L'INSTITUT agro Rennes Angers

La détection d'anomalies de données déséquilibrées : cas du cancer du sein

P. Andrieu, A. Robert, T. Baudat

24 octobre 2024

La détection d'anomalies

Identification d'individus qui présentent un écart par rapport à la normale

Détection de fraudes bancaires, informatiques...

En médecine, faible présence d'individus avec des cas graves (par comparaison à des personnes saines)

-> Problèmes liées à des données déséquilibrés

Jeu de données

Jeu de données sur le cancer du sein (University of Wisconsin – Madison, n.d.)

30 features numériques (caractères géométriques des différentes cellules)

1 variable factorielle : Bénigne (0) ou Maligne (1)

5.5% du jeu de données représente les cellules malignes -> on les considère comme des anomalies

3 méthodes étudiées

- Local Outlier Factor
- Isolation Forest
- One-Class Support Vector Machine

Certaines méthodes ne sont pas adaptées à la situation de nos données (par exemple : DBscan qui est adapté aux données de petites dimensions)

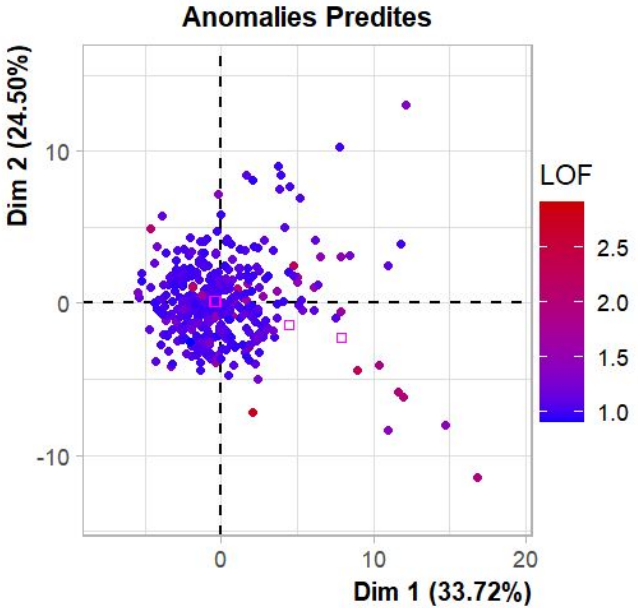
Local Outlier Factor - Principe

Assignation à chaque point d'une valeur de densité locale

Comparaison de chaque densité locale à celle des voisins

Donne un score de "LOF" : plus il est élevé, plus l'observation a de chances d'être une anomalie

Deux hyperparamètres : le seuil du score de LOF et le nombre de voisins considérés

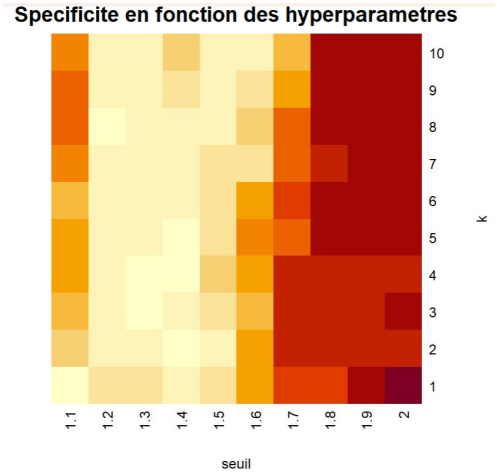


(Jayaswal, 2020)

