International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 46 No. 2 September 2024, pp. 294-306

Fiabilisation Du Machine Learning Pour L'apprentissage De Données Déséquilibrées Combinant La Chaîne De Markov Cachée Et Les Techniques De Rééchantillonnage

Application Sur La Prédiction D'Anomalie

RAMAHEFY Tiana Razefania[1] and RANDRIAMAROSON Rivo Mahandrisoa[2]

[1]Université de Vakinankaratra, Madagascar
ramahefy@yahoo.fr
[2]SE-I-MSDE
ED-STII
[1,2]Antananarivo, Madagascar
Corresponding Author: RAMAHEFY Tiana Razefania; ramahefy@yahoo.fr

Résumé — L'objectif de cet article est de concevoir un modèle capable de rendre l'intelligence artificielle plus fiable en se concentrant sur la gestion de données déséquilibrées. Ce modèle utilise le rééchantillonnage par apprentissage automatique, qui est la « technique de suréchantillonnage synthétique », combiné à des chaînes de Markov cachées pour prédire les anomalies. Le principe est de construire une approche puissante pour améliorer les performances de prédiction, notamment dans des contextes tels que la détection de fraude, qui est considérée comme une sorte d'anomalie.

Mots Clés — Intelligence Artificielle ; Données Déséquilibrées ; Chaines De Markov Cachées ; SMOTE ; Détection D'anomalie

Abstract — The aim of this paper is to design a model that can make artificial intelligence more reliable by focusing on handling imbalanced data. This model uses machine learning resampling, which is the "synthetic oversampling technique," combined with hidden Markov chains to predict anomalies. The principle is to build a powerful approach to improve prediction performance, especially in contexts such as fraud detection, which is considered a kind of anomaly.

Mots Clés — Artificial Intelligence ; Imbalanced Data ; Hidden Markov Chains ; SMOTE ; Anomaly Detection.

I. INTRODUCTION

Concevoir un système qui permet de faire un apprentissage des données déséquilibrées est un défi que l'intelligence artificielle essaie de gérer proprement. Durant les processus d'analyse de classification des données, les données déséquilibrées se définissent comme étant une distribution inégale des classes [1]. Ce déséquilibre peut induire une performance biaisée des modèles d'apprentissage [2]. En outre les modèles de Markov cachés (Hidden Markov Models ou HMM) sont généralement employés pour des tâches telles que la reconnaissance de la parole, la modélisation de séquences temporelles, la bio-informatique et surtout dans la détection d'anomalie comme les fraudes. Nous allons utiliser ces chaînes de Markov cachées et les techniques de rééchantillonnage pour fiabiliser l'apprentissage des données déséquilibrées dans la càs de la prédiction d'anomalie.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 46 No. 2 September 2024, pp. 294-306

II. METHODOLOGIES

Différentes étapes importantes sont nécessaires pour créer un bon modèle capable de gérer proprement l'apprentissage de données déséquilibrées. Les données sont rééchantillonées via une méthode appelée *Synthetic Minority Over-sampling Technique* ou SMOTE. Le but est d'avoir des données équilibrées avant l'entrainement du système qui est assuré par le HMM.

A. Prétraitement des Données

Dans cette étape, l'exercice à faire est d'équilibrer les données en augmentant le nombre d'exemples dans la classe minoritaire et en générant de nouveaux exemples synthétiques. Il est nécessaire ainsi de générer des données déséquilibrées en utilisant une distribution normale. Les observations sont basées sur les deux distributions normales distinctes, une pour les comportements normaux et une pour les comportements anormaux. Les deux paramètres qui définissent les distributions normales sont alors les moyennes et les écarts-types.

Pour la classe normale :

$$X_n \sim \mathcal{N}(\mu_n, \sigma_n^2) \quad (1)$$

Pour la classe anormale :

$$X_{an} \sim \mathcal{N}(\mu_{an}, \sigma_{an}^2) \quad (2)$$

1) Sélection des échantillons minoritaires

Soit D le nombre de la dimensions/caractéristiques de données utilisées et x_i , x_j deux éléments de la classe minoritaire tel que $i \neq j$. La formule du K-plus proches voisins de x_j par rapport à x_i est définie par la distance euclidienne suivante :

$$Dist(x_i, x_j) = \sqrt{\sum_{d=1}^{D}(x_i, d - x_j, d)^2} \quad (3)$$

Lorsque le calcul des distances de x_i à tous les autres points de la classe minoritaire est fait alors nous sélectionnons les k points ayant les distances les plus petites à x_i. Ces points sont les k-plus proches voisins de x_i.

