INTEGRATING PCA WITH DEEP LEARNING MODELS FOR STOCK MARKET FORECASTING

Členovia skupiny: Zuzana Orinčáková & Richard Póša
Zdroj: https://doi.org/10.1016/j.jksuci.2024.102162

Riešený problém

- Článok sa zaoberá presnejším predpovedaním cien akcií na tureckom trhu, ktorý je zložitý z dôvodu vysokej volatility, nelineárnych trendov a šumu vo finančných údajoch, pomocou redukcie počtu metrík metódou metódy PCA a ich navrhnutému modelu. Podľa nich tradičné metódy predikcie majú problém zvládnuť tieto zložitosti za použitia množstva podobných metrík, keďže môžu viesť k preučeniu model, čo často vedie k nepresným predpovediam. Integráciou PCA na redukciu príznakov s pokročilými modelmi hlbokého učenia (ako sú LSTM a CNN) je cieľom štúdie zlepšiť predpovedanie pohybu cien akcií výberom najrelevantnejších technických ukazovateľov a zvýšiť výpočtovú efektívnosť a presnosť týchto predpovedí.

Použitá metóda a jej porovnanie so state of the artom

- Navrhovaná metóda je kombinácia analýzy principiálnych komponentov (PCA) na výber príznakov, v spojení s architektúrami hlbokého učenia LSTM a CNN. Tento hybridný prístup sa zameriava na kľúčové problémy, ako je šum v údajoch, volatilita a nelineárna povaha trendov na akciovom trhu. V porovnaní s tradičnými prístupmi je PCA jedinečným doplnkom pri predpovedaní vývoja na burze, pretože zjednodušuje dimenzie prvkov, čím zefektívňuje modely hlbokého učenia a znižuje riziko preučenia sa modelu odstránením nadbytočných informácií.
- Predchádzajúce state of art metódy sa vo veľkej miere spoliehali buď na samostatné modely strojového učenia, ako sú Support Vector Machines (SVM) a základné RNN, alebo na štatistické metódy, ako je ARIMA, ktoré často zápasia s nelineárnymi charakteristikami údajov akciového trhu. Avšak, z vlastnej skúsenosti vieme povedať pre akcionárov sú metódy založené na matematických modeloch atraktívnejšie, vďaka tomu, že sú vysvetliteľné a vieme sa pozrieť na proces rozhodovanie. Taktiež je okrem samotnej technickej analýzy dôležitejší sentiment informácii a správ, ktorý vychádza o daných firmách, ako aj ich výkonnostné metriky, čo ovplyvňujú rozhodnutia akcionárov.

- Na rozdiel od podobných prístupov hlbokého učenia táto štúdia využíva PCA-LSTM-CNN, čo je kombinácia, ktorá sa spolu bežne nepoužíva. Táto integrácia umožňuje CNN zachytiť krátkodobé závislosti, zatiaľ čo LSTM spracúva dlhodobé vzory, čo vedie k lepšej presnosti predikcie. Najnovšie modely, ako napríklad hybrid CNN-LSTM s mechanizmami pozornosti, pridávajú ďalšie vrstvy na zlepšenie extrakcie príznakov a zlepšenie predikčných schopností pri akciových trendoch. Okrem toho niektoré novšie štúdie použili analýzu sentimentu a empirický rozklad módu (EMD) s PCA a LSTM, čím ďalej spresnili predikciu zohľadnením sentimentu správ v reálnom čase spolu s údajmi o akciách, čo by mohlo byť doplnkovým prístupom k tomuto modelu.

