Optimisation de l'apprentissage des données météorologique déséquilibrées basé sur la modélisation du formule d'Itô et le k-NN

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

[1] *Vakinankaratra University, Madagascar*
[2] *Laboratory Manager, SE-I-MSDE, ED-STII, Antananarivo, Madagascar*

ABSTRACT

Les données météorologiques sont souvent déséquilibrées en raison de la rareté de certains événements extrêmes comme les tempêtes, les sécheresses. L'apprentissage sur de telles données nécessite des techniques spécifiques pour éviter le biais et améliorer les performances du modèle.

Le but de cet article est d'utiliser des données équilibrées avant de faire la prédiction météorologique via les méthodes classique LSTM, ou GCM. Cette méthode est basée sur la modélisation de la formule d'Itô combinée avec la méthode de k-proches voisins.

Keyword: *Apprentissage automatique, données déséquilibrées, formule Itô, k-proches voisins, LSTM*

1. INTRODUCTION

Les données météorologiques déséquilibrées représentent un défi majeur dans l'analyse et la prédiction des événements climatiques. Ces données se caractérisent par une distribution inégale des classes [1], où certains types d'événements sont beaucoup plus fréquents que d'autres. Ce déséquilibre peut affecter la performance des modèles du machine learning en introduisant des biais et en réduisant la capacité des modèles à prédire les événements rares mais souvent critiques.

En combinant la modélisation stochastique avec des techniques d'apprentissage automatique adaptées aux données déséquilibrées, on peut améliorer la précision des prédictions météorologiques, en particulier pour les événements extrêmes.

2. METHODOLOGIE

Quelques étapes importantes sont nécessaires pour faire un bon apprentissage des données météorologiques déséquilibrées dont la définition du jeu de données météorologiques déséquilibrées, le traitement spécifique de ces données via K-NN et enfin l'intégration de la formule d'Itô pour modéliser et prévoir les variables météorologiques comme la température.

2.1 Equilibrage des données

Les données déséquilibres arrivent très fréquemment dans les données météorologiques, et il est vital de les considérer afin d'optimiser la méthode d'entraînement d'un modèle de Machine Learning. Le but est de concevoir un modèle résultant non biaisé vis-à-vis de la (ou des) classes(s) majoritaire(s). Ainsi il est primordial d'équilibrer les données brutes, de les modéliser avant d'appliquer la formule d'Îto.

La méthode de K-proches voisins est utilisée pour la rééchantillonnage de chaque exemple minoritaire. Ainsi pour un point x_i de la classe minoritaire, le travail est de calculer la distance euclidienne avec les autres points de la même classe et de sélectionner les k plus proches voisins [2] basées sur la formule suivante :

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (1)$$

L'étape suivante consiste à générer des données supplémentaires en créant des exemples synthétiques basés sur les k-voisins. Pour chaque voisin x_j de x_i , tel que i≠j, alors nous allons générer un nouveau point synthétique x_{synt} tel que :

$$x_{synt} = x_{moyenne} + \lambda\,(x_j - x_i) \quad (2)$$

Avec λ est un nombre aléatoire entre 0 et 1 et $x_{moyenne}$ la moyenne de x_i prix en fonction de k qui est le nombre de voisin à considérer autour de x_i

2.2 Modélisation des Variables Météorologiques

a) Variable météorologiques

La modélisation des processus stochastiques en météorologie est une approche conçue pour la variabilité et l'incertitude intrinsèque aux phénomènes atmosphériques. Ces modèles sont nécessaires pour simuler l'évolution des variables météorologiques clés tel que la température, la précipitations, l'humidité, les vents, de manière aléatoire et surtout en fonction des données historiques, tout en considérant les variations dues à des facteurs non déterministes. Plusieurs approches peuvent être utilisées dont les plus connues sont : processus de Wiener et modèles diffusionnels, processus de poisson et comptage des événements rares, modèles équations différentielles stochastiques, Modèle ARMA et ARIMA pour les séries temporelles, la chaîne de Markov en météorologie et la simulation de Monte Carlo.

b) Evénements rares

Les événements suivants sont considérés comme événements rares en météorologie : que les tempêtes, les ouragans, les tornades, les inondations, ou les vagues de chaleur extrêmes. Ces phénomènes météorologiques peu fréquents mais potentiellement très impactant peuvent être analyser par les déviations stochastiques des événements rares. Ces déviations sont cruciales pour la prévision, la gestion des risques et la planification des mesures d'adaptation. Voici quelques approches pour avoir cette déviation : distribution de poisson et processus poisson, distribution des valeurs extrèmes (extreme value theory, EVT), processus de diffusion stochastiques (EDP), modèles ARIMA et leurs extensions, simulation de Monte Carlo, le modèle de Markov cachés (Hidden Markov Models, HMM).

