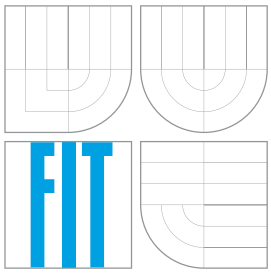
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

VYUŽITÍ VEŘEJNÝCH OBCHODNÍCH INFORMACÍ PRO AUTOMATICKÝ TRADING

USAGE OF PUBLIC BUSINESS INFORMATION FOR AUTOMATIC TRADING

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — MARTIN GRÁCA
AUTHOR

VEDOUCÍ PRÁCE — JAN ČERNOCKÝ, doc. Dr. Ing.
SUPERVISOR

BRNO 2016

Abstrakt

V dnešní době moderních technologií a výkonných počítačů již klasické obchodní modely přestávají fungovat. Pro úspěšné obchodování na burze, generující konzistentní zisky, je proto vhodné využít nových možností a technologií. Cílem této práce je právě díky těmto novým technologiím vytvořit fungující automatický obchodní systém. Tato práce využívá veřejně dostupných dat uložených v databázi Americké Komise pro cenné papíry (SEC), historické ceny akcii a rekurentní neuronové sítě k vytvoření takového modelu. Výsledný obchodní systém je schopný úspěšně obchodovat a vykazovat zisk.

Abstract

In the era of modern technology and high performance computers, the classical trades model getting insufficient. For successful trading, generating stable profit, it is good to use modern technologies and opportunities. The main goal of this work is to develop a trading system based on modern technologies. This work uses public business data from Edgar database managed by U.S. Securities and Exchange Commission (SEC), historical shares prices and recurrent neural network to create such model. The final system is able to trade successfully and generate profit.

Klíčová slova

automatické obchodování, obchodování, burza, trh, neuronová síť, rekurentní neuronová síť, predikce, fundamentální analýza, SEC filings

Keywords

automatic trading, trading, stock exchange, market, neural network, recurrent neural network, prediction, fundamental analysis, SEC filings

Citace

GRÁCA, Martin. *Využití veřejných obchodních informací pro automatický trading.* Brno, 2016. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Černocký Jan.

Využití veřejných obchodních informací pro automatický trading

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana Doc. Dr. Ing. Jana Černockého. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Martin Gráca
18. května 2016

Poděkování

Děkuji svému vedoucímu, Doc. Ing. Janu Černockému, za cenné rady, věcné připomínky, odbornou pomoc a trpělivost při vypracovávání této bakalářské práce.

Obsah

Kapitola 1

Úvod

V dřívější dobách bylo obchodování na burze výsadou pouze velkých světových bank a velkých hráčů, kteří měli dostatek financí na založení obchodního účtu. ve 21. století je situace poněkud odlišná. S rozšířením internetu mezi běžné lidi a s příchodem obchodních platforem jako je MetaTrader nebo NinjaTrader se otevřely brány i pro menší obchodníky, kterým stačí kapitál v hodnotě několika tisíc korun k založení obchodního účtu.

1.1 Motivace

Hlavní motivací této práce je využít dostupné ekonomické informace firem obchodovaných na burze k vytvoření spolehlivého obchodního systému vytvářející zhodnocení, které je mnohem zajímavější než u konzervativních, investičních respektive spořících produktů. S vyššími zisky je ale spojená vyšší míra rizika a ztráty investovaných peněz. Mnoho lidí, kteří se snaží vydělávat na burze, končí se stoprocentní ztrátou, která je zapříčiněná obchodováním na páku (vysvětlení v podkapitole 2.2.2), špatnou strategií nebo psychikou. Více o psychologii v obchodování se dočtete v kapitole 2.6.

Z toho důvodu jsou v této práci použity nejmodernější technologie, což přispívá k vytvoření ziskového a spolehlivého systému, který není ovlivněn psychikou a funguje pouze podle racionálních pravidel.

1.2 Struktura práce

V kapitole 2 se dozvíte nutné základy o obchodování, které jsou důležité pro tuto práci. Řekneme si rozdíly mezi automatickým obchodováním a obchodováním manuálním, vysvětlíme si význam psychologického aspektu a význam Dowovy teorie. Kapitola 3 popisuje teorii neuronových sítí, obzvlášť rekurentních neuronových sítí a jejich typů, způsoby normalizace sítě. Kapitola 4 obsahuje popis veřejných finančních dat, typy finančních dokumentů a standardizovaný formát XBRL. Kapitola 5 obsahuje popis dat pro učení neuronové sítě. Kapitola 6 popis algoritmu. Kapitola 7 obsahuje návrh a implementaci celého modelu a veškerých vytvořených skriptů. Kapitola 8 se věnuje testování a 9 kapitola zahrnuje shrnutí a směr budoucího vývoje a vylepšení.

Kapitola 2

Obchodování na burze

Burza je vysoce organizovaný a regulovaný trh, kde je možné prodat nebo koupit cenné papíry, komodity, měnové páry, future kontrakty nebo opce. Je součástí kapitálového trhu, což je podmnožina finančního trhu. Na burze se pohybují investoři a emitent. Eminenti jsou státní, investiční a veřejné instituce nebo podniky, které nabízejí cenné papíry k prodeji. Mezi investory se řadí banky, podílové nebo penzijní fondy, pojišťovny, ale také fyzické osoby, což je pro nás nejdůležitější.

Historie burzy sahá do 15. století, kdy byla roku 1531 založena první burza v nizozemských Antverpách. Obchodovalo se zde se směnkami, zlatými a stříbrnými mincemi. Začátek newyorské burzy je datován k roku 1817, kdy se obchodníci začali scházet na slavné Wall Street. Dnes je newyorská burza s názvem New York Stock Exchange (NYSE) největší burza na světě dle tržní hodnoty cenných papírů v amerických dolarech. Z tohoto důvodu jsou akciové tituly vybrané pro tuto práci právě z této burzy. Použité akcie se vyznačující vyšší mírou volatility a vysokými objemy obchodů. Jednou z výhod obchodování na burze je možnost vstoupit do obchodu, jak v růstovém období tak v době poklesu, což umožňuje traderům obchodovat i v době recese nebo krátkodobé korekce.

2.1 Pojmy

Zde si vysvětlíme důležité pojmy, které jsou spojené s obchodováním na burze a jsou nutné k pochopení dalších kapitol.

2.1.1 Finanční produkty a deriváty

Mezi finanční produkty se například řadí stavební spoření, penzijní připojištění nebo některé formy životního pojištění. Do burzovně obchodovatelných produktů lze investovat prostřednictvím burzy a patří sem akcie, dluhopisy, futures, opce a ETF (Burzovně obchodovatelný veřejný fond). Tyto produkty jsou vysoce standardizované a nabízí řadu výhod, především vyšší likviditu, standardizovanost a také vyšší transparentnost transakcí.

Mezi nejznámější burzovně obchodovatelné deriváty patří Futures, Opce a CFD. Deriváty jsou odvozené například od cenných papírů nebo komodit. Mají termínový charakter, to znamená, že čas vypořádání obchodu je rozdílný od času sjednání a za sjednání derivátu se neplatí nic nebo jen minimální část vzhledem k celkovému aktivu.
Futures kontrakt je smluvní ujednání, které se obecně sjednává za účelem koupě nebo prodeje určité komodity nebo finančního instrumentu za předem stanovenou cenu v budoucnosti. Některé termínové kontraktry mohou vyvolat fyzické dodání aktiva, zatímco jiné jsou

vypořádány v hotovosti.
CFD neboli Contract For Difference je smlouva o rozdílu, kdy jsou rozdíly ve vypořádání zaplaceny prostřednictvím hotovostních plateb. Tento kontrakt využívá hodně spekulativních obchodníků, neboť jim dovoluje i spekulaci na pokles aktiva, otevřít tzv. "short" pozici.

2.2 Broker

Broker, neboli obchodník s cennými papíry je licencovaná právnická osoba, která zprostředkovává přístup na burzu. Za to účtuje zákazníkovi provizi za své služby. Většina velkých firem jsou tvůrci trhu takzvaní market makeři, což znamená, že jsou prostředníci mezi kupujícími a prodávajícími. Market maker je ochotný ve stejnou chvíli nakoupit i prodat, čímž může dodávat na trh likviditu.

2.2.1 Časový rámec

Časový rámec (anglicky time frame) nám říká, za jak dlouho se vytvoří jedna svíčka nebo čára na zobrazovaném grafu. Může být založen na intradenních - sekundových, minutových, pětiminutových, desetiminutových, patnáctiminutových, půlhodinových, hodinových grafech. Dále na denních grafech, týdenních, měsíčních nebo ročních.

2.2.2 Pákový efekt

Finanční páka (anglicky leverage) znamená použití cizího kapitálu za účelem otevření většího obchodu a tím navýšení svého zisku. S tím je však spojena vyšší možnost ztráty investovaných peněz. Pokud například nakoupíme akcie s pákou 1:2, potřebujeme pouze polovinu vlastních peněz k nákupu, další polovinu nám poskytne broker. Při špatném scénáři, kdy se cena akcie propadne o 10% přicházíme o 20% svého kapitálu. U CFD kontraktů však brokeři nabízejí i páku 1:50 nebo 1:100, což výrazně zvyšuje jak zisky, tak i možné ztráty.

2.2.3 Indikátory

Indikátory jsou součástí technické analýzy a jedná se o matematický výpočet, který se aplikuje na cenu nebo objem finančního aktiva. Na základě indikátorů můžeme předpovídat budoucí změnu ceny. V dnešní době existují stovky těchto indikátorů, které jsou určeny jak pro specifickou část tržního cyklu tak i pro všeobecné použití. Mezi nejznámější patří RSI (Relative Strenght Index), Stochastic, Bolling Band, MACD (Moving Average Convergence Divergence) nebo klouzavé průměry (průměr za několik předešlých uzavíracích cen). RSI například indikuje přeprodanost nebo překoupenost trhu. Klouzavé průměry slouží k určení trendu. Indikátory jsou mezi tradery hojně využívány, nicméně neměly by sloužit jako primární ukazatel, ale spíše jen jako konečné potvrzení ke vstupu do obchodu.

2.3 Technická analýza

Technická analýza je metoda pro předpovídání cenového pohybu, ke kterému používá systematické zkoumání, analyzování a vyhodnocování minulých tržních hodnot daného aktiva. Je hojně používaná u všech finančních produktů a derivátů. K predikci používá pouze numerické informace, zejména cenu, objem, volatilitu nebo počet otevřených kontraktů.

Novodobá technická analýza se začínala rozvíjet od 19. století na základě Dowovy teorie, která vznikla z písemných poznámek Charlese Dowa a jeho společníků Williama Hamiltona a Roberta Rhea [14]. Tato teorie je považována za základ moderní technické analýzy. Z toho důvodu je Dowova teorie detailněji popsána v kapitole 2.7.

Ke grafickému znázornění se při této metodě nejčastěji používají různé typy grafů a svíčkových technik, které začali používat asijští obchodníci.

2.4 Fundamentální analýza

Fundamentální analýza na druhou stranu využívá pro předpovídání budoucího směru vývoje ceny různá ekonomická, účetní, statistická data. Dále zahrnuje politické, historické nebo demografické faktory. Největší část fundamentální analýzy tvoří účetní závěrky firem, známá také jako kvantitativní analýza. Patří sem hospodářské výsledky firem jako jsou tržby, hodnota pasiv, výdaje, zisky přepočtené na akcii a ostatní finanční aspekty společnosti.