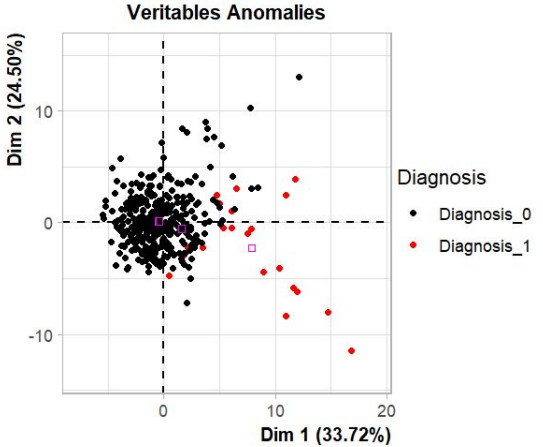
Local Outlier Factor - Résultats

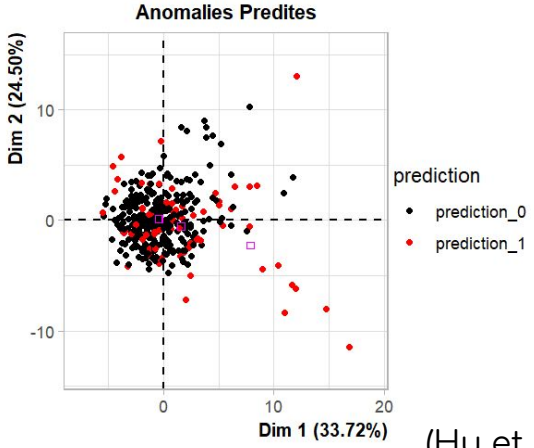
AUC : 0.88

Spe : 0.82



Référence	0	1
Prediction		
0	341	16
1	8	13

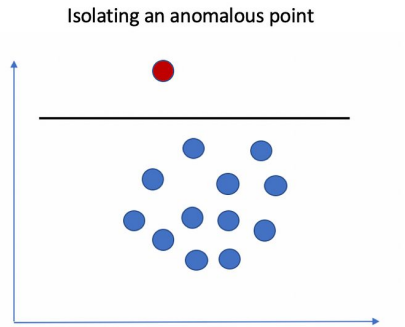


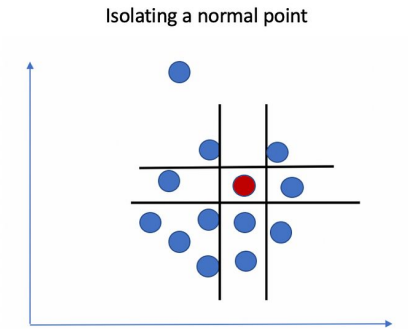


(Hu et al., 2022)

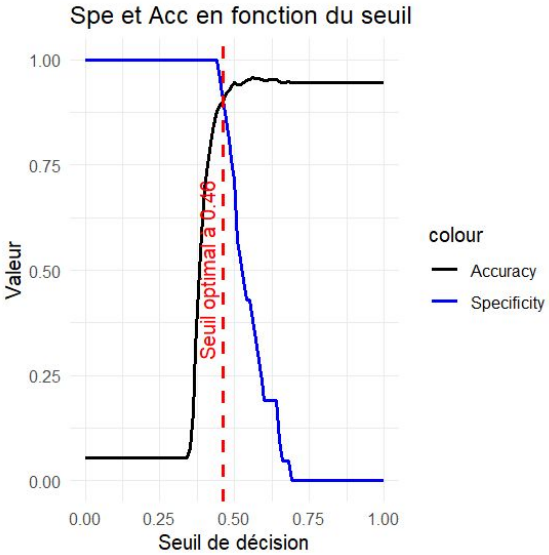
Isolation Forest - Principe

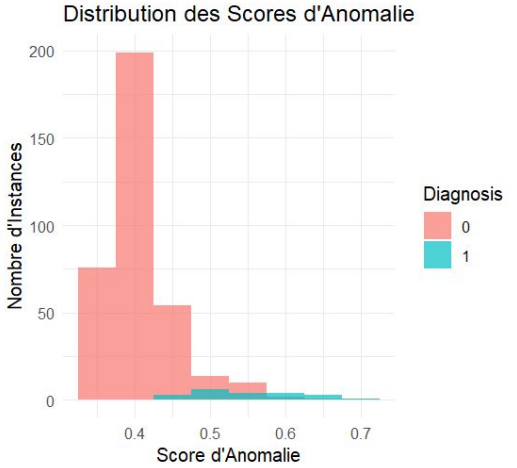
- Principe
  - Anomalies facilement isolables, car rares et différentes
- Concept clé
  - Isolation par partitionnement aléatoire (construction d'arbres de manière aléatoire)
  - Profondeur d'isolement
- Etapes de l'algorithme
  - Echantillonnage aléatoire
  - Construction de l'arbre d'isolation
  - Calcul du score d'anomalie
- Score d'anomalie
  - compris entre 0 et 1
  - d'autant plus grand que l'isolement est rapide
  - seuil à déterminer

