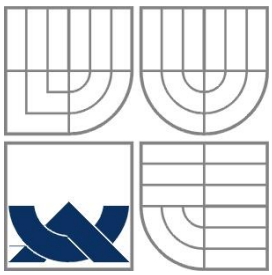
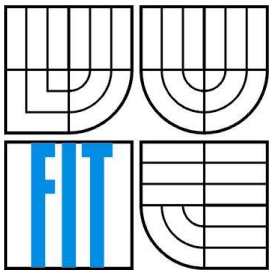


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

PREDIKCE HODNOT V ČASE

PREDICTION OF VALUES ON A TIME LINE

DIPLOMOVÁ PRÁCE
MASTER‘S THESIS

AUTOR PRÁCE
AUTHOR

Bc. Michal Marša

VEDOUCÍ PRÁCE
SUPERVISOR

Prof. Dr. Ing. Pavel Zemčík

BRNO 2016

Abstrakt

Tato práce se zabývá predikcí číselných řad, jejichž aplikace je vhodná i pro predikci vývoje cen na burze. Jsou vysvětleny postupy analýzy a práce s cenovými grafy. Také jsou objasněny způsoby strojového učení. Znalosti jsou využity k sestavení programu, který v řadě nalezne vzory umožňující predikci.

Abstract

This work deals with the prediction of numerical series whose application is suitable for prediction of stock prices. They explain the procedures for analysis and works with price charts. Also explains the methods of machine learning. Knowledge is used to build a program that finds patterns in numerical series for estimation.

Klíčová slova

Číselné řady, časové řady, cenové řady, predikce, klasifikace, strojové učení, regresní analýza, neuronové sítě, genetické algoritmy, vzory, získávání znalostí, předzpracování dat, čištění dat, technická analýza, fundamentální analýza, automatické obchodní systémy, komoditní trhy.

Keywords

Numeral series, time series, price series, prediction, classification, machine learning, regression analysis, neural network, genetic algorithm, patterns, knowledge extraction, data preparation, data cleaning, technical analysis, fundamental analysis, automated trading system, commodity market.

Citace

Marša Michal: Predikce hodnot v čase, diplomová práce, Brno, FIT VUT v Brně, 2016

Predikce hodnot v čase

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením Prof. Dr. Ing. Pavla Zemčíka.
Další informace mi poskytli Monika Bartošová, Ondřej Šťastný, Marek Kaleta a Gabriela Novotná.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Michal Marša
24. 5. 2016

Poděkování

Rád bych poděkoval svému vedoucímu, za veškerou trpělivost a cenné rady, které pomohli dodat mé práci formu a směr. Velkou psychickou oporou mi byla partnerka Marta, která mi pomohla dodat sílu a potřebnou energii k napsání celé práce. Také bych rád poděkoval všem mým přátelům, kteří aktivně na burze obchodují a mnoha hodinám jejich času, který obětovali, aby mě seznámili s danou problematikou.

Obsah

1 Úvod

Lidé se už od pradávna snaží odhadovat nejrůznější aspekty budoucnosti. Například podle počasí, se pokoušeli predikovat, jaká bude úroda. V průběhu desetiletí, pak lidé nasbírali dostatečné množství dat k tomu, aby mohli vytvořit pravidla, která by dané poznatky zobecňovala. Asi nejznámějšími pravidly jsou lidová pořekadla, která zohledňují dlouhé zkušenosti pozorování a vývoje přírody. V dnešních dobách jsou lidé obklopeni stále více informacemi a s nástupem počítačů se mnohonásobně zvýšila schopnost lidí s těmito informacemi pracovat. Lidé však mnohem raději naučí počítač, jak za ně dané vzory objevovat a následně s jeho pomocí zpracují miliony záznamů (například o počasí z celého světa, za posledních sto let). Výsledné vzory jsou vytvořeny nad ohromným množstvím předchozích vzorků, a proto většinou bývají v predikci mnohem přesnější. Jelikož každá správná predikce znamená výhodu nad ostatními, je celkem logické, že se o ni lidé pokoušejí i v obchodní sféře. Asi nejrozsáhlejší oblastí pro predikce se tak staly akciové a burzovní trhy.