2.5 Styly obchodování

Jednotlivé obchodníky můžeme rozdělit podle několika kritérií. Jednu skupinu můžeme rozdělit podle směru trhu, ve kterém zpravidla obchodují:

- Trendoví obchodníci
- Protitrendoví obchodníci
- Swingoví obchodníci

Trendoví obchodníci otevírají obchody zásadně ve směru trendu, a to jak v rostoucím tak v klesajícím. Trend určíme nejsnáze například pomocí klouzavých průměrů, což je aritmetický součet několika předešlých hodnot. Pokud je klouzavý průměr rostoucí, je i trend rostoucí. Jestliže je klouzavý průměr klesající, je trend taktéž klesající.

Druhou skupinu můžeme rozdělit podle časového rámce, ze kterého otevírají nejvíce obchodů:

- Skalpeři
- Intradenní obchodníci
- Krátkodobí obchodníci
- Střednědobí obchodníci
- Dlouhodobí obchodníci

Řazení je od obchodníků pracující s nejnižšími rámci (sekundy, minuty) až po obchodníky využívající rámce od měsíčních po roční. Obecně platí, že čím menší rámec, tím více příležitostí k nákupu nebo k prodeji, ale zde se objevuje nejvíce falešných signálů a velké množství obchodů je spojeno s větší částkou zaplacenou za poplatky. Na druhou stranu, čím větší rámec, tím méně příležitostí k obchodování, ale za to signály pro vstup mají mnohem větší váhu [15].

2.6 Psychologie v obchodování

Psychologie traderů je v některých případech největším kamenem úrazů. Může se stát, že i když máme dobře postavenou obchodní strategii, tak nemusíme dosahovat takových zisků jakých bychom chtěli. To je zapříčiněno nedodržováním určených pravidel, které jsou porušovány, neboť je snadné ztratit koncentraci a dělat unáhlená rozhodnutí, která v lepším případě jen sníží náš profit. V horším případě může dojít ke smazání účtu pokud otevíráme příliš velké pozice nebo po řadě úspěšných obchodů vstupujeme do pozic, které se neslučují s naší strategií. Na vině může být také špatná psychika zapříčiněná problémy v práci, s partnerem atd.. Tento aspekt obchodování u automatických obchodních systémů odpadá. Proto některým traderům můžou automatické nebo alespoň poloautomatické systémy pomoct ke zvýšení profitu na burze.

2.7 Dowova teorie

Dowova teorie se formulovala v podobě článku v časopise Wall Street Journal, které vydával Charles Henry Dow od roku 1900 do roku 1902, kdy zemřel. Vzhledem k předčasné smrti nebyla Henryho teorie kompletně dokončena, ale jeho následovníci publikovali studie, které jeho články rozšiřovaly.

Dow věřil, že akciový trh jako celek odráží celkové obchodní podmínky v rámci hospodářství, a že pomocí analýzy trhu jako celku je tyto podmínky možné zhodnotit a dle nich určit směr hlavních tržních trendů a také kam bude směřovat vývoj jednotlivých akciových titulů. Snažil se najít odpověď na otázku, kdy přesně koupit akcie, ale jeho idea se dá použít na všechny typy finančních trhů jako jsou měnové páry, komodity a jiné. Dow předpokládal, že každá akcie je součást nějakého portfolia, a růst portfolia je tedy vyvolán růstem tržní ceny každé akcie obsažené v tomto portfoliu. Z toho vyplývá, že pokud známe vývoj průměrné ceny portfolia, můžeme určit, zda se cena jednotlivých akcii zvýšila nebo snížila. Dow poprvé použil svou teorii k vytvoření indexu průmyslových odvětví - Dow Jones Industrial Index a k indexu železničního odvětví - Dow Jones Rail Index. Tyto dva indexy byly vytvořeny proto, protože Dow věřil, že když pokrývají dva hlavní ekonomické segmenty - průmysl a přepravu, budou odrážet i obchodní podmínky celého hospodářství. Dnešní indexu už obsahují jiné akciové tituly než ty původní, ale Dowova teorie je aplikovatelná stále stejně [6].

Jelikož technická analýza vychází z Dowovy teorie, uvedeme si jejich šest základních principů:

1. Trh zohledňuje všechny dostupné informace
2. Trh se neustále pohybuje v jednom ze tří směrů - primárním (nejdelší trend, který se formuje více než rok), sekundárním (střednědobý trend, reprezentuje korekci u primárního trendu) a minoritní (krátkodobý pohyb, který trvá méně než tři týdny)
3. Trendy se skládají ze tří fází - akumulace (první fáze, kdy na trh vstupují informovaní investoři), participace (ceny vstoupají vzhůru, prognózy trhu jsou optimistické) a distribuce (fáze, kdy trh dosáhl svého vrcholu a zkušení investoři začínají prodávat)
4. Tržní indexy se musejí vzájemně potvrzovat
5. Objem obchodů musí potvrzovat trend
6. Trend je platný, dokud nedojde k jasnému signálu zvratu

2.8 Proč automatický obchodní systém (AOS)?

Uveďme si zde pár bodů, které hrají ve prospěch automatickým obchodním systémům, oproti systémům manuálním.

+ AOS se drží své strategie, nedělá unáhlená rozhodnutí ovlivněná psychikou.
+ Strategie je jednoduše testovatelná na historických datech.
+ Pracuje i v době, kdy spíme.
+ Šetří čas - tradeři obchodují několik hodin denně, někteří celý den, automatický obchodní systém obchoduje sám
+ Neváhá s otevřením obchodu - někteří tradeři mají problém s otevřením obchodu a váhají tak dlouho, až prováhají nejlepší příležitost pro vstup.
  Pokud Automatický systém detekuje signál, ihned pošle obchodní příkaz.

Když jsme si uvedli klady automatického obchodního systému, v rámci objektivity si uvedeme i zápory těchto systémů.

– Složitost - obchodní systém vyžadují určitou míru technického pochopení, i v případě, že nevyvíjíme svůj vlastní systém. Je nutné pochopit při nejmenším parametry systému
– Větší vstupní investice - pokud nejsme schopni si takový systém sami vytvořit, musíme platit navíc za vývoj tohoto systému
– Delší vývoj - tedy i v případě pokud tento systém vyvíjíme sami
– Někteří můžou mít problém s důvěrou takového systému, přece jen ztráta peněz je citlivá záležitost
– V krizových situacích na trhu, jako jsou například politické události, krach firmy nebo nečekané rozhodnutí centrální banky, mohou být systémy méně výnosné. Takové riziko však hrozí i traderům a na druhou stranu složitější automatické obchodní systémy mohou mít ochranu i proti takovým situacím.

Kapitola 3

Neuronová síť

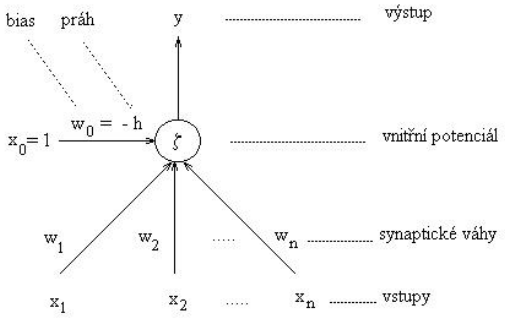
Neuronová síť je matematický model, který je inspirován neuronovým systémem živých organismů. Je specifická svými vlastnostmi, strukturou - jak jsou jednotlivé neurony propojeny, způsobem vybavování - jak probíhá výpočet při jednom průchodu, a v neposlední řadě způsobem učení.

Učení může být buď s učitelem nebo bez učitele. Při učení s učitelem, učitel hodnotí výsledek, který buď chválí nebo kritizuje. U takového učení však musíme předem znát požadovaný výstup. Samotné učení probíhá tím způsobem, že se nastaví počáteční váhy, které se poté při špatném výstupu upravují. Učení končí, pokud jsme s výsledkem spokojení nebo pokud už nemáme žádné další vzorky, tzv. trénovací data. Obecně platí, že časová náročnost učení se zvyšuje s počtem vrstev sítě.

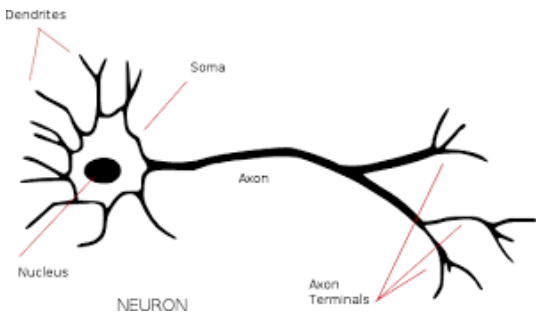
Než se dostaneme k teorii rekurentní neuronové sítě, která byla využita v této práci, vysvětlíme si nejdříve klasickou neuronovou síť, která je základem pro síť rekurentní [12].

3.1 Neuron

Základní jednotkou neuronové sítě je stejně jako u člověka neuron, respektive formální neuron. Modelové schéma můžete vidět níže na obrázku 3.1. Formální neuron se skládá z několika reálných vstupů x_1,x_2...x_n, které modelují dendrity (krátké výběžky neuronu, které přijímají vstupní informaci - nervový vzruch). Všechny vstupy jsou ohodnoceny obecně reálnými váhami w_1,w_2...w_n, které stanovují jejich propustnost.



Obrázek 3.1: Formální neuron (převzato z [7])



Obrázek 3.2: Lidský neuron

Vážená suma vstupních hodnot s váhou w_n představuje vnitřní potenciál neuronu:

$$\xi = \sum_{i=1}^{n} w_i x_i \tag{3.1}$$

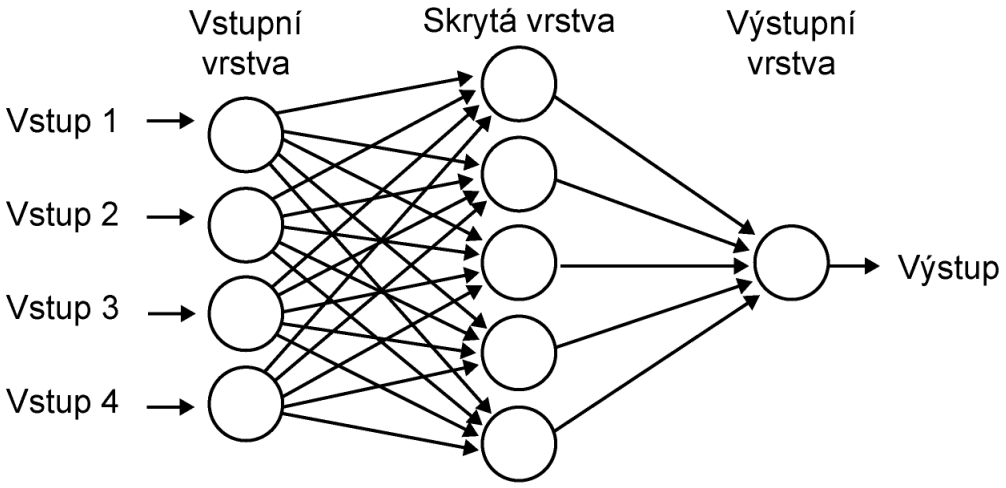
Výstup neuronu y (modelující elektrický impuls axonu) závisí na hodnotě vnitřního potenciálu ξ a hodnotě prahové hodnoty h. Výstupní hodnotu určuje aktivační neboli přenosová funkce. Nejjednodušším typem této funkce je ostrá nelinearita:

$$f(x) = \begin{cases} 1, & \xi \geq h \\ 0, & \xi < h \end{cases}$$

3.2 Vrstevnatá neuronová síť

Ve vrstevnaté neuronové síti jsou neurony rozděleny do několika vrstev. Sousední vrstvy jsou mezi sebou propojeny tím způsobem, že neuron z jedné vrstvy tvoří vazbu se všemi neurony z následující vrstvy. Neurony, které jsou ve stejné vrstvě však mezi sebou žádné vazby netvoří. První vrstva se nazývá výstupní. Poté následuje několik skrytých vrstev a poslední vrstvě se říká výstupní. Jednotlivé vrstvy mohou obsahovat různý počet neuronů.