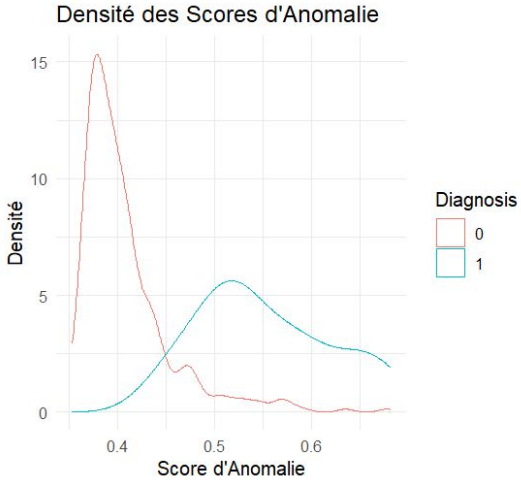




Isolation Forest - Résultats

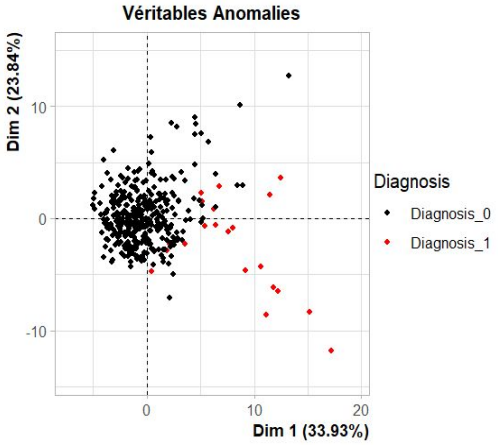


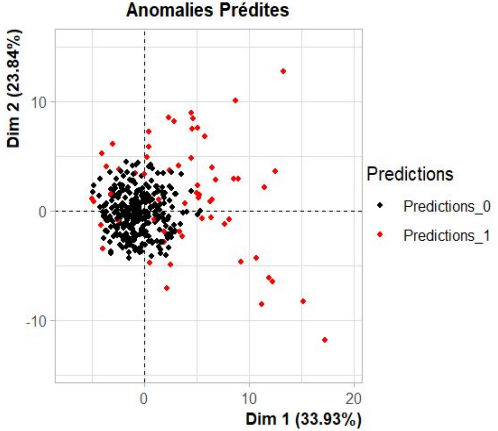




Isolation Forest - Résultats

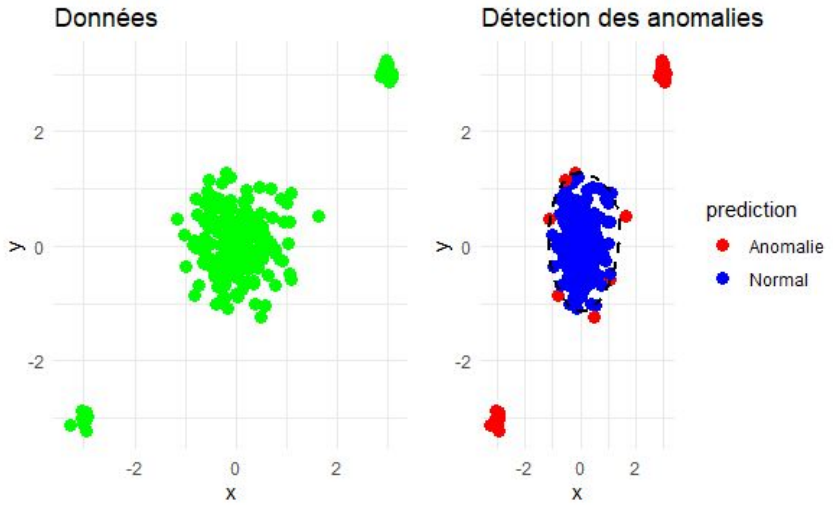
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 321   2
         1  36  19