2.3 Application de la Formule d'Itô

Dans le contexte des données météorologiques, la formule d'Itô est utilisée pour modéliser les variations des variables météorologiques en fonction du temps. Ensuite, il est nécessaire de définir les conditions initiales et les paramètres de dérive et de volatilité spécifiques à chaque variable météorologique. La formule d'Itô peut être ainsi adaptée pour modéliser des phénomènes stochastiques tels que les variations de température ou les précipitations.

La formule d'Itô est une généralisation de la règle de dérivation pour les fonctions des processus stochastiques [4]. En réalité, elle est essentielle pour tracer la variation d'une fonction d'un processus stochastique, comme le mouvement brownien. Alors, elle est généralement utilisée pour modéliser les phénomènes aléatoires.

Pour un processus stochastique P_t, la formule d'Îto pour une fonction f (P_t, t) est comme suit :

$$df\,((P_t, t) = \left(\frac{\partial f}{\partial t} + \mu(t, P_t)\frac{\partial f}{\partial P_t} + \frac{1}{2}\sigma^2(t, P_t)\frac{\partial^2 f}{\partial P_t^2}\right)dt + \sigma(t, P_t)\frac{\partial f}{\partial P_t}\,dW_t \quad (3)$$

avec :

- P_t est un processus stochastique comme la température, la pression ou l'humidité
- $\mu(t, P_t)$ est la dérivé du processus
- $\sigma(t, P_t)$ est la volatilité du processus
- W_t est un mouvement brownien standard

μ utilisé dans la dérivé du processus est la moyenne des changements quotidiens du processus stochastique (T_i dans notre exemple est la température du jour i)

$$\mu = \frac{1}{N-1}\sum_{t=1}^{N-1}(T_{t+1} - T_t) \quad (4)$$

σ utilisé dans la volatilité est l'ecart-type des changements quotidiens de la température

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{t=1}^{N-1}(T_{t+1} - T_t - \mu)^2} \quad (5)$$

3. RESULTATS

Supposons que nous avons les données fictives suivantes pour 30 jours d'observations météorologiques :

Jour	Température
1	24
2	22
3	23
4	21
5	25
6	26
7	28
8	27
9	29
10	30
11	35
12	36
13	30
14	29
15	28
16	27

17	26
18	25
19	24
20	23
21	22
22	21
23	20
24	19
25	18
26	17
27	16
28	15
29	14
30	5

Les températures extrêmes sont :

- Élevée : 36°C
- Basse : 5°C

3.1 Équilibrage par sur-échantillonnage (SMOTE)

L'équilibrage des données se fait par génération de données synthétiques. Pour les températures extrêmes, nous allons utiliser KNN avec régularisation λ.

a- Pour la température la plus élevée qui est 36°C nous avons :

- Identification des voisins pour k=4
  k=4 plus proches voisins : [35, 30, 30, 29]
- Calcul de la Moyenne des Voisins :

$$M = \frac{35+30+30\,29}{4} = 31$$

- Génération de Points Synthétiques avec λ=0.5
  Valeurs aléatoires x_i : [1, -1, 2, -2]
  Variations $\lambda\, x_i$: [0.5, -0.5, 1, -1]
  Nouveaux Points :
  31 + 0.5 = 31.5
  31 - 0.5 = 30.5
  31 + 1 = 32
  31 - 1 = 30

b- Pour la température la plus basse qui est 5°C nous avons :

- Identification des voisins pour k=4
  k=4 plus proches voisins : [14, 15, 16, 17]
- Calcul de la Moyenne des Voisins :

$$M = \frac{14+15+16+17}{4} = 15{,}5$$

- Génération de Points Synthétiques avec λ=0.5

Valeurs aléatoires x_i : [1, -1, 2, -2]
Variations λx_i : [0.5, -0.5, 1, -1]
Nouveaux Points :
15,5 + 0.5 = 16
15,5 - 0.5 = 15
15,5 + 1 = 16,5
15,5 - 1 = 14,5

3.2 Prévision avec la Formule d'Itô

- Calcul des paramètres
  - Variations Quotidiennes (ΔT_i) :