Aby stroj mohl predikovat nějaké výsledky nebo jevy, musí nejdříve „pochopit“, jakým způsobem zkoumaný systém pracuje. K tomuto procesu se využívá strojové učení a klasifikace. Strojové učení a získávání nejrůznějších znalostí o jiném systému (například reálném světě kolem nás) je složitým úkolem, a aby bylo možné dosáhnout alespoň přijatelných výsledků, musí během něj tvůrce programu provést velké množství úkonů. Teprve po úspěšném učení (přenesením reálného problému do modelu, se kterým dokáže počítač pracovat) může systém začít s predikcí následujícího vývoje. Práce se pokouší typické postupy objasnit a ukázat, jak se výsledný program s jednotlivými problémy vypořádal.

Pro strojové učení je vhodné mít dostatečné množství informací. Tato práce využívá cenové řady, konkrétně vývoj ceny nad burzovními komoditami. Ceny komodit se mění poměrně často (i několikrát do minuty), a proto nebyl problém získat rozsáhlý vzorek dat směrem do historie. Vytvořený program tak pracuje s několik let dlouhou číselnou řadou, která mu umožnuje zbavit se statistických odchylek v datech.

Postup učení nad číselnými řadami, je obecný, a měl by být využitelný i nad jinými číselnými řadami. Například při dostatečném vzorku dat, by mohl program provádět předpověď počasí. Každá číselná řada vzniká v jiném prostředí, a v procesu učení jsou využívány heuristiky, které dané prostředí zohledňují. Tyto heuristiky nemohou být obecné a pro využití programu nad jinými řadami, by tedy musely být nalezeny jiné heuristiky pro zpřesnění výsledků.

Přestože se využívají postupy pro zpřesnění výsledků přímo pro dané simulované prostředí, předpovědi nebývají nikdy 100% přesné, a vždy se může vyskytnout neočekávaná událost, která výsledky ovlivní. Ze samotného vývoje ceny není pro systém možné predikovat pád Řecké ekonomiky, přesto daný pád nastal a výrazně ovlivnil vývoj cenové řady.

Složitost vývoje cen v reálné ekonomice a ještě na celosvětové úrovni, je však procesem natolik komplexním, že nebylo v silách autora, aby obsáhl všechny proměnné, které cenu ovlivňují (některé jsou dodnes skryty i předním odborníkům), a proto predikce ani nemohou být dogmaticky přesné. Výsledný program, tedy není orákulum, předpovídajícím vývoj budoucnosti, ale přesto by měl být schopen pomoci vytvořit systém dostatečně robustní, aby se s ním dalo v reálném světě pracovat a generovat zisk.

Přestože predikce nemůže být 100%, neznamená to, že by nemohla být využita. Ostatně k práci profesionálů se systémy s jistou mírou odchylky běžně používají, a protože predikce nejsou vždy naplněny, musí být využívány další metody pro snižování rizika. Přesto by bylo vhodné, aby každý čtenář věděl, že dokud tyto systémy nebudou schopné zpracovávat a modelovat kompletní dění celého

světa a každého jedince v něm, nebudou ani schopné s naprostou jistotou odhadovat jeho následný vývoj.

Predikce čehokoliv, na základě dostatečného vzorku dat, je v dnešní době velmi rychle se rozvíjející směr v oblasti počítačových technologií. Predikce vývoje cen je téma, o kterém se velmi často mluví. Mnoho firem a společností, které jsou na vývoji burzovní ceny závislé, investují nemalý kapitál, aby byly schopny odhadovat růst, či pád svých konkurentů. Přesto mnoho podrobných prací na dané téma neexistuje, protože takto získané znalosti jsou soukromým „know-how“, které firmám poskytují značnou konkurenční výhodu (a to i v případě, že predikce jsou přesné jen z části). Existuje sice mnoho knih, které se problematikou zabývají, ale jen málokterá z nich odhalí čtenáři svůj využívaný obchodní systém a postupy přesné analýzy. Práce se proto snaží ukázat jednotlivá úskalí, která sebou analýza daných dat přináší a zároveň umožnit získání přesných vzorů, které zpracovaný program vytvoří.

Jelikož se na burze (ale i ve světě obecně) vyskytuje velké množství dat, která se neustále mění, je důležité zpracovávat data rychleji než ostatní. Existuje mnoho pokusů predikci zpracovávat na specializovaném hardwaru, který by byl schopen reagovat v reálném čase. Rychlost pak poskytuje výhodu, která obzvláště v obchodním světě, kde je v sázce nemalý kapitál, hraje velmi důležitou roli.