               Accuracy : 0.8995
                 95% CI : (0.8646, 0.9279)
    No Information Rate : 0.9444
    P-Value [Acc > NIR] : 0.9998

                  Kappa : 0.4563

 Mcnemar's Test P-Value : 8.636e-08

            Sensitivity : 0.8992
            Specificity : 0.9048
         Pos Pred Value : 0.9938
         Neg Pred Value : 0.3455
             Prevalence : 0.9444
         Detection Rate : 0.8492
   Detection Prevalence : 0.8545
      Balanced Accuracy : 0.9020

       'Positive' Class : 0
```





One-Class Support Vector Machine - Principe

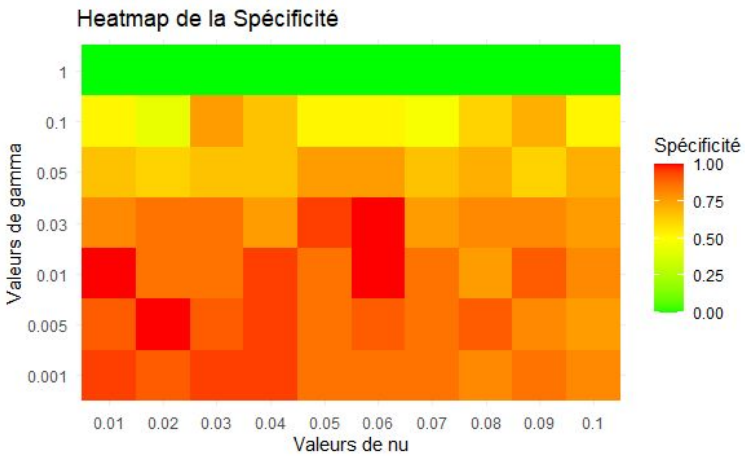
- Apprentissage sur une seule classe (points normaux) pour déterminer une frontière qui enveloppe la majorité des points normaux
- Transformation des données via un noyau (comme les SVM)
- Tous les points qui de sont pas dans la zone normale dans l'hyperplan est considéré comme une anomalie
- **Paramètre ν :** contrôle la proportion d'erreurs acceptées (points hors de la frontière) et la fraction de support vectors.
- **Paramètre γ :** détermine l'influence d'un point de données individuel ; un γ élevé génère une frontière plus complexe.

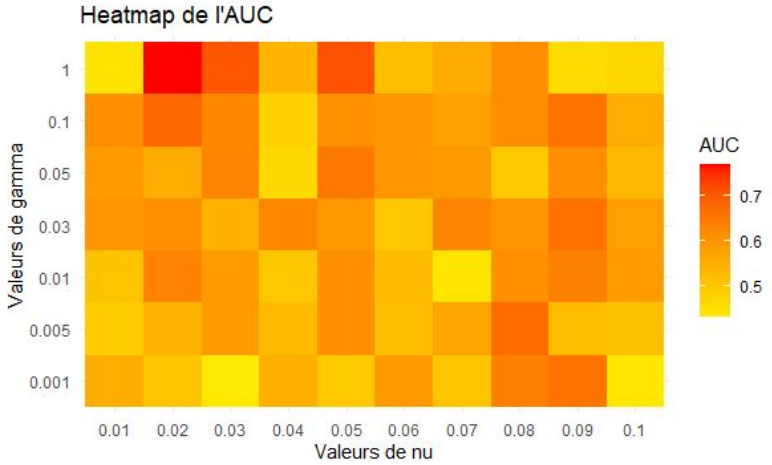


One-Class Support Vector Machine - Résultats

Sélection de paramètres :

- Paramètre ν : en lien avec la proportion d'anomalies -> inférieur à 10%
- Paramètre γ : inversement proportionnel au nombre de features (ici 30 donc valeur de base = 0.03)

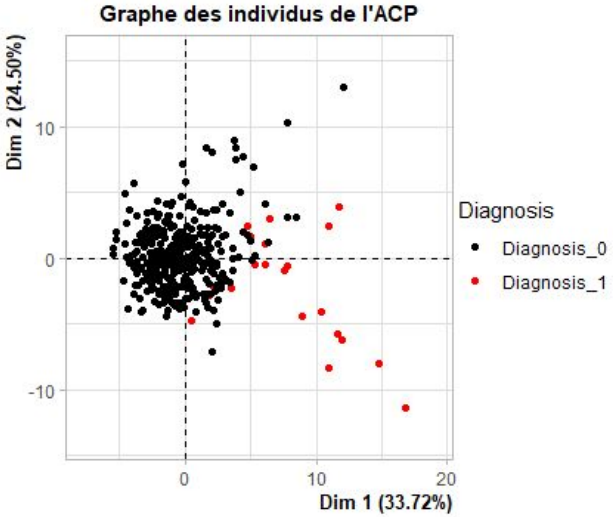


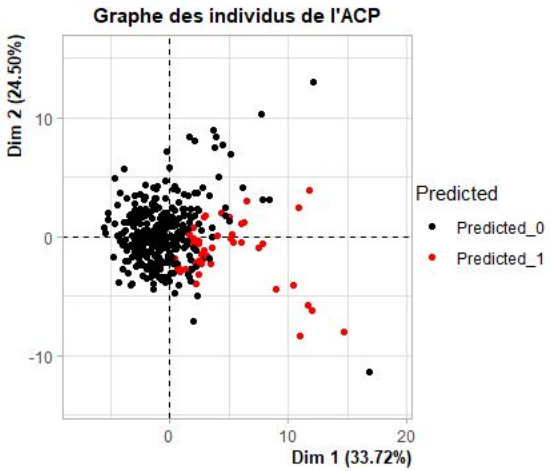


One-Class Support Vector Machine - Résultats

AUC = 0.96

Spe = 0.86





Comparaison des méthodes

	LOF	Isolation Forest	OC-SVM
AUC	0.88	0.96	0.48
Spécificité	0.82	0.90	0.90