2) Génératoin d'un nouvel échantillon synthétique

La génération proprement dite des échantillons synthétique se fait comme suit :

- On détermine et sélectionne aléatoirement un k-plus proches voisins, noté $x_{i\text{-voisin}}$
- On génère une nouvelle instance synthétique x_{synt} en interpolant x_i et $x_{i\text{-voisin}}$

$$x_{synt} = x_i + \gamma * (x_{i\text{-voisin}} - x_i) \quad (4)$$

Avec γ est un nombre aléatoire entre 0 et 1

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/ Vol. 46 No. 2 September 2024, pp. 294-306

B. Définition du modèle de Markov Caché

Soient :

- P la probabilité initiale des états,
- $S = \{S_1, S_2, S_3, \ldots, S_n\}$ l'ensemble des états cachés
- $Q = \{q_1, q_2, q_3, \ldots, q_n\}$ l'ensemble des séquences d'états cachés

1) Matrice de transition des etats (A)

La matrice $A = \{a_{ij}\}$ de transition des états cachés est définie comme suit :

$$a_{ij} = \mathcal{P}(q_{t+1} = S_j \mid q_t = S_i) \quad (5)$$

où a_{ij} est la probabilité de transition de l'état S_i à l'etat S_j et $T = \{t_1, t_2, t_3, \ldots, t_n\}$ la longueur totale de l'observation

2) Matrice d'émission (B)

Soit $O = \{o_1, o_2, o_3, \ldots, o_n\}$ l'ensemble des observations possibles alors

La matrice d'émission (B) est définie comme suit:

$$b_j(o_t) = \mathcal{P}(o_t \mid q_t = S_j) \quad (6)$$

$$b_j(o_t) = \frac{1}{\sqrt{2\pi 2\sigma_j^2}} \exp\left(-\frac{(o_t - \mu_j)^2}{2\sigma_j^2}\right) \quad (7)$$

où :

μ_j est la moyenne de la distribution gaussienne pour l'état j.

σ_j est l'écart-type de la distribution gaussienne pour l'état j.

3) Probabilités initiales des états (π)

La probabilité est en fonction de la séquence des états cachés tels que

$$\pi = \mathcal{P}(q_1 = S_i) \quad (8)$$

C. Initialisation et Entraînement du Modèle HMM

L'entraînement du HMM se fait à partir de la maximisation de la probabilité conjointe des observations et des états cachés. Dans la plupart des càs, l'algorithme Expectation-Maximization (EM) constitué par les deux étapes forward-backward (algorithme de Baum-Welch) est utilisé. [3]

1) Forward algorithme

Cet algorithme est nécessaire pour calculer la probabilité de la séquence d'observations *O* donnée du modèle λ tel que λ définit l'ensemble des paramètres qui définissent complètement le modèle HMM [4].

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 46 No. 2 September 2024, pp. 294-306

$$\alpha_t(i) = \mathcal{P}(o_1, o_2, o_3, \dots, o_t, s_t = s_i \mid \lambda) \quad (9)$$

La valeur initiale est :

$$\alpha_1(i) = \pi_i b_i(O_1)$$

La valeur suivante est :

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right) b_j(o_{t+1}) \quad (10)$$

Ainsi la probabilité de la séquence d'observation est alors :

$$\mathcal{P}(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \quad (11)$$

2) Backward algorithme

L'algorithme backward est essentiel pour ajuster les probabilités de transition et d'émission pour mieux correspondre aux données observées. Ainsi γ et ξ sont des quantités intermédiaires importantes à trouver pour la réestimation des paramètres du HMM, permettant ainsi

$$\beta_t(i) = \mathcal{P}(o_{t+1}, o_{t+2}, o_{t+3}, \dots, o_T \mid q_t = s_i, \lambda) \quad (12)$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, b_j(o_{t+1})\, \beta_{t+1}(j) \quad (13)$$

$\gamma_t(i)$ représente la probabilité que l'état à l'instant t soit i, étant donné la séquence d'observations entière *O* et le modèle λ.