V následujících kapitolách jsou vysvětleny číselné řady, práce s nimi a postupy, které se využívají k analýzám v obchodním světě (kapitola 2). Poté, co se čtenář seznámí s burzovním prostředím, mu jsou představeny postupy strojového učení (kapitola 3). Po představení možných postupů se autor pokusí o shrnutí nejvhodnějších metod (kapitola 4). Následně je vše aplikováno pro vytvoření praktického programu, jehož výsledky jsou následně zhodnoceny (kapitola 5).

2 Prostředí číselných řad

V následujících kapitolách, se čtenář nejprve seznámí s číselnými řadami, základními operacemi s nimi a možnými způsoby jejich predikce. Ve druhé části kapitoly jsou předvedeny některé postupy, které se v burzovním světě uplatňují v souvislosti s tvorbou obchodního systému. Obchodní systém obsahuje skupiny vzorů, které se snaží vývoj číselných řad predikovat. Kapitola představuje shrnutí postupů, které s predikcí číselných řad souvisejí. Nejedná se však o výčet všech možností, neboť takto podrobný seznam by převyšoval rozsah této práce.

2.1 Číselné řady

Tato práce je zaměřena na predikci číselných řad. Číselné řady a jejich vlastnosti jsou základem pro ostatní řady, které jsou z nich odvozené (například časové řady a cenové řady). Tyto cenové řady mají jisté specifické odlišnosti, které jsou probrány v následujících kapitolách. Než se však autor pustí do bližšího rozboru jednotlivých vlastností, bylo by dobré seznámit čtenáře s obecnými číselnými řadami, ze kterých veškerá další práce vychází. Číselná řada je definována takto: [32]

$$a_1, a_2, \dots, a_n; \text{kde } n \in \mathbb{N} \tag{2.1.1}$$

Jedná se tedy o posloupnost různých hodnot, s tím že tato posloupnost má nějakou obecnou délku. Podlé své délky se řady dělí na konečné, nebo nekonečné a nad oběma typy číselných řad je definováno mnoho operací, které usnadní jejich pochopení a práci s nimi. Jednotlivé operace jsou velmi důležité, neboť pomáhají řadu charakterizovat, což je jeden ze základních prvků k pochopení dané řady. Teprve když analytik dokáže číselnou řadu pochopit, může se pokusit ji predikovat.

Operace s číselnými řadami

V této podkapitole práce objasní mnoho operací, které se nad číselnými řadami vykonávají, aby byla získána nějaká charakteristika o dané číselné řadě. Výčet vzhledem k rozsahu práce není úplně kompletní, ale jsou zde zmíněny nejčastější postupy, které analytici využívají. Před samotným výčtem operací by se čtenář měl dozvědět a zamyslet nad výpočetní náročností dané operace. Při hledání vhodných prediktivních vzorů musí být zpracováno velké množství historických dat. Tyto výpočty by mohly zabrat nepřiměřeně dlouhou dobu a je proto vhodné je rozdělit mezi více strojů. Distribuovaný výpočet na více počítačích sice výrazně urychlí čas analýzy, ale pro některé matematické operace není toto rozdělení možné. Je tedy nutné zvážit míru vhodné distribuce. Veškeré operace se dělí do tří kategorií podle vhodnosti pro distribuci:

Distributivní míra

Jedná se o míru, při které je možné výsledek operace získat pomocí snadného rozdělení na menší části a provedením dané operace nad každou touto částí. Výsledná hodnota je pak sumou jednotlivých dílčích výsledků. [13] Jako příklad může sloužit operace pro získání počtu slov v souboru. Soubor může být rozdělen na jednotlivé řádky, každý řádek může být zpracován na jiném stroji a celkový počet slov v celém souboru je pak získán sumou slov na jednotlivých řádcích.

Distributivní míra je velmi vhodná pro paralelní zpracování více vlákny/jádry/stroji.