Topologie takové sítě se nejčastěji zapisuje ve tvaru VstV-SV_1-SV_2-VysV. Například 100-80-60-42, což značí síť se 100 vstupními neurony ve vstupní vrstvě $VstV$, 42 výstupními neurony ve výstupní vrstvě $VysV$ a s 80 a 60 neurony, které jsou ve skrytých vrstvách SV_1,SV_2.



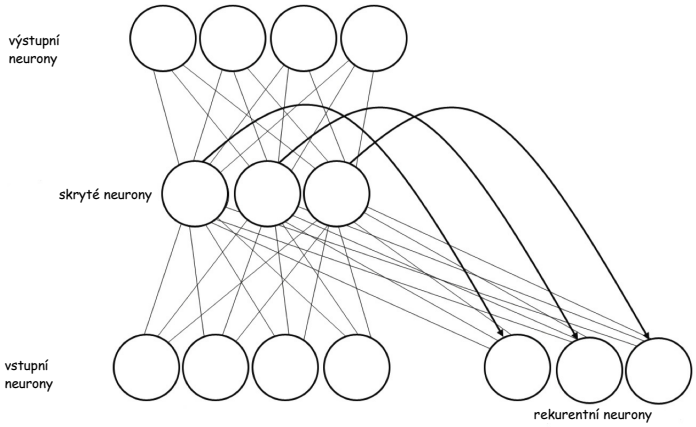
Obrázek 3.3: Vrstevnatá neuronová síť (Převzato z [11])

Algoritmus vybavování funguje tak, že vstupní vrstva předává signály do první skryté vrstvy. Zde všechny neurony spočítají svůj výstup, který zase pošlou do další vrstvy jako vstup pro tuto vrstvu. Tak se pokračuje, dokud nedosáhneme konečného výstupu.

3.3 Rekurentní neuronová síť

Rekurentní vícevrstvá neuronová síť má oproti síti, která byla popsána v předešlé kapitole 3.2, několik vlastností navíc, liší se také její struktura a jsou mnohem více podobné lidskému mozku. V této sítí je zahrnut i časový kontext, jinými slovy tento model se dokáže dívat do minulosti a do výpočtu výstupní hodnoty zahrnuje i předešlé výpočty, které jsou uloženy ve speciálním rekurentním neuronu - má tedy možnost uchování informace. To je umožněno díky zpětnovazebnému přenosu informace z vyšší vrstvy do vrstev nižších. Například z první skryté vrstvy vedou spoje i do vstupní vrstvy směrem k rekurentnímu neuronu. Počet rekurentních neuronů je shodný s počtem neuronů, které jsou o jednu vrstvu výše. Příklad takovéto sítě můžete vidět níže na obrázku 3.4.

Nevýhodou rekurentních neuronových sítí je však obtížnost trénování a větší paměťové požadavky.



Obrázek 3.4: Rekurentní vrstevnatá neuronová síť (Převzato z [10])

Rekurentní neuronové sítě se začaly rozvíjet v 80. letech 20. století a od té doby vzniklo několik modifikací, které jsou optimalizované pro specifické využití. Pro nás jsou důležité zejména bránové rekurentní neuronové sítě LSTM (Long-Short-Term-Memory), popsána v kapitole 3.4 a GRU (Gated Recurrent Units), popsána v kapitole 3.5. Tyto sítě jsou mimo jiné vhodné pro předpovídání časových řad a zlepšují tak přesnost výsledků.

3.4 LSTM

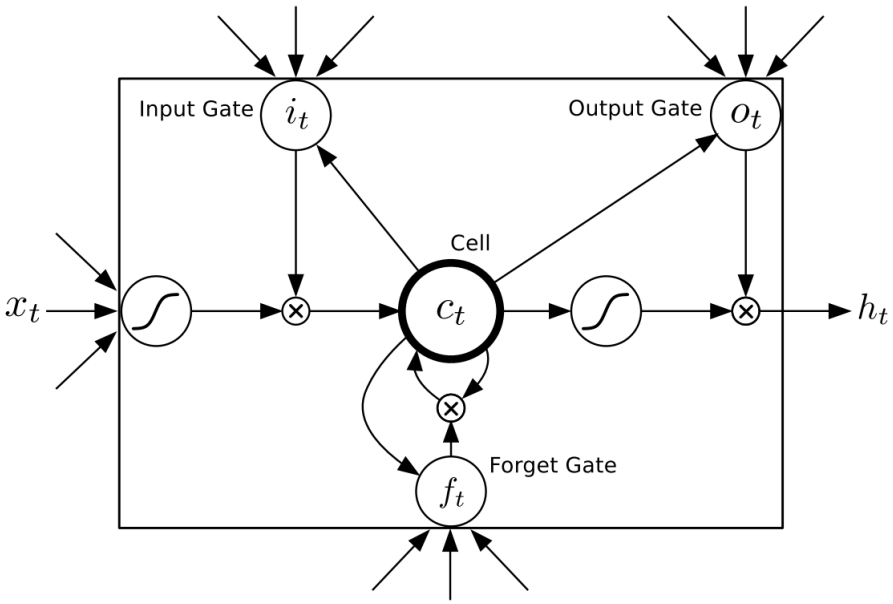
Tato architektura byla poprvé použita v článku z roku 1997 od Hochreitera a Schmidhubera viz [5]. LSTM řeší takzvaný *vanishing gradient problem* , který se vyskytuje u učení umělých neuronových sítí, zejména u gradientně založených metod a metody backpropagation (využívána u učení dopředných neuronových sítí s učitelem). Tento problém způsobuje velmi pomalé učení přední vrstvy se zmenšující se hodnotou gradientu, který může být pouze v rozsahu [0,1].

Výhodou této sítě je, že vzpomínky mohou být uloženy až po nekonečně dlouhou dobu, zatímco u klasických rekurentních sítí složené pouze ze sigmoidních funkcí může dojít k rychlé ztrátě stavu nebo paměti. U sítí RTRL (Real Time Recurrent Learning) a BPTT

(Backpropagation through time - Algoritmus zpětného šíření chyby v čase) nemůže být informace uchována po delší dobu než 10 až 12 časových kroků.

Níže na obrázku 3.5 můžete vidět schéma jedné paměťové buňky, která se skládá z pěti prvků. Vstupní (input) a výstupní brány (output gate), paměťového bloku (memory block), zapomínací brány (forget gate) a skrytého výstupního stavu (hidden state output). Brány můžeme jednoduše brát jako spínač, který může být buď sepnutý nebo rozepnutý.

Paměťový blok se může skládat z jedné nebo více paměťových buněk. V jádru buňky se nachází lineární jednotka s jednoduchou vlastní rekurentní vazbou nastavenou implicitně na počáteční hodnotu 1. Jelikož zde není žádný jiný vstup, tato buňka si dokáže zachovat aktuální stav z jednoho momentu na další. Jednotlivé buňky dostávají vstup ze vstupních jednotek, jiných buněk a bran. Zatímco buňky jsou zodpovědné za udržování informací přes dlouhá časová období, odpovědnost za rozhodování o tom, jaké informace ukládat a kdy se tyto informace použijí mají vstupní a výstupní bránové jednotky.



Obrázek 3.5: Paměťová buňka LSTM (převzato z: [9])

Vstup do buňky prochází skrz nelineární funkci *g(x)*, typicky sigmoidní funkce, která leží v rozmezí [-2,2] a výsledek je násoben s výstupem vstupní brány. Aktivace brány leží v rozsahu [0,1], to znamená pokud je její aktivace blízko nuly, nic do buňky nevstoupí. Jenom pokud je vstupní brána dostatečně aktivována, je signál vpuštěn dovnitř. Podobně nic nevystoupí z buňky dokud není výstupní brána aktivní neboli "sepnutá". Jelikož je vnitřní stav buňky uchováván v lineární jednotce, její aktivační rozsah je neomezený a proto výstup buňky je znova prohnán přes aktivační funkci *h(x)* (typicky sigmoidní funkci v rozmezí [-1,1]).

Brány samotné nejsou nic jiného než konvenční jednotky používající sigmoidní funkci v rozmezí [0,1] a každá z nich dostává na vstup vstupy sítě (vzorky) a výstupy jiných buněk. Rekurentní spoj vnitřní buňky je řízen třetí zapomínací bránou (forget gate). To znamená, že pokud je tato brána zavřená, nedojde k uchování kontextu.

Pro výpočet dostáváme tyto rovnice:

Výstup brány $y^{c_j}(t)$ je

$$y^{c_j}(t) \quad = \quad y^{out_j}(t)h(s_{c_j}(t)) \tag{3.2}$$

kde $y^{out_j}(t)$ je aktivací výstupní brány stav $s_{c_j}(t)$ a je dán vztahem

$$s_{c_j}(0) \quad = \quad 0 \tag{3.3}$$

$$s_{c_j}(0) \quad = \quad s_{c_j}(t-1) + y^{in_j}(t)g(net_{c_j}(t)) \tag{3.4}$$

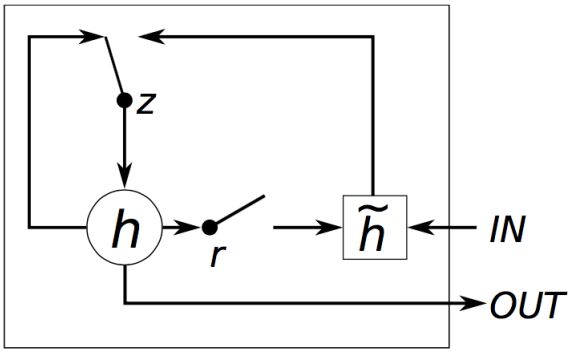
pro $t > 0$

Aplikaci této sítě najdeme například v robotice, rozpoznávání řeči, skládání rýmu, skládání písní, učení gramatiky, rozpoznávání ručně psaného textu nebo, pro nás nejdůležitější, v predikci časových řad.

3.5 GRU

GRU neboli Gated Recurrent Unit je obdobou LSTM sítě popsanou v předešlé kapitole 3.4. Jedná se o poměrně mladý a plně neprozkoumaný model. První zmínka je z roku 2014 v práci Junyonga Chunga [**?**]. GRU má pouze dvě brány, resetovací bránu (reset gate) a aktualizovací bránu z (update gate), kdežto LSTM měla 3 brány. Resetovací brána určuje, jak kombinovat nový vstup s předchozí pamětí a aktualizovací brána určuje, jak moc předchozí paměť zachováváme.

