Master 2 Machine Learning for Data Science

TP2
Network reconstruction - Evaluation

Réalisé par:
Amira KOUIDER - 21904040 FI
Nadia RADOUANI – 21911973 FI

Enseignante: Mme Séverine AFFELDT

Année universitaire: 2020-2021

INTRODUCTION:

Il existe plusieurs méthodes de reconstruction de réseau, à travers ce TP, nous allons en comparer trois et évaluer les réseaux construits : Hill-Climbing, l'aproche PC et Aracne.

1. *Préliminaires*

Data

La base de données utilisée est ***insurance dataset*** proposée par le package de R ***bnlearn,*** elle est composée de 2000 lignes et 27 colonnes.

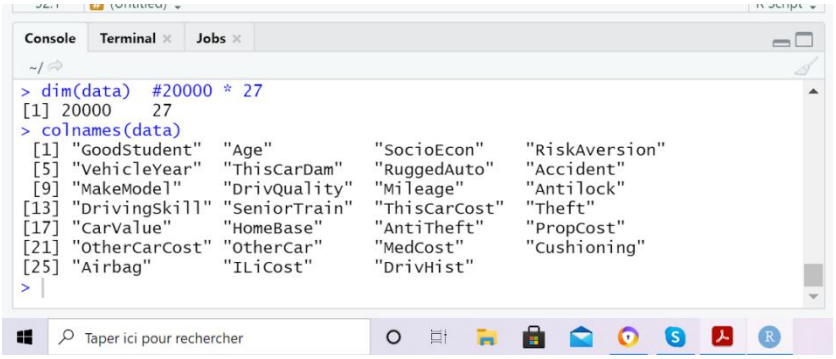

Figure 1 - Aperçu de la base

Insurance ground truth model

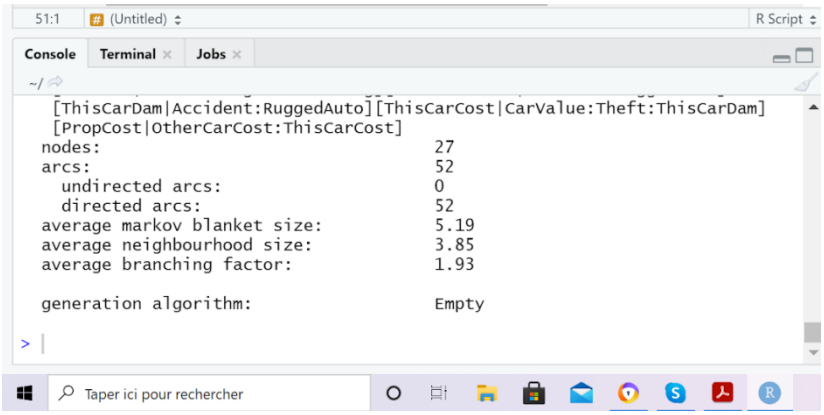

Figure 2 - Dag Content

Le vrai graphe est un graphe direct, composé de 27 noeuds et 52 arêtes.

Directed igraph network

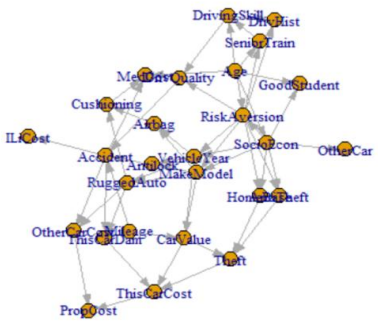

Figure 3 - Dag Graph

2. ***Score-based method (hill-climbing)*** :

Un algorithme de recherche locale qui se déplace continuellement dans le sens d'une augmentation de la valeur pour trouver la meilleure solution au problème. Il se termine lorsqu'il atteint une valeur de pic où aucun voisin n'a une valeur supérieure.

C'est la variante de la méthode **Generate and Test**, qui produit un retour d'information aidant à décider de la direction à prendre dans l'espace de recherche.

Il est caractérisée comme étant une approche gourmande, car, l'algorithme se déplace dans la direction qui optimise le coût, et il ne fait pas de retour en arrière dans l'espace de recherche, car il ne se souvient pas des états précédents.

Insurance ground HC model

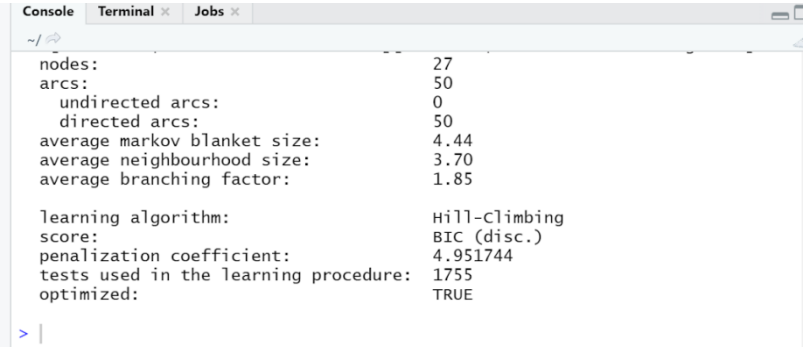

Figure 4 - HC Content

Le graphe obtenu est un graphe direct, composé de 27 noeuds et 50 arêtes.