Algebraická míra

Tuto míru lze získat aplikací nějaké algebraické operace, nad několika distributivními mírami. [10] Příkladem může být průměr hodnot v souboru. Nejdříve se nad všemi čísly v souboru musí provést jejich suma (což je distributivní míra) a zároveň je nutné spočítat počet všech čísel, které se v souboru nacházejí (opět distributivní míra). Každá z operací může být zpracována odděleně několika stroji. Po získání hlavního výsledku (po sečtení dílčích výsledků) je provedena algebraická operace dělení, která již musí být zpracována na jednom stroji. [13]

Holistická míra

Vlastností holistické míry je jistá nepříjemnost pro paralelní zpracování a to v tom, že se daná operace musí provádět nad celým seznamem všech hodnot. Operace mající holistickou míru, tak musí vždy pracovat se všemi daty a jsou časově náročné, neboť není možné dané výpočty provádět distribuovaně. [13]

Charakteristiky řad

Nyní práce představí jednotlivé operace, které se pro analýzu číselných řad využívají. Jedná se především o způsoby získání nějaké charakteristiky řad (průměry, kvantily a míra přírůstku), uhlazení řad (klouzavé průměry) a podobnost řad, kterou představuje korelace.

Absolutní přírůstek

Jedná se o absolutní přírůstek v řadě mezi dvěma hodnotami. Absolutní přírůstek spadá do míry dynamiky dané řady neboli jejího vývoje:

$$\Delta a_m = a_m - a_{m-1} \tag{2.1.1}$$

Pro celou řadu je pak vhodné spočítat průměrný absolutní přírůstek: [32]

$$\bar{\Delta} = \frac{1}{n-1} * \sum_{m=2}^{n} \Delta a_m \tag{2.1.3}$$

Je možné využít i jinou metodu než je aritmetický průměr (metody vysvětleny později).

Koeficient růstu

Neboli také tempo růstu řady opět přináší údaje o dynamice řady a jejím růstu [32].

$$k_m = \frac{a_m}{a_{m-1}} \tag{2.1.4}$$

Průměrný koeficient růstu řady se pak získá jako exponenciální průměr jednotlivých koeficientů.

$$\bar{k} = \sqrt[n-1]{k_2 * k_3 * \dots * k_n} \tag{2.1.5}$$

Aritmetický průměr

Jedná se asi o nejznámější veličinu, určující polohu. Aritmetický průměr slouží k přiblížení se středu různých hodnot (určuje polohu středu), které mají nějaký rozptyl.

Pro *n* hodnot s hodnotou a_i kde i je od 0 po *n* se aritmetický průměr spočítá pomocí: [13]

$$\bar{a} = \frac{1}{n} * \sum_{i=0}^{n} a_i \qquad (2.1.6)$$

Nevýhodou aritmetického průměru je, pokud jedna z hodnot výrazně vybočuje a tím posune výsledek dále od středu. Tato vlastnost normálně vadit nemusí, ale může být vhodné mít představu o tom, ve které části se hodnoty převážně pohybovaly.

Vážený průměr

Jedná se o typ průměru, který se velmi hodí, pokud jednotlivé položky mají různou důležitost (váhu). [13] Dokonce, i když mají všechny hodnoty stejnou váhu, může být výhodné nějak penalizovat starší hodnoty pomocí snižování váhy směrem do historie.

Vážený průměr se spočítá jako součet součinu hodnoty a váhy dané hodnoty, podělený součtem vah: [31]

$$\bar{a} = \sum_{i=1}^{n} w_i * a_i \Big/ \sum_{j=1}^{n} w_j \qquad (2.1.7)$$

Exponenciální průměr

Tento typ průměru, se snaží o snížení oné odchylky od středu hodnot. Je vhodné ho využívat u tzv. přírůstkových hodnot, což cenové grafy jistě jsou. Geometrický průměr se spočítá jako N-tá odmocnina ze součinu N prvků: [4]

$$\left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}} = \sqrt[n]{a_1 * a_2 * \ldots * a_n} \qquad (2.1.8)$$

Median

Medián je velmi vhodnou mírou pro vychýlená (asymetrická) data, protože zohledňuje množství prvků dané hodnoty. Medián je prostřední hodnota ze seřazených dat. Pokud je počet dat lichý, je medián prostřední z těchto hodnot. Pokud je počet prvků naopak sudý, je mediánem aritmetický průměr dvou prostředních hodnot.

Mezi největší nevýhody mediánu pak patří, že se jedná o holistickou míru, což znamená, že se špatně distribuuje (vždy se musí seřadit celá posloupnost prvků). Medián se určí dle následujícího vztahu: [13]

$$median = L_1 + \left(\frac{n/2 - (\sum freq) * l}{f_{median}}\right) * c \qquad (2.1.9)$$

L_1 – je dolní hranicí intervalu mediánu
n - je počet prvků v seznamu
$(\sum freq) * l$ - je součet frekvencí všech intervalů obsahujících hodnoty menší než L_1
f_{median} – je frekvencí intervalu mediánu
c – je šířka intervalu mediánu

Modus

Další mírou je modus, který vystihuje hodnotu s největší frekvencí výskytu. [31] Pro symetrické rozložení dat je průměr, medián i modus stejný, u vychýlených dat se již výsledné hodnoty liší. [13] Pro různé analýzy se pak hodí práce s různými mírami nebo jejich kombinace.