Pokud tedy nastavíme všechny resetovací brány na 1 (brána bude otevřená) a všechny aktualizovací brány na 0, dostaneme model klasické rekurentní neuronové sítě.



Obrázek 3.6: Paměťová buňka GRU (Převzato z [8])

Základní myšlenka a použití bran je stejné jako u LSTM. Vyskytuje se zde však několik hlavních odlišností:

- GRU obsahuje místo tří bran pouze dvě
- neuchovává vnitřní paměť c_t, což je rozdíl odkrytého oproti skrytému stavu. Není zde totiž výstupní brána, která je u LSTM
- vstupní a zapomínací brána je sjednocena do jedné aktualizovací brány z a resetovací brána je umístěna přímo za předchozí skrytý stav.
- u GRU chybí výpočet druhé nelinearity u výstupu

Teď, když jsme si vysvětlili dvě podobné rekurentní sítě, nabízí se otázka, která z nich je lepší. To obecně nemůžeme říci. V mnoha problémech se zdají být oba modely rovnocennými soupeři. Vzhledem k méně parametrům u GRU, se však tato síť o něco málo rychleji učí a potřebuje tak ke generalizaci méně dat. Na druhou stranu, pokud máme vstupních dat dostatek, může být vhodnější LSTM.

3.6 Metriky

Neméně důležité je vyhodnocení kvality vytvořeného modelu a přesnosti predikce. Tu můžeme měřit například predikční chybou, která vypočítává míru nepřenosti mezi predikovaným a originálním výstupem. Uveďme například střední kvadratickou chybu MSE (Mean Squared Error), která se vypočítává z n predikovaných hodnot $\hat{y}_i$ a reálných hodnot y_i. Její vzorec je definován takto:

$$MSE = \frac{\sum_{i=1}^{n}(\hat{y}_i(t) - y_i(t))^2}{n} \tag{3.5}$$

Obdobou je RMSE neboli střední kvadratická chyba (z anglického root mean square error). Někdy též RMSD. RMSD je dobrým měřítkem přesnosti, ale pouze k porovnání prognostické chyby různých modelů pro konkrétní proměnné, ale nikoli mezi různými proměnnými, neboť je závislá na měřítku.

RMSE predikovaných hodnot $\hat{y}_t$ pro čas t

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i(t) - y_i(t))^2}{n}} \tag{3.6}$$

Kapitola 4

Finanční data

Jak napovídá název této práce, cílem je využít veřejných obchodních informací pro predikci akciového vývoje a následně vytvořit obchodní systém, který bude tyto informace využívat. Mezi finanční data patří také různé statistické údaje týkající se ekonomiky firem.

Analýzou těchto dat se zabývá finanční analýza, někdy též bilanční analýza. Cílem této analýzy je odhalit silné a slabé stránky firmy, zjistit ekonomickou situaci ve firmě a její zdraví. Finanční analýza se dělí do dvou skupin, interní a externí. Interní analýza je prováděna ekonomy uvnitř firmy, kteří mají k dispozici veškeré účetní a ekonomické informace, různé dokumenty, finanční plány a jiné statistické údaje. Tato práce se však zaměřuje na externí finanční analýzu, která bere v úvahu pouze dostupné, veřejně známe informace. Zaměřuje se pouze na velké americké korporace, které mají ze zákona nutnost zveřejňovat údaje o své ekonomické situaci. Jednou z možností jak tyto data využít je přímo z amerického SECu, který tyto data shromažďuje.

4.1 SEC

SEC (U. S. Securities and Exchange Commission) je americká komise pro cenné papíry a burzu, vytvořená kongresem. Někdy též označována jako "hlídací pes Wall Streetu" (Watchdog of the Wall Street). Vznikla v roce 1934 po velké finanční krizi v roce 1929. Její hlavní činností je regulovat americký finanční trh. Jak SEC uvádí, jejím hlavním posláním je ochrana investorů; zajišťovat férový, řádný a efektivní trh a podporovat tvorbu kapitálu. Snaží se o tržní prostředí, které je hodné důvěry veřejnosti. Přispívá k lepší regulaci a ochraně. Monitoruje také proces, kdy jedna firma kupuje jinou. SEC se skládá z pěti komisařů, které volí prezident spojených států a schvalují senátoři.

Pokud například nějaký subjekt koupí více než 5% kapitálu nějaké společnosti, musí do deseti dnů od koupi obeznámit SEC, aby se zamezilo nelegálnímu převzetí.

Celá komise je rozdělena do pěti divizí. Mezi ně patří: *Divize korporátních financí* (Division of Corporation Finance), *Divize obchodování a trhu* (Divison of Trading and Markets), *Divize investičního managementu* (Division of Investment Management), *Divize vymáhání* (Divison of Enforcement) a *Divize ekonomické analýzy a analýzy rizik* (Division of Economic and Risk analysis).

Pro nás je nejdůležitější *Divize korporátních financí*, která dohlíží na zpřístupnění důležitých firemních informací investorské veřejnosti. Také zkoumá veškeré dokumenty, které jsou společnosti povinné podat této komisi. Zde patří sdělení o registraci, čtvrtletní a roční

hlášení nebo výroční zprávy. Monitoruje účetní činnosti, které vedou k formulaci obecně uznávaných účetních zásad GAAP (generally accepted accounting principles). Tyto zásady pak používají firmy k sestavení své účetní závěrky.

4.2 EDGAR

Nezkráceně, the Electronic Data Gathering, Analysis, and Retrieval system. Tedy Elektronický systém pro získávání, analýzu a sběr dat. EDGAR provádí automatizovaný sběr, validaci, indexování, přijímání a předávání firemních dokumentů, které mají tyto firmy povinně dle zákona dodávat SECu. Snaží se zvýšit účinnost a spravedlnost trhu s cennými papíry a přinést tak lepší podmínky pro investory, korporace a ekonomiku tím, že urychlí příjem dokumentů, šíření a analýzu firemních informací.

4.3 SEC Filings

SEC filings zahrnuje různé typy souborů, které dávají náhled do historie společnosti, pokroku a malý náhled do budoucnosti. Patří sem sdělení o registraci, formální a pravidelné zprávy a další formuláře, které jsou poskytovány SECu. V této práci byly analyzovány pouze periodické zprávy, ze kterých byly získávány potřebná finanční data. Patří sem:

10-K

Je komplexní souhrnná zpráva o výkonnosti společnosti, které musí být předloženy každoročně SECu. Zahrnuje historii společnosti, organizační strukturu, vlastní kapitál, hospodáření, zisky na akcii, dceřinné společnosti a jiné.

Podstatný fakt je, že musí být podána do 60 dnů po skončení fiskálního roku.

10-Q

10-Q je zkrácená verze 10-K formuláře. Podobně jako předchozí formulář, obsahuje komplexní zprávu o výkonnosti společnosti, které musí být předloženy čtvrtletně všemi věřejně obchodovatelnými společnostmi SECu. Firmy jsou povinné zveřejňovat relevantní situace o jejich finanční situaci a informace by měli být poskytnuty všem zainteresovaným stranám.

Čtvrtletní zpráva musí být podána do 45 dnů po skončení všech tří fiskálních čtvrtletí. 10-Q zpráva za čtvrté čtvrtletí se nevydává, neboť je součástí roční zprávy.

8-K

Většina hlavních informací je uvedena v předešlých dvou dokumentech. V tomto formuláři se však uvádí neplánované konkrétní události a uvádí další podrobnosti, například datové tabulky a tiskové zprávy. Korporace zde, mimo jiných dalších významných událostí, zveřejňují dokončení akvize, nakládání s majetkem nebo jmenování nových vedoucích pracovníků.

Existuje ještě mnoho dalších formulářů, například Proxy statement (obsahující platy manažerů), Schedule 13D (odhaluje držitele nejvíce akcií), Form 144, Foreign Investments. Ty však nejsou nejlepší volbou, protože obsahují málo numerických dat, které jsou pro nás klíčové.

4.4 Finanční ukazatele

Do jedné části finační analýzy spadají také poměrové finanční ukazatelé. Patří sem ukazatelé likvidity (schopnost podniku splácet), ukazatele rentability, ukazatele obratu, ukazatele aktivity (schopnost podniku rychle využívat svého majetku) a ukazatele zadluženosti a finanční struktury. Mezi ukazatelé likvidity patří:

- Běžná likvidita = krátkodobá aktiva / krátkodobé dluhy
- Pohotová likvidita = (krátkodobá aktiva - zásoby) / krátkodobé dluhy
- Okamžitá likvidita = peněžní prostředky / okamžitě splatné dluhy
- krátkodobé dluhy ku jmění = okamžitě splatné dluhy / vlastní kapitál

Mezi ukazatelé obratu patří:

- hrubá zisková marže = hrubý zisk / čisté tržby
- obrat čistého jmění = čisté prodeje / vlastní kapitál
- obrat aktiv = čisté tržby / celková aktiva
- současný obrat aktiv = (čisté tržby - zisky) / současná aktiva

a mezi ukazatelé zadluženosti patří:

- dluh k vlastnímu kapitálu = celkové dluhy / vlastní kapitál
- dluhy k celkovým aktivům = celkové dluhy / celková aktiva

Posledním a velmi důležitým ukazatelem je ROI, někdy též ROI index. Zkratka pochází z anglického *Return On Investment*, což znamená návratnost investice. Jedná se o základní ukazatel v podnikové ekonomice, marketingu, prodeji a investování. ROI udává kolik jednotek finančních prostředků vydělala jedna utracená jednotka.

výpočet je následující:

$$ROI = \frac{z - investice}{investice} \tag{4.1}$$

kde z je čistý zisk

Tímto dostaneme výsledek v absolutní hodnotě. Pokud bychom chtěli znám procentuální hodnotu, je třeba upravit vzorec na

$$ROI(\%) = \frac{z}{investice} \times 100 \tag{4.2}$$

4.5 XBRL

XBRL (eXtensible Business Reporting Language) je obchodní sdělovací standard. Přesněji se jedná o otevřený, mezinárodní standard pro digitální obchodní výkazy. Spravován je globální neziskovou společnosti XBRL International. XBRL se používá ve více než 50 zemích světa. Ročně se vytvoří miliony těchto dokumentů.

Vychází ze značkovacího jazyka XML. Vznikl v roce 2003 jako standard pro odevzdávání finančních dat při pravidelných kvartálních výsledcích, které jsou americké firmy obchodovatelné na burze povinné zveřejňovat.

Pro všechny typy obchodních informací umožňují přesné datové značky:

- přípravu
- validaci
- publikaci
- změnu
- používání
- a analýzu

Kapitola 5

Data

V této kapitole je popsáno, jaká data byla použita pro trénování rekurentní neuronové sítě, tedy jako její vstupy. Zaměříme se na datovou sadu akciových titulů a na finanční data, které byly získány a které pak byly dále použitý pro výpočet finančních ukazatelů ekonomického vývoje firem.