Jour	Température	$\Delta T_i = T_{i+1} - T_i$
1	22	ΔT_1 =24-22 = 2
2	24	ΔT_2 =23-24 = -1
3	23	ΔT_3 =21-23 = -2
4	21	ΔT_4 =25-21 = 4
5	25	ΔT_5 =26-25 = 1
6	26	ΔT_6 =28-26 = 2
7	28	ΔT_7 =27-28 = -1
8	27	ΔT_8 =29-27 = 2
9	29	ΔT_9 =30-29 = 1
10	30	ΔT_{10} =35-30 = 5
11	35	ΔT_{11} =36-35 = 1
12	36	ΔT_{12} =30-36 = -6
13	30	ΔT_{13} =29-30 = -1
14	29	ΔT_{14} =28-29 = -1
15	28	ΔT_{15} =27-28 = -1
16	27	ΔT_{16} =26-27 = -1
17	26	ΔT_{17} =25-26 = -1
18	25	ΔT_{18} =24-25 = -1
19	24	ΔT_{19} =23-24 = -1
20	23	ΔT_{20} =22-23 = -1
21	22	ΔT_{21} =21-22 = -1
22	21	ΔT_{22} =20-21 = -1
23	20	ΔT_{23}=19-20 = -1
24	19	ΔT_{34} =18-19 = -1
25	18	ΔT_{25} =17-18 = -1
26	17	ΔT_{26} =16-17 = -1
27	16	ΔT_{27} =15-16 = -1
28	15	ΔT_{28} =14-15 = -1
29	14	ΔT_{29} =5-14 = -9

30	5	ΔT_{30} =31,5-5 = 26,5
31	31,5	ΔT_{31} =30,5-31, 5= -1
32	30,5	ΔT_{32} =32-30,5 = 1,5
33	32	ΔT_{33} =30-32 = -2
34	30	ΔT_{34} =16-30 = -14
35	16	ΔT_{35} =15-16 = -1
36	15	ΔT_{36} =16,5-15 = 1,5
37	16,5	ΔT_{37} =14,5-16,5 = -2
38	14,5	

- Moyenne des changements quotidiens

$$\mu = \frac{1}{N-1}\sum_{t=1}^{N-1} \Delta T_i$$

$\mu = \frac{1}{37}$ (2-1-2+4+1+2-1+2+1+5+1-6-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-9+26,5-1+1,5-2-14-1+1,5-2)

$\mu = \frac{1}{37}(-7,5) \approx -0,20$

- Écart-type des Variations (σ) :

$$\sigma = \sqrt{\frac{1}{37}}((2+0,203)^2 + (-1+0,203)^2 + (-2+0,203)^2 + (4+0,203)^2 + (-1+0,203)^2 + (2+0,203)^2 + (2+0,203)^2 + (2+0,203)^2 + (2+0,203)^2 + (2+0,203)^2 + (2+0,203)^2 + (2+0,203)^2 + (-1+0,203)^2 + (-1+0,203)^2 + (-1+0,203)^2 + (-1+0,203)^2 + (-1+0,203)^2 + (-1+0,203)^2 + (-1+0,203)^2 + (-1+0,203)^2 + (-1+0,203)^2 + (-1+0,203)^2 + (-1+0,203)^2 + (-1+0,203)^2 + (-1+0,203)^2 + (-1+0,203)^2 + (-9+0,203)^2 + (26,5+0,203)^2 + (-1+0,203)^2 + (1,5+0,203)^2 + (-2+0,203)^2 + (-14+0,203)^2 + (-1+0,203)^2 + (1,5+0,203)^2 + (-2+0,203)^2)$$

$$\sigma = \sqrt{175,16} \approx 13,24$$

- Prédiction des Températures avec la Formule d'Itô

L'intégration de l'équation d'Itô est :

$$P_{t+1} = P_t + \mu .dt + \sigma.\sqrt{dt} . Z_t$$

où Z_t est une variable aléatoire tirée d'une distribution normale standard η(0,1)).

Simulation

Supposons que nous souhaitons prédire les températures pour les 10 prochains jours à partir de P_0=14.5. Nous utiliserons μ≈ -0.20 et σ≈13.24.

Pour chaque jour t, nous calculons :

$P_{t+1}=P_t - 0.20 + 13{,}24 \cdot Z_t$

Les valeurs de Zt dans l'exemple précédent sont des variables aléatoires tirées d'une distribution normale standard, notée η(0,1)

Z_t=[0.5,−0.3,1.2,−0.8,0.7,0.4,−0.1,-1.5,−0.4,0.3]

Soit

$P_1 = 14{,}5 - 0{,}20 + 13.24*0{,}5 \approx 20{,}92$
$P_2 = 20{,}92 - 0{,}20 + 13.24*(-0{,}3) \approx 16{,}75$
$P_3 = 16{,}75 - 0{,}20 + 13.24*1{,}2 \approx 32{,}44$
$P_4 = 32{,}44 - 0{,}20 + 13.24*(-0{,}8) \approx 21{,}65$
$P_5 = 21{,}65 - 0{,}20 + 13.24*0{,}7 \approx 30{,}72$
$P_6 = 30{,}72 - 0{,}20 + 13.24*0{,}4 \approx 35{,}82$
$P_7 = 35{,}82 - 0{,}20 + 13.24*(-0{,}1) \approx 34{,}23$
$P_8 = 34{,}23 - 0{,}20 + 13.24*(-1{,}5) \approx 14{,}17$
$P_9 = 14{,}17 - 0{,}20 + 13.24*(-0{,}4) \approx 8{,}67$
$P_{10} = 8{,}67 - 0{,}20 + 13.24*0{,}3 \approx 12{,}44$

