février 2025 au plus tard**. Ce travail se composera d'un dossier retraçant vos démarches et résultats entrepris pour traiter le sujet, de plus amples explications vous sont données ci-dessous. Vous êtes libres d'utiliser le langage de votre choix pour effectuer ce travail, R ou Python. Choisissez celui avec lequel vous êtes le plus à l'aise.

Les données à employer sont disponibles à l'adresse suivante :

Accès aux données

Elles sont associées à une publication plutôt récente portant sur les thématiques de l'**Adapation de Domaine** et de l'**Apprentissage dans un contexte déséquilibré**.

Un code vous permettant de reproduire les expériences du papier est également disponible à cette même adresse, mais l'objectif et que vous réalisiez vos propres expériences.

1 A propos des données

Les données sur lesquelles vous allez travailler sont des données réelles. Il s'agit d'une version anonymisée d'un jeu de données utilisé par un doctorant lors d'une thèse

effectuée chez *ATOS Wordline*.

2 Travail à effectuer : première partie

Dans cette première partie, on souhaite simplement traiter le problème d'apprentissage dans un contexte déséquilibré. Pour cela, on va considérer les jeux de données suivants :

kaggle_source_cate_i, pour i allant de 0 à 3

Plus précisément, pour chaque $i \in [\![0, 3]\!]$, nous avons

- un fichier *train* qui contient les données d'entraînement (mais uniquement les descripteurs), *i.e.* $\mathbf{X}_{train}$
- un fichier *train_label* qui contient les **étiquettes relatives aux données d'entraînement** , *i.e.* $\mathbf{y}_{train}$
- un fichier *train* qui contient les données tests (mais uniquement les descripteurs), *i.e.* $\mathbf{X}_{test}$
- un fichier *test_label* qui contient les **étiquettes relatives aux données tests** , *i.e.* $\mathbf{y}_{test}$

Les jeux de données d'entraînement contiennent environ $40,000$ exemples et une cinquantaine de variables. On dénombre ensuite $15,000$ exemples dans les jeux de données tests.

3 Quelques suggestions

Idéalement, le travail effectué devrait comprendre au moins 5 procédures ou méthodes différentes vues en cours : une méthode peut être un algorithme de classification seul ou encore couplé ou non à une méthode d'échantillonnage par exemple. Je vous donne ci-dessous une liste non exhaustive des méthodes que vous pourriez utiliser pour votre travail.

- **Pre-Process sur les données** : utilisation d'algorithmes d'over-sampling (random - SMOTE - Adasyn - ...) ou encore des approches d'under-sampling (random - Tomek Link - Edited Nearest Neighbout - NearMiss - ...) vous pouvez aussi utiliser des méthodes dites *cost-sensitive* qui vont accorder plus de poids aux exemples d'une classe donnée, voire même des poids spécifiques à chaque exemple.
- **Algorithmes** : vous pourrez utiliser des algorithmes non supervisés comme des méthodes de clustering (k-means, clustering hiérarchique ou encore les auto-encodeurs) pour détecter les fraudes. Vous disposez également d'un large éven-

tail d'algorithmes de classification supervisés que vous pouvez utiliser : decision trees, random forests, gradient boosting, nearest-neighbors, réseaux de neurones profonds, SVM (linéaire ou non ...), analyse discriminante, boosting, ...
- **Post-traitement :** combinez les résultats issus de différents modèles (bagging) afin de créer un modèle potentiellement plus puissant.

N'hésitez pas à regarder sur internet quelques exemples d'utilisations des algorithmes sus-mentionnés et votre objectif sera de les adapter au contexte des données (consulter des sites comme Kaggle - MachineLearningMastery ou encore Medium qui seront pour vous une bonne source d'inspiration). Vous verrez que toutes les méthodes ne sont pas forcément applicables à ce type de données : si tel est le cas, n'hésitez pas à préciser dans votre rapport pourquoi une méthode n'a pas fonctionné selon vous.

4 Travail à effectuer : deuxième partie

Dans cette deuxième partie, on maintenant traiter le même problème mais sur des jeux de données différents pour mettre en œuvre une technique d'**adaptation de domaine non supervisé**.

Pour cela, on se concentrera uniquement sur les fichiers :

kaggle_source_cate_0.

On utilisera donc que le jeu de données 0.

kaggle_target_cate_i.

Soit un ensemble de 4 fichiers pour les données sources mais uniquement de **trois fichiers** pour les données cibles !