5.1 Technická data

Jako technická data byla zvolena sada sto amerických akcií s denním časovým rámce od 1. 1. 1980 až do 11. 1. 2016. Jedná se o ceny 100 vybraných firemních akcií obchodovatelných na burze NYSE. Data jsou stažena z yahoo finance. Použita byla cena *adjusted close*, což je upravená zavírací cena dne, která zahrnuje všechny změny a firemní akce, které byly provedeny před otevřením trhu následující den a často se používá při historické analýze dat.

Denní časový rámec nám poskytuje dostatek dat pro trénování sítě a jak bylo popsáno v kapitole 2.5, vyhneme se tak většímu množství falešných signálů, které jsou způsobené dočasnou volatilitou a je tedy větší pravděpodobnost, že se síť naučí lépe rozpoznávat správné vzory. Větší množství akcii je vhodné použít proto, aby se síť naučila rozpoznávat kontext mezi jednotlivými akciemi, které se navzájem mohou ovlivňovat. Vybrané akcie nám také poskytnou dostatečnou volatilitu, která pomáhá k lepšímu trénování sítě.

5.2 Fundamentální data

Fundamentální data, která byla použita pro trénování neuronové sítě vychází z kapitoly 4.3 a 4.4. Pro každou akcii bylo použito 9 fundamentálních dat - 8 poměrových ukazatelů a index návratnosti investice. Více v kapitole.

Rozdělení dat

Stejně jako použití technických dat, rozdělení dat na trénovací, validační a testovací byla zvolena ve stejném poměru jakém používá kolega Petr Huf, zabývající se ve své práci také algoritmickým obchodováním. To znamená, že 64% dat bylo použito na trénování sítě, 16% na validaci a a zbylých 20% na testování.

Kapitola 6

Popis algoritmu

Algoritmus můžeme rozdělit do tří částí. Připravu dat pro neuronovou síť; vytvoření modelu neuronové sítě, natrénování sítě na připravených datech a na vytvoření automatického obchodního systému.

6.1 Příprava dat

V první části bylo potřeba připravit veškerá data pro neuronovou síť. To zahrnovalo stažení všech finančních souborů z amerického SECu. Vzhledem k tomu, že se XBRL formát začal používat poměrně nedávno (až od roku 2005) a pro neuronovou síť byla použita historická data akcií už od roku 1980, bylo nutné dvojí zpracování finančních dat. Zpracování finančních dat od roku 1980 až do cca roku 2005, kdy jsou data méně strukturovaná a uložena ve formátu HTML. Postupem času se také měnil formát i těchto souborů. Dříve byly tyto soubory méně obsáhlé, později začali přibývat například tabulky a velikost jednoho textového souboru se tak dostala k několika Megabytům. Od roku 2005 jsou SEC data dostupná ve formátu XBRL. To mnohem usnadnilo extrakci dat.

Po stažení všech finančních dat z databáze SECu bylo potřeba data zformátovat do jednoho souboru a to tím způsobem, aby odpovídaly technickým datům. Vzhledem k tomu, že finančních dat je poměrně o dost méně než dat technických, byla prázdná časová místa ze začátku vyplněna nulami. Později v rámci normalizace byly nuly odstraněny a finanční data byla lineárně roztažena. To znamená, že aktuální hodnota byla nakopírována i do pozdějších dat, ve kterých nemáme žádný údaj ze SECu. Nakonec tak vznikly 2 matice se stejným počtem řádků.

6.2 RNN

Pro predikci akciových kurzů byla vybrána rekurentní neuronová síť, přesněji bránová síť GRU (popsána v kapitole 3.5). Navržená síť se skládá ze vstupní vrstvy, z jedné skryté vrstvy a z vrstvy výstupní.

Vstupem sítě pro jednu akcii je vektor 10 hodnot, skládající se z historické ceny akcie a fundamentálních dat. V rámci normalizace byla vstupní data transformována do intervalu [0,1].

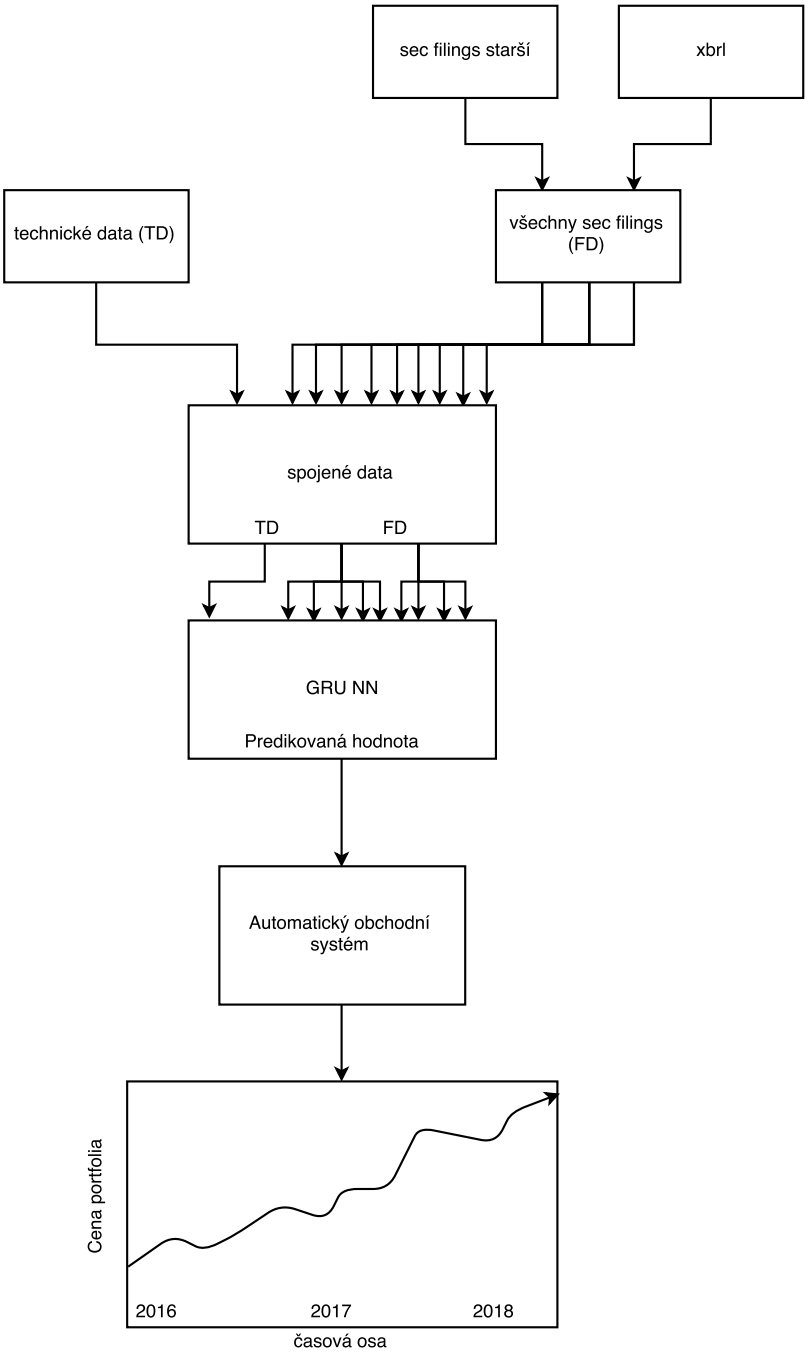
Výstupem je pouze jedna predikovaná hodnota pro následující den, která je predikovaná z predešlých denních hodnot.

6.3 Obchodní strategie

Výstup z neuronové sítě bylo třeba dále zpracovat a zakonponovat ho do obchodní strategie. Jako vzor byla použita jednoduchá strategie, která využívá klouzavých průměru (MA) 2.2.3 jako signálu pro otevření nebo uzavření obchodu. Nejprve dojde k vypočítání padesáti (MA50) a dvě stě denního (MA200) klouzavého průměru a pokud se dostane MA50 nad MA200 jsou nakoupeny akcie o objemu 2% z celkové hodnoty portfolia. V opačném případě dojde k prodeji všech nakoupených akcií.

Po přidání neuronové sítě do obchodní strategie dojde nejprve k predikci hodnoty akcie pro následující den. Nakoupení akcií dojde v případě, pokud predikovaná hodnota je větší než krátkodobý klouzavý průměr MA50. K prodeji akcií dojde tehdy, pokud je predikovaná hodnota menší než dlouhodobý klouzavý průměr MA200. Výnosnost a výsledky této strategie pro různé typy neuronových síťí jsou uvedeny v kapitole 8.

Do obchodů byly také zahrnuty poplatky za koupi akcií. Ty byly stanoveny na 0.5% z celkového nakoupeného množství. Pro akcie v hodnotě $1000 tedy zaplatíme $5 za poplatky.



Obrázek 6.1: Blokové schéma obchodního systému)

Kapitola 7

Implementace

V této kapitole je popsáno vše, co se týká samotné implementace. Jaký software byl použit pro zpracování dat, tvorbu neuronové sítě a tvorbu automatického obchodního systém.

7.1 Použitý software

Veškeré skripty a programy jsou napsané v jazyce Python. Ten poskytuje potřebnou funkcionalitu pro práci s daty, tvorbu neuronové sítě i testování obchodní strategie.

7.1.1 knihovny pro stažení a práci s daty

Pro stažení SEC filings byly využity 2 externí knihovny. *SECEdgar* [13]. Tato knihovna stáhne všechny dostupné 8-K,10-K a 10-Q soubory od zadaného data pro konkrétní firmu. Druhou knihovnou je *sec-xbrl* [1], tato knihovna stáhne všechny dostupné soubory ve formátu xbrl od - do zadaného data.

Další knihovna, která byla použita se jmenuje *xbrl* [2]. Ta dokáže rozparsovat soubor ve formátu XBRL a vrátit jako výsledek přes 40 ekonomických dat.

7.1.2 Keras

Keras [3] je minimalistická, modulární knihovna pro tvorbu a práci s neuronovými sítěmi. Jedná se o nadstavbu nad knihovnou Theano, která slouží k numerickým výpočtům. Keras byl vyvinut s cílem umožnit rychlé experimentování a minimalizovat tak čas mezi nápadem a výsledkem.

Podporuje jak konvoluční, tak rekurentní neuronové sítě, pro nás je však klíčová síť rekurentní. Důležité je, že podporuje libovolné schéma připojení, včetně mnoho-vstupního a mnoho-výstupního zapojení. Samotný výpočet pak může běžet jak na CPU, tak na GPU.

7.1.3 Quantopian - Zipline

Softwarů, poskytujících zpětné testování (testování strategie na historických datech) a tvorbu automatického obchodního systému existuje několik. Hojně používané jsou platformy NinjaTrader nebo Metatrader4. Ty jsou však napsány v jiných jazycích než v Pythonu. Propojení neuronové sítě s obchodní strategii by tak bylo poměrně složité. Z toho důvodu tak byla použita knihovna *zipline* [4] napsaná v pythonu od společnosti Quantopian.