Les températures prédites pour les 10 prochains jours seraient donc approximativement :

$T_1 \approx 20{,}92$
$T_2 \approx 16{,}75$
$T_3 \approx 32{,}44$
$T_4 \approx 21{,}65$
$T_5 \approx 30{,}72$
$T_6 \approx 35{,}82$
$T_7 \approx 34{,}23$
$T_8 \approx 14{,}17$
$T_9 \approx 8{,}67$
$T_{10} \approx 12{,}44$

4. INTERPRETATION ET DISCUSSION

- Variations quotidiennes :

Basé sur les résultats précédents, les prédictions nous présentent des variations significatives d'un jour à l'autre, ce qui est spécifique pour les processus stochastiques. En réalité, la température passe de 20.92 à 16.75, puis de 16.75 à 32.44, ce qui montre des augmentations et des diminutions importantes, reflétant la volatilité de la température.

• Tendance moyenne :

La tendance moyenne, dérive de μ, est très faible avec une valeur de $\mu \approx 0.20$. Ainsi, en l'absence de fluctuations aléatoires, la température augmenterait très légèrement chaque jour.

• Fluctuations stochastiques :

La volatilité σ qui est environ ≈ 13,24 introduit des fluctuations significatives autour de la tendance moyenne. Les valeurs de Zt provenant d'une distribution normale standard η (0,1) accentuent aussi ces fluctuations aléatoires.

• Comportement des extrêmes :

Des températures plus extrêmes ont été observées. Nous avons une augmentation à 35.82 (T_6) et une diminution à 08.67 (T_9). Ces températures extrêmes reflètent la capacité du modèle à capturer des événements météorologiques rares ou extrêmes, qui sont vraiment important dans les études climatologiques.

• Consistance avec les données historiques :

Les prédictions montrent des comportements similaires en termes de fluctuations. Les données historiques étalaient des variations importantes, et ces variations sont également visibles dans les prédictions.

5. CONCLUSIONS

En utilisant la formule d'Itô et les techniques d'équilibrage des données, nous avons pu fiabiliser la modélisation et la prédiction des températures dans un jeu de données météorologiques déséquilibrées. Notre approche permet de gérer les données déséquilibrées et d'améliorer la précision des prévisions, en particulier pour les événements météorologiques rares. La formule d'Itô est ensuite appliquée à ces données équilibrées et nous avons obtenu des prévisions plus précises et robustes.

L'application de la formule d'Itô en continue permet de modéliser les prévisions des températures en intégrant à la fois une tendance moyenne et des fluctuations aléatoires. Les prédictions de la formule d'Itô montrent comment les températures peuvent varier de manière significative et stochastique autour d'une tendance moyenne très faible. Les valeurs obtenues sont influencées par des variables aléatoires qui introduisent de la volatilité, reflétant ainsi la nature imprévisible des conditions météorologiques.

Cette approche permet de mettre en évidence l'importance de l'équilibrage des données avant d'utiliser un modèle mathématique. Elle nous permet ainsi de modéliser seulement les tendances générales et aussi les événements extrêmes, offrant ainsi une vue d'ensemble des possibles évolutions de la température dans un cadre stochastique. Dans l'analyse des données météorologiques, il est courant de rencontrer des phénomènes non stationnaires et non linéaires. Pour bien modéliser ces séries temporelles, il est nécessaire d'appliquer des techniques adaptées à ces comportements complexes, comme la différenciation pour la stationnarité, et des modèles non linéaires pour capturer la complexité des relations climatiques dont les modèles SARIMA, réseaux de neurones (LSTM) et des forêts aléatoires pour les prévisions météorologiques à court termes et les modèles physiques (GCM) pour les prévisions météorologiques à long termes ou des modèles hybrides utilisant des vraies données historiques.

6. REFERENCES

[1]. Laurent. Rouvière, "Donnés déséquilibrées," CRNS, cours, Novembre 2023.
[2]. Tom M. Mitchell, "Machine Learning". McGraw-Hill Science/Engineering/Math, p.232, March, 1997
[3]. Garry Toth et Al. "Environnement Canada - Manuel sur le brouillard et la prévision du brouillard", p.18
[4]. Jean-Christophe Breton, "Calcul stochastique", Université Rennes 1, cours, Octobre -Décembre 2021