Použité dáta

- Štúdia využíva historické údaje o cenách akcií piatich spoločností (ASELS.IS, TUPRS.IS, THYAO.IS, SISE.IS, FROTO.IS) na Istanbulskej burze cenných papierov za obdobie 10 rokov, od 1. januára 2014 do 1. januára 2024. Tento súbor údajov pochádza z Yahoo Finance a obsahuje polia dátum, otváracia cena, najvyššia cena, najnižšia cena, záverečná cena, upravená záverečná cena a objem obchodovania.
- Okrem zdrojových údajov o cenách akcií sa v štúdii vypočítali hodnoty desiatich bežne používaných technických ukazovateľov a to vážnený kĺzavý priemer (WMA), exponenciálny kĺzavý priemer (EMA), index relatívnej sily (RSI), Chandeho oscilátor hybnosti (CMO), Williamsov percentuálny rozsah (WILLR), miera zmeny (ROC), kĺzavý priemer trupu (HMA), trojitý exponenciálny kĺzavý priemer (TEMA), priemerný smerový index (ADX) a psychologická čiara (PLine).

Metodológia trénovania a vykonaných experimentov

- Metóda ktorú zvolili je veľmi zaujímavá a principiálne vhodná, 10 rokov dát a technické ukazovatele je dostatočné množstvo informácii na technickú analýzu trhu, no keď sa pozrieme na graf týchto ukazovateľov môžeme si všimnúť že poniektoré sú zle vyrátané (ako napríklad PLine, ktorá podľa vzorca nemôže nadobúdať hodnotu vyššiu ako 100, avšak v ich grafe dosahuje hodnotu 250).
- Nadmerná redukcia dát taktiež uberá z významu technickej analýzy a keď sa pozrieme na graf 3 týchto ukazovateľov, vidíme inverziu v rokoch na ktorých je vykonávané testovanie a teda je trocha otázne ako sa s tým model vysporiada. Možno by bolo rozumnejšie nájsť mieru podobnosti, pri ktorej by sme zlúčili metriky ktoré sa dostatočne podobajú, aby nedošlo k preučeniu, a zároveň by sme tým urobili predspracovanie dát flexibilnejším.

- Princíp CNN na zachytenie krátkodobého trendu a LSTM na prepojenie týchto krátkodobých trendov z dlhodobého hľadiska, je principiálne perfektný nápad a dalo by sa to otestovať na zachytávanie trendu namiesto samotnej predikcie presných hodnôt.
- Porovnanie štyroch modelov s cieľom určiť najvýkonnejší model je to prínosné, pretože to umožňuje izolovať model, ktorý najlepšie zovšeobecňuje rôzne akcie a časové rámce.
- Taktiež by to chcelo testovanie na iných medzinárodných trhoch, čo je spomenuté v ich poznámkach a viacej ladenia siete. Rovnako by pomohlo pridať ďalšie faktory ovplyvňujúce trh, čo taktiež spomenuli, a ako už bolo spomenuté skôr, tými by boli určite sentiment globálnych a pre daný podnik lokálnych informácii a určité výkonnostné metriky spoločnosti.
- Výsledky nie sú pretransformované spätne na pôvodné hodnoty a teda nevieme o akú odchýlku sa reálne jedná.

Očakávané ťažkosti pri replikácii výsledkov

- Hneď prvá vec ktorú sme si všimli bol chaotický popis rozdelenia trénovacích, validačných a testovacích dát, kde súčet rozdelenia mal výsledok 120%.
- Taktiež je chaotický popis vstupov a ich úprava, v ktorom kroku robíme normalizáciu, kedy štandardizujeme, počítame PC z normalizovaných hodnôt keď v grafe 2 sú v zdrojovej forme a následne na grafe 3 v štandardizovanej forme.
- Výpočty technických ukazovateľov sú veľmi zle zapísané a v mnohých prípadoch aj chybné či s nezadefinovanými hodnotami. Taktiež je asi predpoklad že čitateľ pozná dané vzorce, keďže ich popis je minimálny a museli sme si ich naštudovať z inej literatúry.
- Taktiež ak do grafu 3 pridali min-max normalizovanú záverečnú cenu, ako to že nadobúda hodnoty mimo rozsahy [0, 1].
- Po hlbšej analýze nám táto štúdia príde neaplikovateľná pre produkciu, tiež pochybujeme o daných výsledkoch, no jedná sa o zaujímavý experiment a postup.