7.2 Skripty pro přípravu dat

skript pro nexbrl data - *download_sec_edgar_data.py*

Tento skript slouží ke stažení 10-Q,10-K a 8-K souborů z databáze EDGARu uložené v souboru *txt*. Stahují se pouze data k firmám, které jsou uložené v *BP_shares_8.txt* , který obsahuje 4 sloupce, které obsahují jméno akciového titulu, CIK číslo (identifikační číslo firmy), datum, do kterého mají být data stažena, nejvyšší počet stažených souborů. K připojení do databáze EDGAR slouží knihovna crawler, která také obsahuje samostatné API pro jednotlivé typy SEC filings. Jsou jimi metody *seccrawler.filing_10Q*, *seccrawler.filing_10K* a *seccrawler.filing_8K*. Stažené soubory se uloží do složky SEC-Edgar-Data.

loadSECfilings.py

Tento skript nebyl napsán, ale je dostupný z xxx a slouží ke stažení všech dostupných xbrl souborů z databáze EDGARu. Vytvoří složku *sec*, která obsahuje složku pro každý rok a v ní složku pro každý měsíc, v které jsou všechny stažené xbrl soubory v zip formátu. Skript se spouští s parametrem *-f od_roku* a parametrem *-t do_roku*.
příkladem spuštění může být:

python3 loadSECfilings.py -f 2005 -t 2016

V tomto případě se stáhnou všechny dostupné XBRL soubory. Stáhne více než 130 tisíc dokumentů o celkové velikosti více než 20 GB, což zabere několik hodin, než dojde ke stažení všech dat. Stahování se doporučuje spouštět nejlépe přes noc newyorského času, kdy nejsou servery tolik vytížené.

Pokud bychom chtěli stáhnout data pouze pro konkrétní měsíc, spustíme skript s parametrem *-y rok* a *-m mesic*.
Například:

python3 loadSECfilings.py -y 20010 -m 9

extract_xbrl_filings.py

Jelikož jsou všechny stažené XBRL soubory zkomprimované, byl napsán tento skript, který projde všechny podsložky ve složce *sec* a vyhledá pro akciové tituly uložené v souboru *BP_shares_8.txt* všechny příslušné soubory, které odzipuje pomocí knihovny *zipline* a uloží do složky *sec_xbrls* a do podsložky pro daný titul.

parse_sec_filings.py

Tento skript prochází postupně všechny podsložky *SEC-Edgar-Data*, které byly stažené pomocí *download_sec_edgar_data.py*. Každý název souboru je pak předán funkci *process_sec_file*, ve které dojde k rozdělení obsahu souboru na část s hlavičkou a na textovou část za pomocí funkce *unpack_pem*. Ta rozděluje SEC filings podle řetězce "———BEGIN PRIVACY-ENHANCED MESSAGE———"a "———END PRIVACY-ENHANCED MESSAGE", které oddělují hlavičku od těla. Pokud se zde tyto řetězce nenachází, jedná se o nevalidní soubor, který není zpracováván.

I když mají všechny procházené soubory přípony *.txt*, uvnitř obsahují HTML dokument. Z toho důvodu byla použita knihovna lxml, umožňující parsování HTML. Konkrétně byla

použita metoda *html.fromstring*, která umí zpracovat soubor z řetězce. Nejdříve bylo třeba získat datum, kdy byl dokument zveřejněný. K tomu slouží funkce *find_sec_date*, která přijímá jako parametr obsah jednoho ze tří tagů, ve kterém se může datum vyskytovat. Jedná se o tagy <acceptance-datetime>, <sec header> nebo <ims header>. Funkce vrací datum v potřebném formátu, který má tvar DDDD.MMMM.YYYY 0:0. Například 31.6.2000 0:0. Tento formát je shodný s datem u technických dat a je potřebný k následnému správnému spojení technických a fundamentálních dat.

Všechna ekonomická data jsou ukládána do slovníku se jménem *data*, kde jako klíče slouží názvy jednotlivých ekonomických dat. Například liabilities, stockholders_equity, total_equity, gross_profit a jiné.

Po získání všech dat je do pole stats_array uloženo nejprve získané datum a následně ekonomická data s čárkou jako oddělovačem na začátku, což umožní vytvoření konečného souboru. Toto pole je vráceno jako výsledek funkce *process_sec_file*.

V hlavním cyklu, kde se prochází všechny soubory, se toto pole pro jeden soubor ukládá do pole *all_stats_array*, obsahující informace o všech finančních datech. Po skončení cyklu dochází k seřazení pole *all_stats_array* podle dat.

Výsledek pro jeden akciový titul je uložen do souboru se jménem jmeno_stats.csv a do složky *stats_files*, například DD_stats.csv

parse_xbrl_filings.py

Tento skript, stejně jako předchozí, prochází všechny podsložky sec_xbrls a využívá knihovnu *xbrl* k rozparsování jednotlivých souborů. O parsování se stará funkce *parse_xbrl*, která přijímá 2 parametry. Soubor obsahující xbrl dokument a jako druhý parametr datum vydání tohoto dokumentu. Datum je získán z názvu procházeného souboru.

Vyextrahovaná data jsou uložena do csv souboru, například *dd_xbrl_stats.csv* , který obsahuje zformátované datum, kdy byla finanční data zveřejněna a dále 49 sloupců ekonomických dat. Výsledek je uložen v adresáři *xbrl_stats_files*

merge_sec_filings.py

Slouží ke spojení všech získaných dat z databáze EDGAR, které jsou uloženy ve složkách *SEC-Edgar-Data* a *xbrl_stats_files*. Nejprve je získán seznam všech historických dat *al_dates* ze souboru FullData_dot.csv, obsahující technická data. Poté se prochází pole all_dates a při každém kroku je volána funkce find_sec_stats s aktuálním datem jako parametr. Tato funkce hledá záznam pro daný datum v obou složkách Pokud ho nalezne vrátí celý řádek souboru a ten uloží do výstupního souboru. V případě, že žádná záznam nebyl nalezen, jsou do výstupního souboru uloženy pouze nuly se středníky jako oddělovači.

count_ratios_and_merge_data.py

Při spuštění tohoto skriptu dojde ke správnému spojení technických dat s fundamentálními daty a jsou spočítány všechny poměrové ukazatele a ROI index. Výpočet probíhá na základě vzorců z kapitoly 4.4.

7.3 tvorba modelu a trénování neuronové sítě

Veškerá implementace se nachází v souboru *rnn_with_sec.py*. Ten obsahuje 4 nejdůležitější funkce. *_load_data* , *_train_test_split*, *_create_and_learn_rnn_model* a *save*. Ve funkci

_load_data probíhá rozdělení dat (načtené například ze souboru *DD_stats_and_ratios.csv*) na vstupní a výstupní hodnoty neuronové sítě. Cyklus prochází postupně data od nejstarších až po nejnovější a rozděluje je vždy na matici o 10 sloupcích a 100 řádcích jako vstup pro neuronovou síť (matice uloženy v proměnné *docX*) a pouze na jedno prvkovou matici, která obsahuje predikovanou cenu akcie z 200 předešlých hodnot, jako výstup sítě (matice uloženy v proměnné *docY*).

Ukázka kódu z funkce _load_data pro načtení dat:

```
docX, docY = [],[]
steps = 100
for i in range(0,int(data.shape[0]-steps)):
    docX.append(data[i:(i+steps),:])
    docY.append(data[i+steps,:1])
```

Ve funkci *_train_test_split* dochází pouze k rozdělení vstupních a výstupních dat sítě na trénovací a testovací podle x-ové osy matice (axis 0). proměnná *ntrn* obsahuje počet vzorků pro trénování.

Ukázka kódu funkce _train_test_split:

```
X, Y = \_load\_data(data)
ntrn = round(X.shape[0] * (1 - test_size))
perms = np.random.permutation(X.shape[0])
X_train = X.take(perms[0:ntrn], axis=0)
Y_train = Y.take(perms[0:ntrn], axis=0)
X_test = X.take(perms[ntrn:], axis=0)
Y_test = Y.take(perms[ntrn:], axis=0)
```

Ve funkci *_create_and_learn_rnn_model* probíhá tvorba modelu rnn, její překlad a trénování. Jako první je provedena normalizace všech dat na rozsah [0,1], čeho je dosažené pomocí metody *MaxAbsScaler()* ze třídy preprocessing, dostupné z knihovny sklearn. Tato normalizace je vhodná pro data, kde se vyskytuje hodně nulových hodnot, což v našem případě u fundamentálních dat platí. Poté se volá funkce *train_test_split* popsána výše, která uloží výsledek do proměnných x_train, y_train, X_test a y_test.

Následně je vytvořen model neuronové sítě. počty neuronů jsou uloženy v proměnných in_neurons, hidden_neurons a out_neurons. Objekt modelu je vytvořen z třídy *Sequential*. Jednotlivé vrstvy a aktivační funkce jsou pak přidávány pomocí metody *add*. Překlad probíhá po volání metody *compile*.

Uložení modelu a váh

Aby nebylo nutné vytvářet model a trénovat ho pokaždé pro případy testování obchodní strategie, jsou model i natrénované váhy modelu uloženy. Model je uložen ve formátu json do souboru rnn_sec_model_architecture.json, což umožňuje metoda to_json() třídy *Sequential*, která po zavolání vrací json v podobě řetězce. K uložení váh slouží metoda *save_weights* z téže knihovny. Přijímá název h5 souboru, do kterého se váhy mají uložit. V našem případě do *rnn_sec_model_weights.h5*.

Výsledné schéma modelu rekurentní neuronové sítě se nachází níže na obrázku 7.1.

Výpočet chyby

Po natrénování a spuštění predikce na testovacích datech je vypočtena chyba RMSE, podle vzorce uvedeného v kapitole 3.6.

kód výpočtu vypadá následovně:

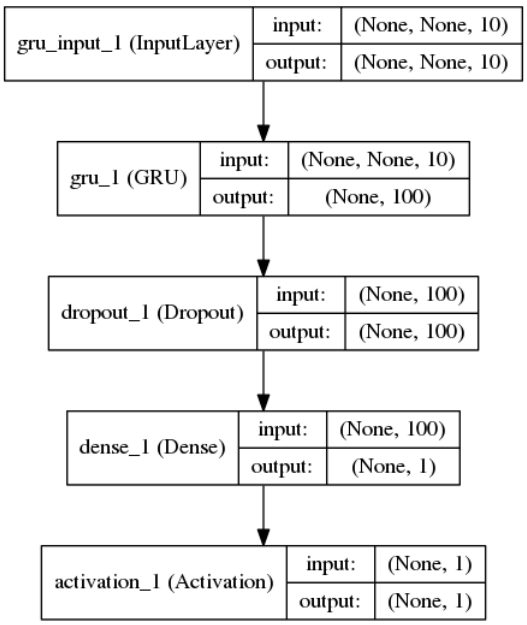
```
np.sqrt(((predicted - y_test) ** 2).mean(axis=0)).mean()
```

7.4 Implementace obchodního systému

Obchodní systém je implementován v souboru *mg_trade_strategy.py*. V tomto zdrojovém souboru se nachází čtyři funkce. Funkce *handle_data* a *initialize* jsou přesněji metody, které je nutné doplnit ke správnému spuštění obchodní strategie využívající knihovnu zipline. Metoda *initialize* se volá pouze jednou při spuštění strategie a slouží k inicializaci počátečních hodnot. Metoda *handle_data* se pak volá v každém kroku a obsahuje celou obchodní strategii. Pokud jsou splněné podmínky pro obchod, je volána funkce order_target s 1. parametrem, obsahující název akcie a druhým parametrem, obsahující počet akcií, které chceme koupit.

Pomocná funkce *get_sec_data* slouží k získání příslušných ekonomických ukazatelů, například ze souboru "DD_and_ratios.csv".

