


Detectarea anomaliilor în serii de timp



Eftimie Petre-Laurențiu, 351
Buzaș Radu-Gabriel, 351
Coman Tudor, 351
Luculescu Ștefan, 311



Cuprins

- Problema abordată
- Use case
- Soluții (metode statistice, FFT, procese Gaussiene, ARMA, machine learning, deep learning)
- Bibliografie



Problema abordată

- detectarea anomaliilor în serii de timp, aplicată pe prețuri de acțiuni listate la bursă
- în acest caz, termenul de “anomalie” se referă la deviații semnificative de la tendințele tipice ale seriei de timp respective, fiind un indicator pentru evenimente neobișnuite sau pentru oportunități de tranzacționare
- este o problemă care poate fi abordată în multe moduri (procesarea semnalelor, modele AI, modele matematice și stochastice, reinforcement learning șamd.)



Use case-uri pentru detectarea anomaliilor pe bursă

Acest proces este esențial pentru instituțiile financiare, de toate mărimile, care tranzacționează active (pe piața de capital sau nu), pentru că are o arie largă de aplicații:

- detectarea fraudei sau a erorilor de sistem
- urmărirea pieței de către organe de control
- risk management
- trading algoritmic
- analiză predictivă sau macroeconomică
- administrarea/ diversificarea unui portofoliu (în general ca instituție)



Soluții tehnice

- metode statistice
- machine learning
- deep learning



Metode statistice

- eliminarea trendului din seria de timp
- detectarea anomaliilor
 - metoda z-score
 - metoda medie mobilă
 - metoda de deviație medie absolută
 - metoda procentuală



Eliminarea trendului din seria de timp

Metoda regresiei polinomiale

1. Valorile seriei de timp sunt reprezentate într-un vector $\mathbf{v} = [v_0, v_1, \dots, v_{n-1}]^T$
2. Momentele de timp sunt reprezentate într-un vector $\mathbf{t} = [t_0, t_1, \dots, t_{n-1}]^T$
3. Presupunem că trendul seriei este un polinom P de grad mic ($d \in \{1, 2, 3, 4\}$)
4. Rezolvăm sistemul liniar $P(t_i) = v_i, i \in \{0, 1, \dots, n-1\}$ în sensul celor mai mici pătrate pentru a determina coeficienții polinomului
5. Seria fără trend este $\mathbf{r} = [r_0, r_1, \dots, r_{n-1}]^T, r_i = v_i - P(t_i)$



Eliminarea trendului din seria de timp

Metoda regresiei polinomiale - rezolvarea sistemului

1. Polinomul P este reprezentat prin vectorul de coeficienți

$$c = [c_0, c_1, \dots, c_d]^T, P = c_0 + c_1X + \dots + c_dX^d$$

2. Matricea sistemului este $A \in M_{n,d+1}(\mathbb{R})$, $A_{i,j} = c_i^j, i \in \{0, 1, \dots, n-1\}, j \in \{0, 1, \dots, d\}$

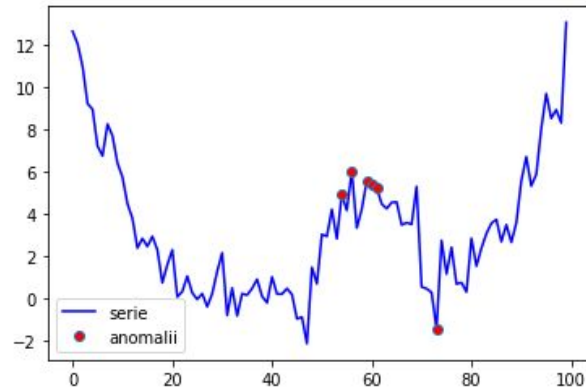
3. În ipoteza $n - 1 > d$ (seria de timp este lungă, iar gradul polinomului este mic), matricea A are rang $d+1$ și sistemul $Ac = v$ admite soluție unică în sensul celor mai mici pătrate (c minimizează norma vectorului $v - Ab$, unde b este un vector din \mathbb{R}^{d+1})

4.
$$c = (A^T A)^{-1} A^T v$$

Detectarea anomaliilor

Metoda z-score

1. Se lucrează pe serii fără trend $r = [r_0, r_1, \dots, r_{n-1}]^T$
2. Se calculează media $m = \frac{1}{n} \sum_{i=0}^{n-1} r_i$ și deviația standard $s = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} (r_i - m)^2}$
3. Se calculează $z = [z_0, z_1, \dots, z_{n-1}]^T$, $z_i = \frac{1}{s} (r_i - m)$
4. Se consideră anomalii valorile z_i cu modul mai mare decât un anumit prag (de obicei 2 sau 3 sau chiar și mai mic)