Ainsi la mise à jour du système est

$$\gamma_t(i) = \mathcal{P}(q_t = s_i \mid o, \lambda) \quad (14)$$

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \quad (15)$$

$\xi_t(i,j)$ représente la probabilité de transition de l'état i à l'instant t à l'état j à l'instant t+1, étant donné la séquence d'observations entière *O* et le modèle λ.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 46 No. 2 September 2024, pp. 294-306

$$\xi(i\,,\,j) = \mathcal{P}(q_t = s_i\,,\, q_{t+1} = s_j \mid o,\lambda) \tag{16}$$

$$\xi(i\,,\,j) = \frac{\alpha_t(i)a_{ij}\,b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}\,b_j(o_{t+1})\beta_{t+1}(j)} \tag{17}$$

La réestimation des paramètres se fait ainsi :

$$\pi_i = \gamma_1(i) \tag{18}$$

$$b_j(k) = \frac{\sum_{t=1,o_t=v_k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)} \tag{19}$$

Où $V = \{v_1, v_2, v_3, \dots, v_n\}$ est l'ensemble des vocabulaire d'observations et v_k est le k-ème symbole dans le vocabulaire d'observations.

D. Prédiction des États Cachés

Une fois le modèle HMM entraîné, nous utilisons l'algorithme de Viterbi pour prédire la séquence d'état caché la plus probable étant donné la séquence d'observations.

Algorithme de Viterbi :

Dans cet algorithme, $\delta_t(i)$ est la probabilité de la séquence d'état la plus probable qui se termine à l'état i à l'instant t, étant donné la séquence d'observations jusqu'à l'instant t. Ensuite $\psi_t(j)$ stocke l'argument maximisant $\delta_t(i)$

$$\delta_t(i) = \max_{q_1,q_2,q_3,\dots,q_{t-1}} P(q_1, q_2, q_3, \dots, q_t = s_i\,, o_1, o_2, o_3, \dots, o_t, \mid \lambda) \tag{20}$$

La valeur initiale est :

$$\delta_1(i) = \pi_i b_i(o_1) \tag{21}$$

La valeur suivante est

$$\delta_{t+1}(j) = \max_{i=1,\dots\dots,N}[\delta_t(i)a_{ij}]b_j(o_{t+1}) \tag{22}$$

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 46 No. 2 September 2024, pp. 294-306

Pour terminer, en récupérant la séquence d'état la plus proche, nous avons

$$q_T^* = arg \max_{i=1,\ldots,N} \delta_T(i) \quad (23)$$

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \quad (24)$$

III. RÉSULTATS

A. *Préparation des données*

Soient les observations pour 7 jours suivantes :

Jour d'observation	O_1	O_2
1	0,2	0,1
2	0,4	0,2
3	5,0	4,8
4	0,3	0,2
5	5,5	5,1
6	0,1	0,3
7	0,3	0,4

Les observations pour les jours 3 et 5 sont classées comme anormales.

Ajoutons quelques colonnes sur le tableau précédent pour inclure les observations "normales" et "anormales" .

Jour d'observation	O_1	O_2	Comportement
1	0,2	0,1	Normal
2	0,4	0,2	Normal
3	5,0	4,8	Anormal
4	0,3	0,2	Normal
5	5,5	5,1	Anormal
6	0,1	0,3	Normal
7	0,3	0,4	Normal

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 46 No. 2 September 2024, pp. 294-306

B. Equilibre des données

Nous avons remarqué que les données du tableau ne sont pas équilibrées car le nombre de la classe « Normal » est petit par rapport au nombre de la classe « Anormal » Pour équilibrer ces données, nous allons générer des échantillons synthétiques pour la classe « Anormal » en utilisant SMOTE.

1) Sélection des échantillons minoritaires

Les échantillons minoritaires sont ceux de la classe « Anormal »:

$x_1 = (5.0, 4.8)$, $x_2 = (5.5, 5.1)$

2) Recherche du k-plus proche voisin

Avec k=1, les plus proches voisins sont simplement les autres points minoritaires. Donc pour x_1 le plus proche voisin est x_2 et vice versa.