Poslední obsažená funkce nese název *analyze* a slouží k zobrazení výnosnosti strategie na zvoleném intervalu historických dat ve formě grafu.



Obrázek 7.1: Schéma modelu síťě GRU pro jeden akciový titul

Kapitola 8

Testování a výsledky

8.1 výsledky natrénované neuronové sítě

Testování rekurentní neuronové sítě probíhalo na mnoha různých parametrech (počet skrytých neuronů, počet epoch, počet kroků)a na různých datech. Na základě vypočtené chyby byly poté parametry optimalizovány. Výsledky testů pro akcie firmy DuPont jsou zobrazeny v tabulce 8.1 níže.

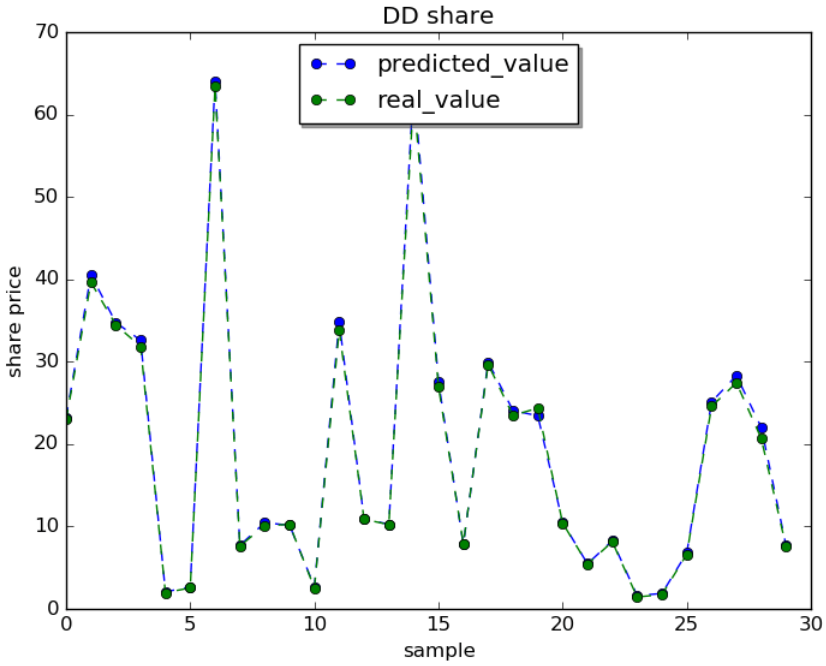
Tabulka 8.1: Výsledky RMSE při různých parametrech RNN

model	počet trénovacích vzorků	skryté neurony	kroků	epoch	čas CPU	rozdeleni dat	LSTM RMSE(%)	GRU RMSE(%)
1	6053	100	200	10	1565 s / 1148 s	68/12/16	1.705	0.87
2	6053	100	200	100	/ 10474 s /	68/12/16	0.93	0.89
3	6053	50	200	10	920 s / 624 s	68/12/16	1.14	1.16
4	7130	50	100	10	509 s / 324 s	68/12/16	1.05	1.53
5	7130	50	100	50	2523 s /1677 s / 616 s	68/12/16	0.83	0.74
6	6430	50	100	50	2172 s / 1472 s	64/16/20	1.22	1.13

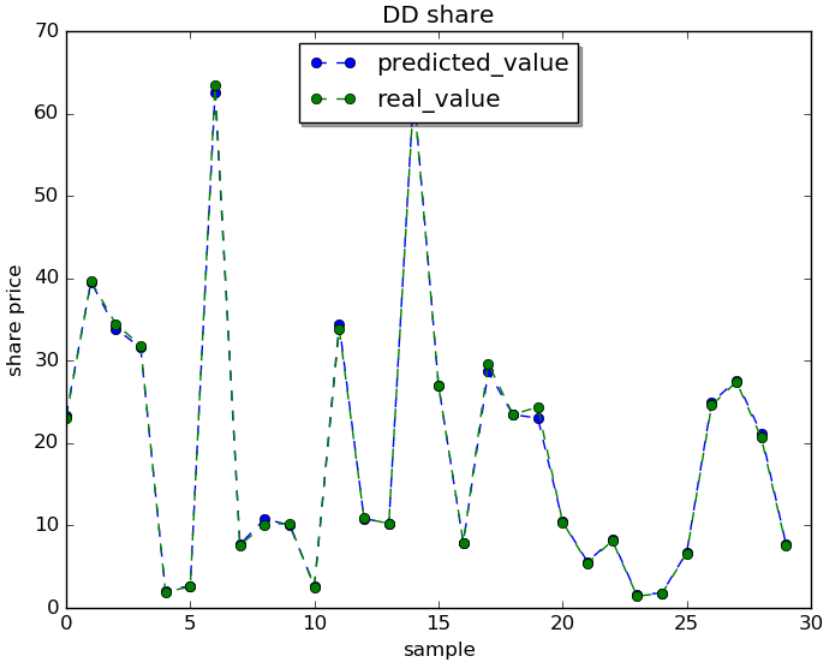
Níže na grafech je zobrazeno několik výstupních testovacích hodnot pro rekurentní neuronovou síť. Pro porovnání jsou zobrazeny 3 grafy pro různé typy neuronových sítí. Výsledky jednoduché rekurentní neuronové sítě se nachází na obrázku 8.1, pro bránovou LSTM síť na obrázku 8.2 a pro bránovou síť GRU na obrázku 8.3. Jedná se o výsledky modelu 5 z předešlé tabulky 8.1, u kterého byla dosažena nejnižší hodnota chyby RMSE. Modré body jsou hodnoty predikované (predicted_value) a zelené jsou reálné hodnoty testovacích dat. Na ose y je zobrazena cena akcie.

8.2 výsledky obchodní strategie

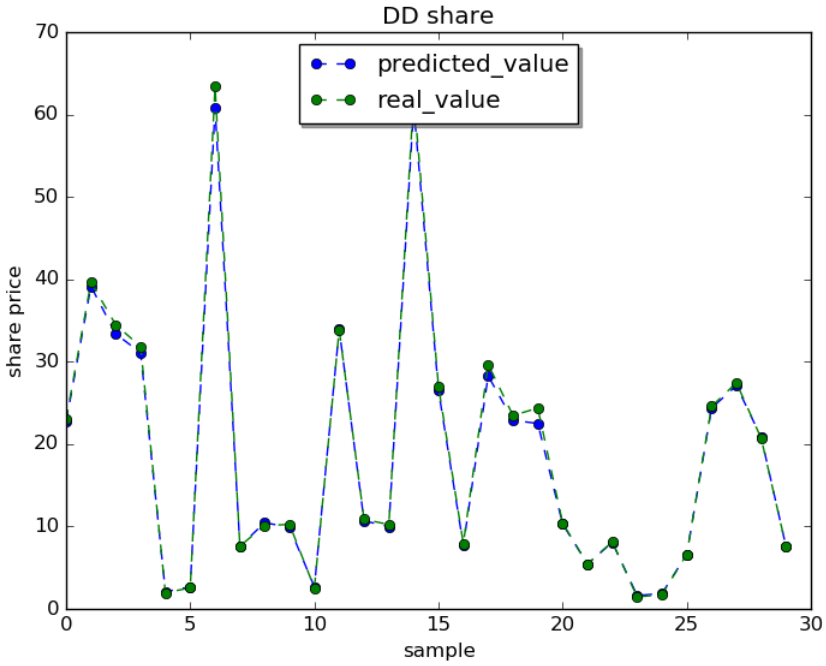
Pro testování obchodní strategie bylo zvoleno období od poloviny roku 2011 do 1.1. 2016. V tomto období se nacházejí všechny tři trendy i směry. V první fázi se trh pohyboval v úzkém rozpětí a nacházel se tedy v době akumulace, poté přišla fáze expanze a dlouhodobý trend. V roce 2015 došlo k prudkým výprodejům a následně zase k prudkému růstu ceny akcie. Pro toto časové období, byla vyzkoušena obchodní strategie s různými typy neuronových sítí. Níž na obrázku 8.4 je výsledek výchozí obchodní strategie, která nám poslouží k porovnání výnosnosti jednotlivých systémů.



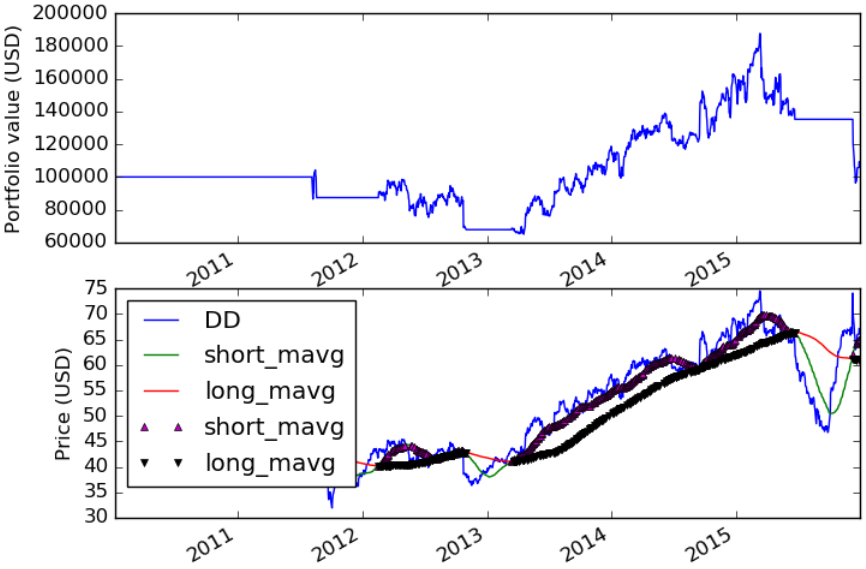
Obrázek 8.1: Výsledek predikce pro jednoduchou RNN



Obrázek 8.2: Výsledek predikce pro síť LSTM



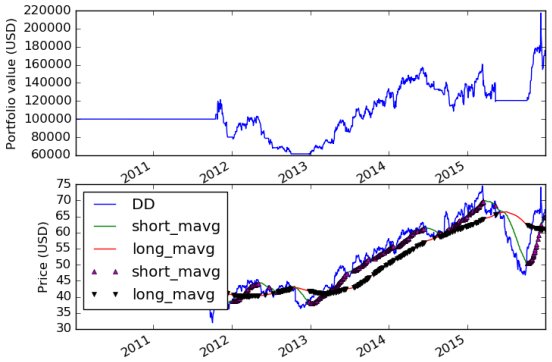
Obrázek 8.3: Výsledek predikce pro síť GRU



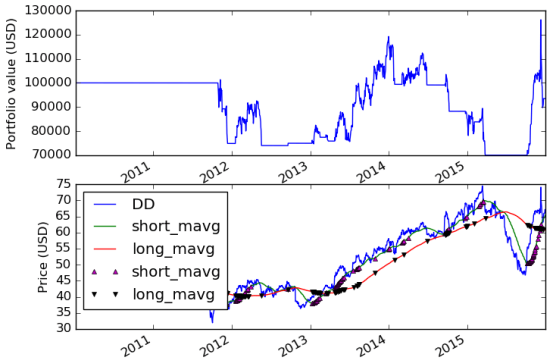
Obrázek 8.4: Výchozí obchodní strategie bez neuronové sítě

Na obrázcích 8.5 a 8.6 jsou zobrazeny výsledky jednoduché rekurentní neuronové sítě využívající v prvním případě pouze technická data a v druhém i data fundamentální. V tomto případě dosáhla lepšího výsledku síť pouze z technickými daty, kdežto síť s fundamentálními daty skončila v mírné ztrátě. Šipka nahoru u legendy značí nákup a šipka dolů značí prodej akcií. Nákupů a prodejů za dané období je hodně, což trochu znepřehledňuje spodní graf na obrázku, ale za to můžeme dobře vidět, že v době prudkého propadu ceny nedochází k otevírání obchodů.