- un fichier *train* qui contient les données d'entraînement (mais uniquement les descripteurs), *i.e.* $\mathbf{X}_{train}$
- un fichier *train* qui contient les données tests (mais uniquement les descripteurs), *i.e.* $\mathbf{X}_{test}$
- un fichier *test_ label* qui contient les **étiquettes relatives aux données tests** , *i.e.* $\mathbf{y}_{test}$

Nous n'allons donc pas utiliser de données étiquetées lors de notre apprentissage.

L'objectif est simple, il va falloir essayer d'obtenir de bonnes performances sur les **target de test**, en construisant un modèle à l'aide des données **sources étiquetées** et des données **target non étiquetées**, *i.e.*, on fait de l'**adaptation de domaine non supervisé**.

On pourra également observer les performances sur le jeu de données test de la distribution *source*.

Résumé

Il s'agit du résumé de votre travail, vous devez donc, en une dizaine de lignes environ, présenter le problème traité ainsi que l'approche proposée dans le rapport. Enfin vous finirez par donnée une idée des conclusions obtenues.

Introduction (et éventuellement état de l'art)

Dans cette section, vous devrez présenter le contexte de l'étude (on pourra par exemple reprendre la description donnée au début du document) ainsi que les difficultés liées à cette étude. On pourra également citer quelques exemples d'applications liées à la problématique traitée.

Si vous avez consulté quelques références, liens sur internet ou autre, c'est le moment de les citer en évoquant en deux ou trois lignes la méthode employée.

Cette section se finira par une présentation de la structure de votre rapport, par exemple.

La Section 2 sera consacrée à l'analyse des données. La Section 3 sera consacrée à la présentation de la méthode proposée pour résoudre ce problème ...

Analyse des données

On évoque rapidement le jeu de données. Vous commencerez par une analyse synthétique des données à l'aide d'outils statistiques élémentaires : vous présenterez quelques graphes pertinents et intéressants s'il y a lieu ainsi que les grandes caractéristiques du jeu de données, sélectionnez les informations intéressantes.

Méthodologie

Vous commencerez par présenter les notations que vous allez employer tout au long de la rédaction de votre rapport.

Vous présenterez ensuite les outils que vous allez utiliser dans la partie expérimentale. On commencera par parler de ce que l'on souhaite maximiser (par exemple AUCROC ou F-mesure) en les définissant avant de s'attaquer à la présentation des algorithmes.

Par exemple, si vous faites une contribution basée sur du boosting et que vous combinez avec des méthodes à noyaux, il faudra rappeler ce qu'est le boosting, ce que sont les méthodes à noyaux (ce sont les approches de bases) et ensuite vous expliquerez comment vous combinez les deux pour résoudre le problème confié.
Il ne faut pas hésiter à présenter le processus de façon *abstraite*, c'est-à-dire avec des notations mathématiques et ne pas être uniquement verbeux.

Un pseudo-code est également appréciable pour synthétiser l'approche proposée.

Dans le cas ou vous proposez plusieurs approches à des fins de comparaisons, il faudra prendre soin de présenter les différentes approches et de justifier pourquoi vous intéressez à ces approches là.

Remarque : il n'est pas nécessaire de présenter tous les algorithmes employés, mais uniquement ceux qui servent à l'élaboration d'une version "exotique".

Expériences

Vous dresserez ensuite votre protocole expérimentale qui présentera la ou les méthodes sélectionnées pour répondre à la tache demandée. Celui-ci comprend en général trois parties

Présentation de données

Vous représentez ici rapidement le jeu de données employées ainsi que ses caractéristiques : nombre d'exemples, dimensions, taux de déséquilibre,....

Protocole expérimental

Vous présentez rapidement les expériences que vous allez faire, *i.e.* les différents algorithmes testés, le range des hyper-paramètres employés ainsi que la façon dont sont optimisées ces hyper-paramètres (cross-validation en k-folds, simple validation ou est-ce que vous faites le choix de les fixer). Quels sont vos ensembles d'entraînement/validation/test ?

Les informations que vous fournissez dans cette section doivent permettre au lecteur de pouvoir reproduire les résultats que vous allez présenter dans la sous-section suivante.

Résultats

Ici vous allez présenter et analyser les résultats obtenus à l'aide de graphiques et/ou tableaux. Outre les performances, on pourra aussi s'intéresser au critère de rapidité d'un algorithme.

L'analyse doit aussi permettre de mettre en exergue les avantages inconvénients des méthodes proposées. Cela peut passer par l'utilisation d'autres mesures de performances/critères pour évaluer/comparer vos algorithmes.

Conclusions

Il s'agit de conclure quant à votre étude. Reprendre le travail proposé et les principales conclusions. Il est également important de proposer des perspectives à votre travail en fonction des résultats obtenus et de l'approche proposée.