3) Génération des échantillons synthétiques

a. Pour $x_1 = (5.0,4.8)$ et $x_{i\text{-voisin}} = (5.5, 5.1)$

$x_{synt} = x_1 + \gamma * (x_{i\text{-voisin}} - x_i) = (5.0 , 4.8) + \gamma * ((5.5,5.1) – (5.0,4.8))$

Et pour $\gamma = 0.5$

$x_{synt} = (5.0 , 4.8) + 0.5 * (0.5,0.3) = (5.0,4.8) + (0.25, 0.15) = (5.25,4.95)$

b. Répétons le processus pour générer un autre point synthétique pour $\gamma = 0.7$

$x_{synt} = (5.0 , 4.8) + 0.7 * (0.5,0.3) = (5.0,4.8) + (0.35, 0.21) = (5.35,5.01)$

Ainsi, après SMOTE, nos données équilibrées pourraient ressembler au tableau suivant :

Jour d'observation	O_1	O_2	Comportement
1	0,2	0,1	Normal
2	0,4	0,2	Normal
3	5,0	4,8	Anormal
4	0,3	0,2	Normal
5	5,5	5,1	Anormal
6	0,1	0,3	Normal

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 46 No. 2 September 2024, pp. 294-306

7	0,3	0,4	Normal
8	5,25	4,95	Anormal
9	5,35	5,01	Anormal

C. Initialisation des Paramètres du HMM

1) Matrice de transition des états (A) :

D'après les états du tableau précédent on peut déduire la matrice des états.

Calculons les transitions entre les états :

De Normal à Normal : 2 fois (Jours 1-2, 6-7) = 2/5 = 0,4

De Normal à Anormal : 3 fois (Jours 2-3, 4-5,7-8) 3/5=0,6

De Anormal à Normal : 2 fois (Jours 3-4, 5-6) 2/3 =0,7

De Anormal à Anormal : 1 fois (Jours 8-9) 1/3 =0,3

$A=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ soit $A=\begin{bmatrix} 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$

2) Probabilité d'émission (B)

$$B=\begin{bmatrix} \eta\,(\mu_1, \sigma_1^2) & \eta\,(\mu_2, \sigma_2^2) \\ \eta\,(\mu_3, \sigma_3^2) & \eta\,(\mu_4, \sigma_4^2) \end{bmatrix}$$

Supposons que :

$\mu_1 = 0{,}3 \;\; \sigma_1^2 = 0{,}1^2 \;\; \mu_2 = 0{,}2 \;\; \sigma_2^2 = 0{,}1^2$

$\mu_3 = 5{,}2 \;\; \sigma_3^2 = 0{,}2^2 \;\; \mu_4 = 5{,}0 \;\; \sigma_4^2 = 0{,}2^2$

$B=\begin{bmatrix} 0.27 & 0.2 \\ 5.25 & 4.95 \end{bmatrix}$

Probabilité initiale (π)

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 46 No. 2 September 2024, pp. 294-306

$\pi = [0,6 \quad 0,4]$

Algorithme forward

Pour chaque observation O_t à chaque instant t, calculons $\alpha_t(j)$

Jour 1 :

$\alpha_1(1) = \pi_1 b_1(o_1) = 0.6 * \eta\,(0.2 \mid 0.3, 0.1^2)$

$\alpha 2(2) = \pi_2 b_2(o_2) = 0.4 * \eta\,(0.2 \mid 5.2, 0.2^2)$

Calcule de la valeur de la fonction de densité de probabilité η

- Pour $\eta\,(0.2 \mid 0.3, 0.1^2)$

$$\eta\,(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi*0.01}} exp\left(-\frac{(0.2-0.3)^2}{2*0.01}\right)$$

$$= \frac{1}{\sqrt{0.02}} exp\left(-\frac{0.01}{0.02}\right)$$

$$= \frac{1}{0.141} exp(-0.5)$$

$\approx 2,82*0.606 = 1.708$

- Pour $\eta\,(0.2 \mid 5.2, 0.2^2)$

$$\eta\,(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi*0.04}} exp\left(-\frac{(0.2-5.2)^2}{2*0.04}\right)$$

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 46 No. 2 September 2024, pp. 294-306