En plus d'obtenir de meilleurs résultats, Isolation Forest reste la méthode la plus rapide (puis OC-SVM et enfin LOF qui reste la plus longue) (Togbe et al., 2020)

Les limites de la détection d'anomalies par le machine learning

Problèmes si trop d'anomalies regroupées qui peuvent ne pas détectées (exemple d'un capteur qui peut générer des données anormales sur une durée t)

Choix des hyperparamètres nécessaire

Une alternative pour les données déséquilibrées ?

Ré-échantillonnage possible avec oversampling ou undersampling

Oversampling (sur-échantillonnage) : Générer des observations pour la classe minoritaire (par exemple : SMOTE, Adasyn)

Undersampling (sous-échantillonnage) : Supprimer les observations de la classe majoritaire (par exemple : Tomek)

(Rouvière, 2023)

Conclusion

La détection d'anomalies est adaptée à tous les domaines (diversité de méthodes avec des contraintes différentes)

Isolation Forest reste la méthode la plus performante par rapport à OCSVM et LOF

No Free Lunch : Aucune méthode idéale de manière générale

Compromis à faire : qu'est-il le plus important de conserver?

Bibliographie

Hu Y, Shan WM and Y, Australia (2022) Rlof: R Parallel Implementation of Local Outlier Factor(LOF).

Liu FT, Ting KM, Zhou Z-H (2008) Isolation Forest. 2008 Eighth IEEE Int. Conf. Data Min. IEEE, Pisa, Italy, pp 413–422

Rouvière L (2023) Données déséquilibrées.

Togbe MU, Chabchoub Y, Boly A, Chiky R (2020) Etude comparative des méthodes de détection d'anomalies.

University of Wisconsin–Madison (n.d.) Breast Cancer Wisconsin (Diagnostic) Data Set. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

Vaibhav Jayaswal (2020) Local Outlier Factor (LOF) — Algorithm for outlier identification, dans Towards Data Science