igraph network

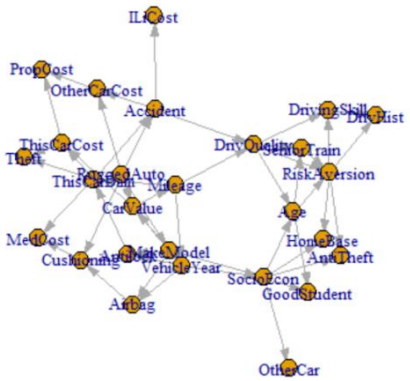

Figure 5 - HC Graph

The number of true positive (TP), false positive (FP) and false negative (FN)

```
> scores(dag, data_hc)
$tp
[1] 38

$fp
[1] 12

$fn
[1] 14

$precision
[1] 0.76

$recall
[1] 0.7307692

$fscore
[1] 0.745098

>
```

Figure 6 - HC Scores

Entre le modèle réel et le modèle obtenu en utilisant HC, 38 arêtes ont été correctement prédites. Le modèle donne une précision de 76%.

FP edges in reconstructed network

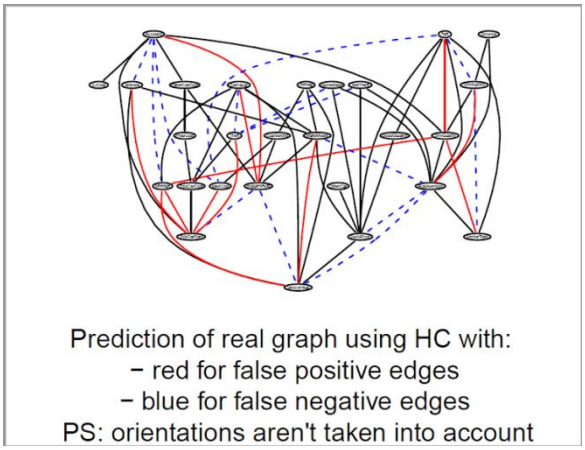

Figure 7 - Graphe avec les FP et FN

On remarque qu'il y a bien 12 FP arêtes en rouge et 14 FN arêtes en bleu (ce qui a été calculé précédemment) et ceci sans prendre en compte les orientations.

FP edges in reconstructed network (orientation are taken into account)

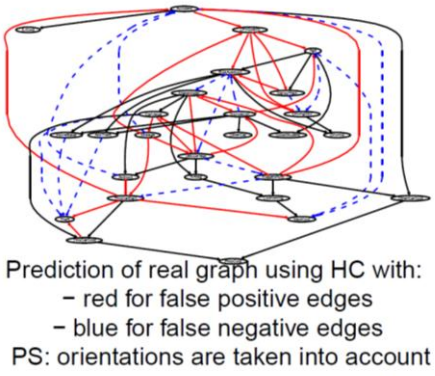

Figure 8 - Graphe avec les FP et FN

On remarque que la prise en compte des orientations change le nombre des arêtes FP et FN.

3. ***Constraint-based method (PC)***

Un algorithme qui vérifie les nœuds deux à deux, et refait le test en conditionnant sur une variable tierce et en grandissant, en plus en plus, son ensemble du conditionnement. Il a l'avantage de pouvoir donner des orientations exprimant les causalités mais son inconvénient réside dans le fait que le test d'indépendance est sensible au bruit des données.

Insurance ground PC model

```
> pcl_bn <- as.bn(pcl)
> pcl_bn

  Random/Generated Bayesian network

  model:
    [partially directed graph]
  nodes:                                27
  arcs:                                 29
    undirected arcs:                    13
    directed arcs:                      16
  average markov blanket size:          2.74
  average neighbourhood size:           2.15
  average branching factor:             0.59

  generation algorithm:                 Empty

>
```

Figure 9 - PC Content

Le graphe obtenu contient 27 noeuds et 29 arêtes dont 13 directes et 16 indirectes.

igraph network

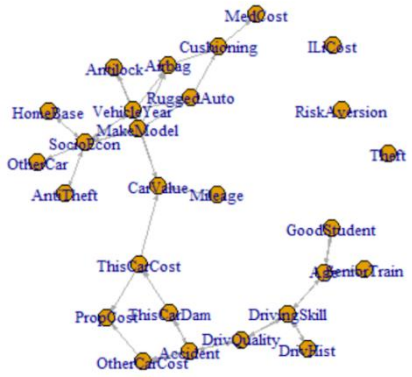

Figure 10 - PC Graphe

The number of true positive (TP), false positive (FP) and false negative (FN)

```
~/
> scores(dag, pcl_bn)
$tp
[1] 29

$fp
[1] 0

$fn
[1] 23

$precision
[1] 1

$recall
[1] 0.5576923

$fscore
[1] 0.7160494

> |
```

Taper ici pour rechercher

Figure 11 - PC Scores

Entre le modèle réel et le modèle obtenu en utilisant PC, 29 arêtes ont été correctement prédites. Le modèle donne une précision de 100%.