$= \frac{1}{\sqrt{0.08}} exp\left(-\frac{25}{0.08}\right)$

$= 1.414\ exp(-312.5)$

$\alpha_1(1) \approx 0.6 * 1.708 = 1.025$

$\alpha_2(1) \approx 0$

Jour2

$$\alpha_2(1) = [\alpha_1(1)\ a_{11} + \alpha_1(2)\ a_{21}]\ b_1(O_2)$$

$$\alpha_2(2) = [\alpha_1(1)\ a_{12} + \alpha_1(2)\ a_{22}]\ b_2(O_2)$$

Calculons $\alpha_2(1)$

$$\alpha_2(1) = [1.025 * 0.7 + 0.04] * \eta(0.4 \mid 0.3, 0.1^2)$$

$$= 0.7175 \frac{1}{0.141} exp\left(-\frac{0.01}{0.02}\right)$$

$$\approx 0.7175 * 2.82 * 0.606 \approx 1.215$$

De même $\alpha_2(2) \approx 0$

Nous pouvons continuer de cette manière pour chaque jour

D. Prédiction des états cachés

1) Algorithme Viterbi

Pour trouver la séquence d'états cachés la plus probable.

$\delta_t(j) = \max_i [\delta_{t-1}(i) a_{ij}] b_j(o_t)$

$\psi_t(j) = \arg\max_i [\delta_{t-1}(i) a_{ij}]$

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 46 No. 2 September 2024, pp. 294-306

Effectuons le calcul pour chaque jour :

Pour j=2 (Anormal) :

$$\delta_1(2) = \pi_2 b_2(O_1) \approx 0$$

Pour j=1

$$\delta_2(1) = \max[\delta_1(1)a_{11}, \delta_1(2)a_{21}]b_1(o_2)$$

$$= \max [1.025 * 0.7, 0* 0.4]] * 1.215 \approx 0.7175 *1.215 \approx 0.872$$

Pour j=2

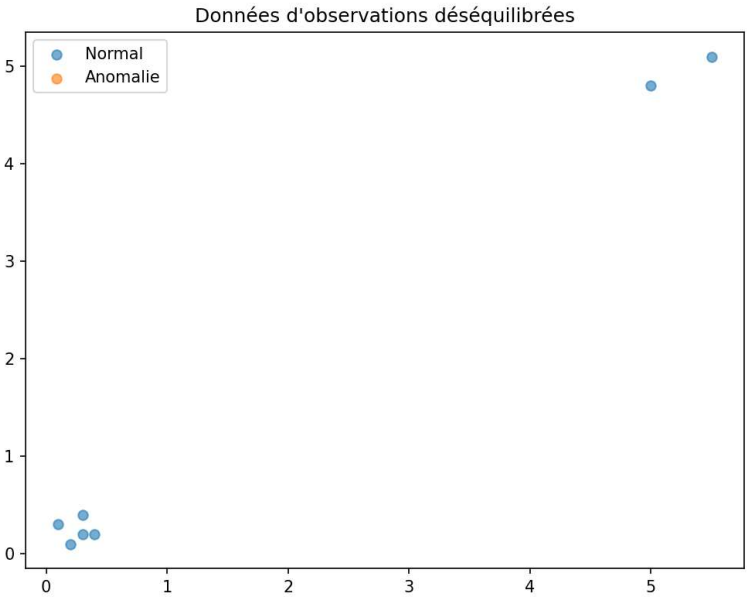
$$\delta_2(2) = 0$$

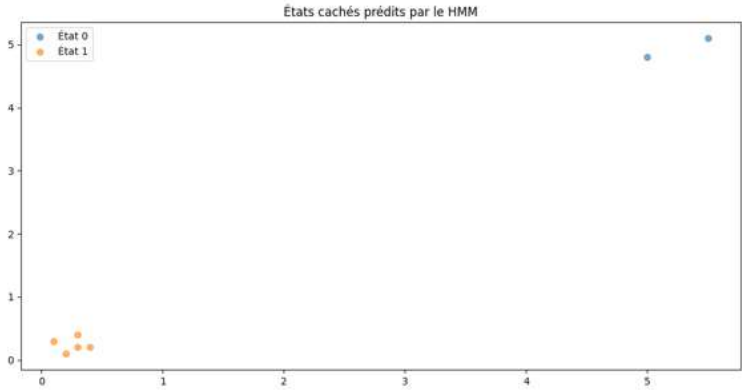
Cette méthode peut être étendue manuellement pour tous les jours et pour ajuster les paramètres du modèle HMM. Les résultats obtenus fourniront la séquence d'états cachés et leur probabilité, permettant ainsi de détecter les anomalies dans les données déséquilibrées. Bien que cette approche manuelle soit pratique pour de petits jeux de données, il est préconisé d'utiliser des outils de programmation pour des jeux de données plus volumineux. Le tableau et la figure suivante montrent à la fois la representation des données d'observations sur les 7 jours et l'etats cachés prédits par le HMM.