Přesné výsledky jsou uvedeny v tabulce 8.2.

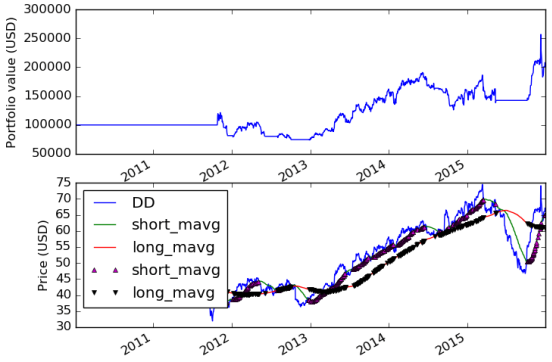


Obrázek 8.5: Jednoduchá rekurentní neuronová síť pouze s technickými daty

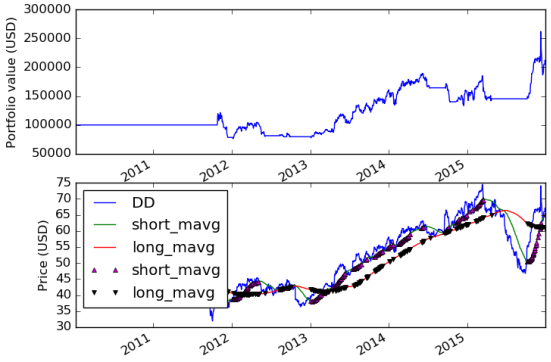


Obrázek 8.6: Jednoduchá rekurentní neuronová síť i s fundamentálními daty

Na dalších obrázcích 8.7 a 8.8 jsou výsledky bránové rekurentní neuronové sítě GRU, u které je vidět výrazné zlepšení oproti jednoduché neuronové sítě. I když to není z grafu úplně patrné, při použití fundamentálních dat bylo dosažené lepšího výsledku než pouze z technickými daty. V porovnání s předešlou neuronovou sítí bylo také zamezeno větším propadům a celkový zisk je o třetinu větší. Přesné hodnoty nalezneme v tabulce 8.2.

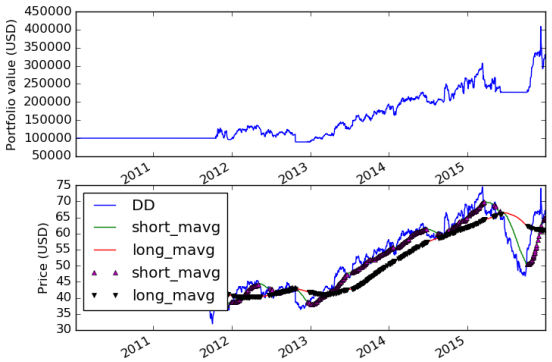


Obrázek 8.7: GRU rekurentní neuronová síť pouze s technickými daty

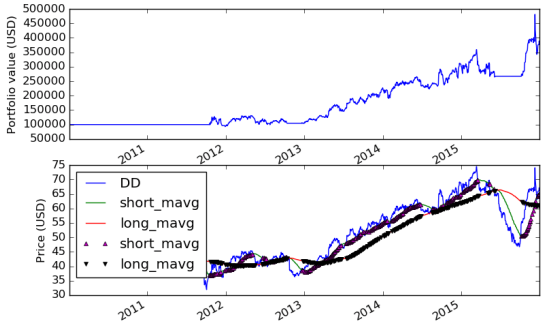


Obrázek 8.8: GRU rekurentní neuronová síť i s fundamentálními daty

Poslední testovanou sítí je sít LSTM, jejíž výsledky vidíme na obrázcích 8.9 pro technická data a 8.10 pro data technická i fundamentální. Při použití tohoto typu sítě bylo dosaženo nejlepšího výsledku oproti předešlým sítím a to poměrně viditelně. Zisky jsou oproti sítě GRU skoro 3x vyšší a je viditelný také rozdíl mezi použitím pouze technických dat a použitím, jak technických dat, tak fundamentálních. Přesné údaje jsou opět uvedeny v tabulce 8.2.



Obrázek 8.9: LSTM rekurentní neuronová síť pouze z technickými daty



Obrázek 8.10: LSTM rekurentní neuronová síť i s fundamentálními daty

Tabulka 8.2: Výsledky obchodní strategie z různými neuronovými sítěmi

portfolio	jednoduchá RNN s TD	jednoduchá RNN s TD i FD	GRU s TD	GRU s TD i FD	LSTM s TD	LSTM s TD i FD
nejnižší hodnota ($)	60 761	70 225	70 833	74 107	85 952	87 500
nejvyšší hodnota ($)	215 810	125 679	252 530	259 007	402857	476 429
konečná hodnota ($)	173 905	**94 250**	203 423	208 333	329 524	**382 143**
konečný zisk (%)	73,9	**-5,75**	103,4	108,3	229,5	**282,1**

TD = technická data, FD = fundamentální data

Kapitola 9

Závěr

Cílem této práce bylo využít fundamentální informace a zahrnout je do obchodního systému, čímž by se zvětšila výnostnost daného systému. Ukázalo se, že fundamentální data napomáhají ke zlepšení přesnosti predikce budoucího vývoje ceny. Z výsledků je taky patrné, že s novými architekturami sítí (LSTM, GRU) lze dosahovat mnohem lepších výsledků, než u starších rekurentních neuronových sítí. Nejvyššího zhodnocení bylo dosaženo u LSTM sítě. Zhodnocení kapitálu o 280% považuji za velmi slušné.

Přestože se vytvořený obchodní systém ukázal jako ziskový, je stále co vylepšovat. Proto uvádím v následující podkapitole možný směr vývoje této práce.

9.1 Budoucí práce

- Sledovat databázi Edgaru a zpracovávat nově přidané SEC filings
- Vylepšit obchodní strategii o další indikátory - RSI, Fibonacciho posloupnost
- Přidat do obchodního systému VSA (Volume Spread Analysis), což je analýza objemu obchodů
- Přidat do portfolia další akciové tituly, což ale zvýší hardwarové nároky
- Vyzkoušet vyšší timeframe například týdenní, popřípadě otevírat obchody podle informací z několika timeframů
- Využít některého z řešení, které poskytuje zpracované fundamentální data a rozšířit neuronovou síť o tyto informace
- Spustit obchodní systém na demo účtu, popřípadě na reálném účtu a sledovat jak si povede při skutečném obchodování

Literatura

[1] Altova: Sec-xbrl. https://github.com/altova/sec-xbrl, 2014.

[2] Cabrera, J.: python-xbrl. https://github.com/greedo/python-xbrl, 2014.

[3] Chollet, F.: Keras. https://github.com/fchollet/keras, 2015.

[4] Hebert, E.; fawce; Wiecki, T.; aj.: Zipline. https://github.com/quantopian/zipline, 2013.

[5] Hochriter, S. and Schmidhuber, J: Long short-term memory. *Neural Computation*. 1997 [cit. 1997], [Online; navštíveno 30.4.2016]. URL http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf

[6] ifcmarkets: *the-dow-theory-in-technical-analysis*. "[Online; navštíveno 1.5.2016]". URL http://www.ifcmarkets.com/pdf/tradingbooks/en/the-dow-theory-in-technical-analysis.pdf

[7] Obrázek: *Formální neuron*. http://programujte.com/galerie/2005/08/200508191445_obr1.jpg.

[8] Obrázek: *GRU buňka*. http://d3kbpzbmcynnmx.cloudfront.net/wp-content/uploads/2015/10/Screen-Shot-2015-10-23-at-10.36.51-AM.png.

[9] Obrázek: *LSTM*. http://blog.otoro.net/wp-content/uploads/sites/2/2015/05/LSTM.png.

[10] Obrázek: *Rekurentní vrstevnatá neuronová síť*. http://www.cogcrit.umn.edu/images/krause/fig3_krause.jpg.

[11] Obrázek: *Vrstevnatá neuronová síť*. https://encrypted-tbn2.gstatic.com/images?q=tbn:ANd9GcSMQjuTpIgCdwjKq2D0--_NEK1GSerUb0DJFSeVX2zMUEgY3ANjKQ.

[12] prof. Ing. Ivo Vondrák, CSc: *Neuronové sítě*. "[Online; navštíveno 25.4.2016]". URL http://vondrak.cs.vsb.cz/download/Neuronove_site.pdf

[13] Ranjan, R.: SEC-Edgar. https://github.com/rahulrrixe/SEC-Edgar, 2014.

[14] Rhea, R.: *The Dow Theory* . Snowballpublishing, 2013, ISBN 978-1607966289.

[15] Turek, L.: *Price Action - jak vidět, co jiní nevidí* . Czechwealth, spol. s r. o., 2014.

Seznam příloh

Kapitola 10

Obsah DVD

- /data - obsahuje technická a fundamentální data
- /data/models_and_weights - obsahuje modely neuronových sítí a natrénované váhy
- /tests - zde jsou skripty pro trénování neuronové sítě a testování obchodní strategie
- /source_codes - obsahuje zdrojové soubory pro přípravu dat a externí knihovny

Kapitola 11

Manual

Trénování sítě:

Pro natrénování GRU neuronové sítě spusťte:

tests/rnn_with_sec.py -type GRU

Pro natrénování LSTM neuronové sítě spusťte:

tests/rnn_with_sec.py -type LSTM

Pro natrénování jednoduché rekurentní neuronové sítě spusťte:

tests/rnn_with_sec.py -type SimpleRNN

Spuštění obchodní strategie:

Pro strategii využívající GRU síť pouze s technickými daty spusťte:

tests/mg_trade_strategy.py -type GRU –onlytech

Pro stejnou síť ale i s daty fundamentálními spusťte skript bez posledního parametru:

tests/mg_trade_strategy.py -type GRU

Pro další typy neuronových sítí je spuštění obdobné, mění se pouze parametr -type.

Stažení finančních dat (není nutné ke spuštění testů):

Pro stažení všech SEC filings spusťte:

source_codes/SEC-Edgar-master/SECEdgar/download_sec_edgar_data.py

Pro stažení všech xbrl souborů spusťte:

source_codes/sec-xbrl-master/loadSECfilings.py -f 2005 -t 2016

pro jejich rozbalení:

source_codes/extract_xbrl_filings.py

poté je možné spustit parsovací skripty:

source_codes/parse_SEC_filings.py
source_codes/parse_xbrl_filings.py

Následně můžete spojit všechna data spuštěním:

source_codes/merge_SEC_filings.py

a vytvořit konečný soubor pro akcie DD spuštěním:

source_codes/count_ratios_and_merge_data.py