Détection de fraude sur les cartes de crédit

AZANGHO Camille,KITIHOUN Adeline, MAKHOKH Lamyae

15/11/2019

Introduction :

La fraude est devenue un phénomène très important pour les institutions financières, les caisses populaires et les banques en particulier. La lutte contre la fraude s'accompagne de techniques de prévention traditionnelles telles que les codes PIN, les mots de passe et les systèmes d'identification, mais elles sont devenues inadéquates dans les systèmes bancaires modernes.

L'objectif de ce travail sera d'implémenter la technique d'apprentissage des SVM et ensuite de faire une compariason de ses performances par rapport à d'autres méthodes de machine learning en classification.

1-Description de la base de données

Les données se basent sur les transactions survenues en 2 jours:

- 492 fraudes.

- 284315 transactions.

La variable cible : class.

Les variables explicatives:

```
##  [1] "Time"   "V1"     "V2"     "V3"     "V4"     "V5"     "V6"
##  [8] "V7"     "V8"     "V9"     "V10"    "V11"    "V12"    "V13"
## [15] "V14"    "V15"    "V16"    "V17"    "V18"    "V19"    "V20"
## [22] "V21"    "V22"    "V23"    "V24"    "V25"    "V26"    "V27"
## [29] "V28"    "Amount" "Class"
```

Echantillionage de nos données:

La question du déséquilibre des classes peut entraîner de graves préjugés en faveur de la classe majoritaire, ce qui réduit les performances de la classification et augmente le nombre de faux négatifs. Comment pouvons-nous résoudre le problème?

Les techniques les plus couramment utilisées sont le rééchantillonnage des données, soit un sous-échantillonnage de la majorité de la classe, soit un sur-échantillonnage de la classe des minorités, ou une combinaison des deux. Cela se traduira par une amélioration des performances de classification.

Dans ce travail, nous avons utilisé le sous-échantillonnage le "Under-sampling", qui consiste à garder toutes les données de la calsse minoritaire et à réduire de manière aléatoire celles de la classe majoritaire afin de réequilibrer le jeu de données.

2- Modélisation du svm

Définition:

Support Vector Machine (SVM) est un algorithme d'apprentissage automatique supervisé, qui peut être utilisé à la fois pour les défis de classification et de régression. Cependant, il est principalement utilisé dans les problèmes de classification. Dans cet algorithme, nous traçons chaque donnée sous forme de point dans un espace à n dimensions (où n est le nombre de caractéristiques que vous avez), la valeur de chaque caractéristique étant la valeur d'une coordonnée particulière. En d'autres termes,à partir de données d'apprentissage étiquetées (apprentissage supervisé), l'algorithme génère un hyperplan optimal qui classe les nouveaux exemples.

1. Configurer les données de formation pour la création de modèles.

2. Configurer les paramètres du SVM.

3. Formateur SVM.

4. Prédicteur SVM.

Kernel:

Les algorithmes SVM utilisent un ensemble de fonctions mathématiques définies en tant que noyau (kernel). La fonction du noyau est de prendre les données en entrée et de les transformer dans la forme requise. Différents algorithmes SVM utilisent différents types de fonctions du noyau. Ces fonctions peuvent être de types différents. Par exemple, une fonction de base radiale (RBF), linéaire, non linéaire, polynomiale et sigmoïde.

Dans ce travail, nous avons demandé à la procédure svm() de construire un classifieur linéaire (kernel = 'linear')

3-Autres méthodes de classification

On a fait 3 autres méthodes de classification pour les comparer avec le SVM et chercher laquelle donne de meilleures performence sur l'échantillon test

- Arbre de classification.

- Gradient boosting.

-Regression logistique.

4- Comparaison des différentes méthodes

Selon la taille de l'échantillon et l'equilibrage du jeu de données on a des résultats différents au niveau des performences. On remarque cependant que la méthode du Gradient Boosting donne de meilleurs résultats.

5- Conclusion:

Notre objectif premier était d'effectuer l'implémentation des SVM pour la détection de la fraude. A la suite de notre étude on peut dire que:

Avantages du SVM:

. Sa grande précision de prédiction.

. Fonctionne bien sur de plus petits data sets.

. Ils peuvent être plus efficace car ils utilisent un sous-ensemble de points d'entraînement.

Inconvénients du SVM:

. Ne convient pas à des jeux de données plus volumineux, car le temps d'entraînement avec les SVM peut être long.

. Moins efficace sur les jeux de données contenant du bruits et beaucoup d'outliers.