Jour d'observation	O_1	O_2	Comportement	Etat caché prédit HMM (etat caché viterbi)
1	0,2	0,1	Normal	0
2	0,4	0,2	Normal	0
3	5,0	4,8	Anormal	1
4	0,3	0,2	Normal	0
5	5,5	5,1	Anormal	1
6	0,1	0,3	Normal	0
7	0,3	0,4	Normal	0

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 46 No. 2 September 2024, pp. 294-306





IV. Interprétation et discussion

Les formules mathématiques de SMOTE permettent de générer des données synthétiques pour équilibrer les classes déséquilibrées. En appliquant ces formules, nous avons équilibré les données de l'observation de 7 jours en ajoutant des instances synthétiques pour la classe minoritaire. Cette approche permet de rendre le modèle de machine learning plus robuste et moins biaisé envers la classe majoritaire.

1) Interprétation des États Cachés Prédits

a- Jour 1, 2, 4, 6, 7 :

- Les observations sont proches de (0.3, 0.2) et (0.2, 0.1), ce qui est cohérent avec la distribution normale de la classe majoritaire (*Normal*).
- Le modèle HMM a prédit l'état "*Normal*" (État 0) pour ces jours.

b- Jour 3, 5 :

- Les observations sont autour de (5.0, 4.8) et (5.5, 5.1), ce qui est significativement différent de la classe *Normale*.
- Le modèle HMM a correctement identifié ces jours comme étant "*Anormal*" (État 1).

2) Limites Actuelles de SMOTE :

Pour la sensibilité aux paramètres, SMOTE dépend fortement du choix des paramètres, tels que le nombre de voisins k et la manière dont les nouveaux exemples synthétiques sont générés. Un mauvais réglage de ces paramètres peut conduire à des résultats non optimaux.

Bien que SMOTE soit efficace pour augmenter la taille de l'échantillon de la classe minoritaire, il peut aussi introduire du bruit et des *artefacts* dans les données synthétiques, ce qui pourrait affecter négativement la performance du modèle.

L'efficacité de SMOTE peut varier en fonction de la distribution spécifique des données et de la nature des classes minoritaires. Dans certains cas, d'autres techniques d'équilibrage des données pourraient être nécessaires pour obtenir de meilleurs résultats.

V. CONCLUSION

L'intégration de SMOTE avec HMM pour traiter les données déséquilibrées représente une approche prometteuse mais nécessite une compréhension approfondie des défis spécifiques et des stratégies d'amélioration adaptées. En explorant des techniques avancées de sur-échantillonnage, en optimisant les paramètres de SMOTE et en évaluant rigoureusement la performance du modèle HMM, il est possible de surmonter les limites actuelles et de maximiser l'efficacité de ces méthodes dans divers contextes d'application en apprentissage automatique.

Bien que SMOTE soit une méthode largement utilisée et efficace pour traiter les données déséquilibrées, il existe des défis et des opportunités d'amélioration à considérer. L'intégration de techniques avancées et l'approfondissement de la recherche permettront de surmonter les limites actuelles et d'adapter ces méthodes aux défis futurs en matière d'apprentissage automatique et de traitement des données. Il est recommandé de choisir la méthode d'équilibrage des données en fonction des caractéristiques spécifiques des données et des objectifs de modélisation pour obtenir les meilleurs résultats possibles.

Enfin pour mettre en œuvre des stratégies d'évaluation rigoureuses pour comparer différentes approches de gestion des données déséquilibrées, il est nécessaire de considérer des autres méthodes de mesures de performance appropriées comme la matrice de confusion, le score F1 et les courbes ROC, spécifiques aux séquences de données et aux modèles HMM.

REFERENCES

[1] Laurent. Rouvière, “Donnés désquilibrées,” CRNS, cours, Novembre 2023.

[2] O. Olawale Awe, “Computational Strategies for Handling Imbalanced Data in Machine Learning”, VP-IASE VP of Global Engagement, -LISA 2020 Global Network, USA. p.10

[3] Alperen Degirmenci, “Introduction to Hidden Markov Models,” Harvard University, cours 2014.

[4] Daniel Jurafsky & James H. Martin, “Hidden Markov Models,” Standford University, Draft of August 20, 2024.