FP edges in reconstructed network

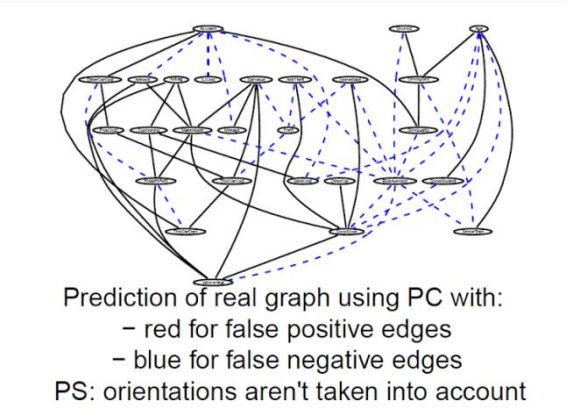

Figure 12 - Graphe avec les FP et les FN

Sans prendre en compte les orientations, le graphe donne bien 29 arêtes TP et 0 FP.

4. ***Local search method (aracne)***

L'algorithme commence par un graphe complet, il estime l'information mutuelle entre deux nœuds et la compare à un seuil, ensuite il supprime les liens inférieurs à ce seuil. C'est une approche qui s'applique sur un squelette de graphe, il ne prend pas en considération l'orientation. Ceci ne permet pas de savoir qui contrôle qui, et peut conduire à une suppression de vrais liens.

Insurance ground Aracne model

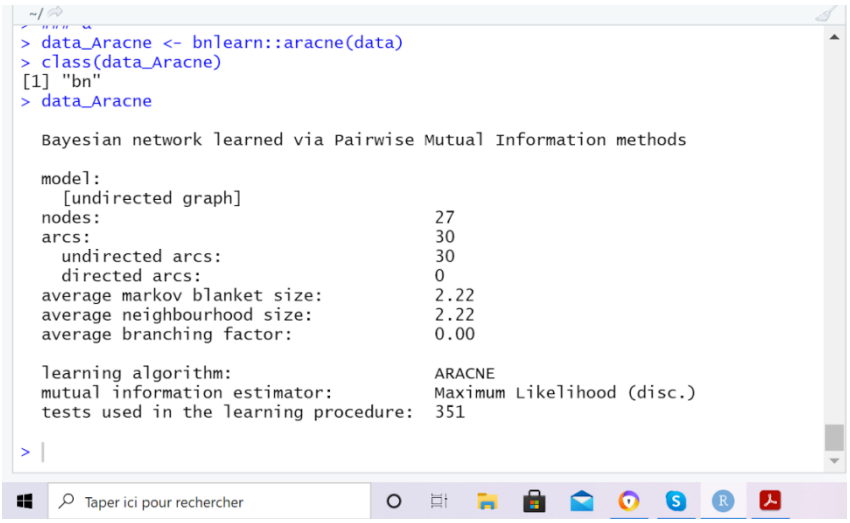

Figure 13 - Arcane content

Le graphe obtenu contient 27 noeuds et 30 arêtes indirectes.

igraph network

Figure 14 - Aracne Graph

The number of true positive (TP), false positive (FP) and false negative (FN)

```
> scores(dag, data_Aracne)
$tp
[1] 28

$fp
[1] 2

$fn
[1] 24

$precision
[1] 0.9333333

$recall
[1] 0.5384615

$fscore
[1] 0.6829268

>
```

Figure 15 - Aracne Scores

Entre le modèle réel et le modèle obtenu en utilisant Aracne, 28 arêtes ont été correctement prédites. Le modèle donne une précision de 93%.

FP edges in reconstructed network

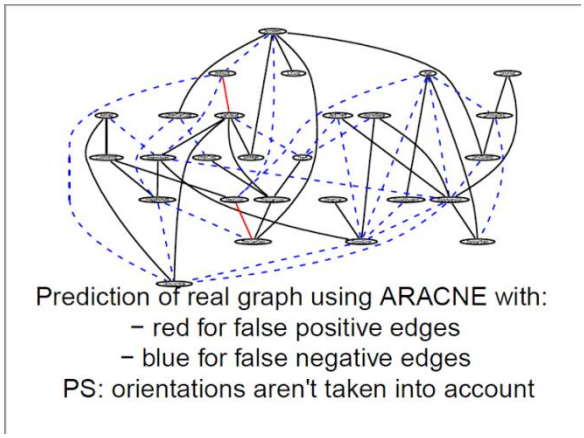

Figure 16 - Graphe avec les FP et les FN

On remarque qu'il y a bien 28 FP arêtes en rouge et 2 FN arêtes en bleu (ce qui a été calculé précédemment) et ceci sans prendre en compte les orientations.

CONCLUSION:

On comparant les résultats de précision des trois méthodes (76% pour HC, 100% pour PC, 93% pour Aracne) on déduit que la meilleure méthode de prédiction pour notre base est l'approche PC).