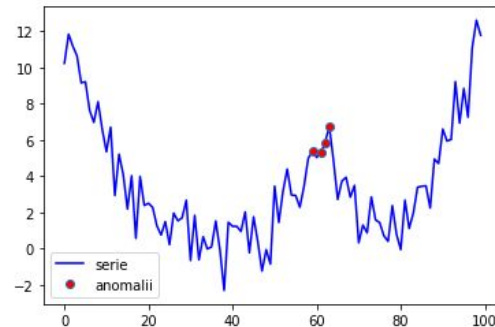
Detectarea anomaliiilor

Metoda medie mobilă

1. Se lucrează pe serii fără trend $r = [r_0, r_1, \dots, r_{n-1}]^T$
2. Se calculează media mobilă pe o fereastră glisantă de lungime fixată d
3. După ce se calculează diferența față de medie pe toate ferestrele, valorile care depășesc un anumit prag sunt considerate anomalii



Detectarea anomaliilor



Metoda de deviație medie absolută

1. Se lucrează pe serii fără trend $r = [r_0, r_1, \dots, r_{n-1}]^T$
2. Se calculează mediana seriei de timp. Mediana (p) este o valoare aleasă astfel încât jumătate din valori să fie $\geq p$ și jumătate din valori să fie $\leq p$ (în practică vom alege mijlocul vectorului sortat). Mediana nu coincide neapărat cu media
3. Se calculează deviația absolută $e = [e_0, e_1, \dots, e_{n-1}]^T, e_i = |r_i - p|$
4. Se calculează deviația absolută medie $m = \frac{1}{n} \sum_{i=0}^{n-1} e_i$
5. Se consideră anomalii valorile e_i mai mari decât un anumit prag stabilit în funcție de m .



Detectarea anomaliiilor

Metoda procentuală

1. Presupunem că apar anomalii pe un anumit procent (cunoscut) din serie (de exemplu, procent aproximat de o metodă Fourier)
2. Se aplică una dintre metodele anterioare cu diferența că pragul se ajustează pentru a obține procentul dorit de anomalii



FFT pentru anomalii regionale

- aproximăm o nouă curba peste semnalul inițial aplicând FFT și apoi IFFT cu un număr redus de parametri (ex. avem 32 de valori inițiale, folosim doar 6)
- pentru fiecare eșantion, calculăm diferența între semnalul original și cel aproximat
- punctele a căror diferență este mai mare decât media diferențelor devin suspecte pentru a fi anomalii
- ne asigurăm că sunt anomalii calculând z-score pentru fiecare punct, având în vedere vecinii săi
- punctele cu z-score peste un prag sunt anomalii
- ca să găsim regiunile ce conțin anomalii, căutăm 2 puncte în succesiune ce au z-score de semn opus (ex. după o rampă, urmează o pantă)
- merge bine dacă schimbările de frecvență sunt bruște



Procese gaussiene / ARMA

- folosim setul de antrenare pentru a găsi hiperparametrii optimi
- medie + covarianță folosind eventual MLE pentru procese gaussiene (Ornstein–Uhlenbeck)
- P parametri autoregresivi și q termeni pentru media glisantă la ARMA folosind Grid Search
- prezicem următoarele n puncte și apoi comparăm cu datele actuale
- dacă diferența pentru un eșantion depășește un prag, atunci este anomalie
- pentru prag putem lua multipli de deviația standard sau quantiles din setul de diferențe
- ARMA merge bine dacă datele din viitor pot fi modelate cu un număr finit de puncte din trecut
- procesele gaussiene merg bine dacă datele provin dintr-o distribuție similară cu cea presupusă



Machine Learning

Supervised:

- Random Forests
- K-NN Regression - simplu

Unsupervised:

- Isolation Forests
- One Class SVM

Atât Isolation Forests cât și One Class SVM au ca scop principal detectarea anomaliilor



Deep Learning?

- Metoda 'eficientă' în privința predicțiilor
- Rețelele Neurale au capacitatea de a aproxima orice funcție
- Rețelele Neurale modelează anumite trăsături care par total aleatoare pentru oameni
- Arhitectura NARX (Non-Linear Autoregressive with exogenous input) a furnizat rezultate promițătoare
- LSTM sunt bune în contextul analizei seriilor de timp prin capacitatea lor de a modela și captura modele de date complexe
- Autoencoders sunt utili în situația de unsupervised learning
- Combinarea mai multor modele poate avea rezultate pozitive (ex.: LSTM + Autoencoders)



Prophet?

- Model dezvoltat de Facebook pentru analiza seriilor de timp
- Model Open-Source
- Non linear regression
- Separarea componentelor de trend, sezoniere și reziduale
- Nu utilizează modele de ML
- Prophet + Deep Learning?



Bibliografie

- <https://dl.acm.org/doi/10.5555/1789574.1789615>(FFT)
- <https://www.researchgate.net/publication/358425639> Metodologie de analiza a seriilor de timp cu aplicatii in modelarea si predictia datelor biomedicale si de sanatate publica
- <https://medium.com/@akashsri306/detecting-anomalies-with-z-scores-a-practical-approach-2f9a0f27458d>
- https://en.wikipedia.org/wiki/Laplace_distribution
- <https://arxiv.org/pdf/2306.12969.pdf>
- <https://peerj.com/preprints/3190.pdf>
- https://web.njit.edu/~usman/courses/cs675_fall18/10.1.1.441.7873.pdf
- <https://www.kaggle.com/code/gauravduttakiit/predicting-stock-prices-using-facebook-s-prophet>