Rozptyl

Rozptyl dané řady značí míru druhých mocnin odchylek od průměru řady. [32]

$$s_a^2 = \frac{1}{n-1} * \sum_{i=1}^{n} (a_i - \bar{a})^2 \tag{2.1.10}$$

Směrodatná odchylka

Jedná se o odmocninu rozptylu: [32]

$$s_a = \sqrt{s_a^2} = \sqrt{\frac{1}{n-1} * \sum_{i=1}^{n} (a_i - \bar{a})^2} \tag{2.1.11}$$

Kvantily

Jedná se o další míru polohy, která se snaží jednotlivé proměnné rozdělit do q skupin. Každá skupina představuje pravidelný interval pravděpodobnosti, který říká, že pro k-tý q-kvantil je hodnota x taková, že pravděpodobnost, že náhodná veličina X bude menší než x je nejvýše k/q. Hodnoty kvantilů mohou být odhadnuty různými způsoby, například:

$$i = k * \frac{n}{q} \tag{2.1.12}$$

Některé známé kvantily mají speciální jména, u nichž si snad každý dokáže představit, jak se s kvantily pracuje, protože se s nimi již setkal: [13]

- 100-kvantil je označován jako percentil.
- 10-kvantil je decil
- 5-kvantil je kvintil
- 4-kvantil je kvartil a právě s nimi se nejvíce pracuje při analýze cenových řad, a to v podobě krabicových grafů (je vysvětleno později v kapitole 2.4 Vizualizace číselných řad)

Korelace

Korelace vyjadřuje míru podobnosti jedné číselné řady nad řadou jinou. Korelace dvou číselných řad *X* a *Y* je dána vztahem: [32]

$$s_{xy} = \frac{1}{s_x * s_y} * \sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y}) \tag{2.1.13}$$

$s_{xy} \in \langle -1 ; 1 \rangle$; a kde hodnoty S_x a S_y
vyjadřují směrodatnou odchylku dané číselné řady.

Pokud je hodnota S_{xy} rovna -1 znamená to, že obě číselné řady mají přesně obrácený směr vývoje. Pokud jedna řada vykazuje stoupající tendenci, druhá řada je klesající a naopak. Naproti tomu hodnota 1 reprezentuje stav, kdy jsou obě číselné řady shodné. Jakákoliv hodnota pak představuje míru mezi těmito možnými stavy.

2.2 Predikce řad

Nyní práce nastíní některé metody používané pro predikci číselných řad. Jedná se o základní metody, které pro efektivní využití v obchodním světě nemusí být použitelné, ale je vhodné některé postupy ukázat, aby měl čtenář ucelenější představu o problematice.

Transformace měřítka

Transformace měřítka může mít využití, například pokud si analytik přeje dávat na starší data menší důraz. Nejčastěji se tato metoda obecně využívá pro snížení rozptylu hodnot [32] (který se v čase exponenciálně zvyšuje). Logaritmováním nebo odmocněním, se pak dají exponenciální tendence řady potlačit.

Při využití této metody se staré záznamy stávají plytkými a hrají nižší rozpoznávací roly. Analytik se tak může soustředit na hodnoty aktuálnější. [2] Hlavní směr starších dat však zůstává zachován a to pomáhá analytikovy odhalit dlouhodobý vývoj řady.

Vyhlazování řad

K vyhlazování se využívají nejrůznější klouzavé průměry o různé šířce a slouží analytikům hlavně ke stanovení trendu řady. Tato technika se využívá hlavně k odstranění šumu, který do řady zanesla drobná odchylka měření z důvodu příliš častého měření. [34] Jedná se o stav, kdy se samotná řada změnila jen minimálně (časový interval byl pro změnu příliš krátký), ale změna je způsobena odchylkou čidla, které hodnotu řady zaznamenává. Poté klouzaví průměr zahladí tyto náhodné odchylky a řada bude zobrazovat jen reálné hodnoty.

Asi nejčastější metodou pro vyhlazování řady je klouzavý průměr z předchozích hodnot. Jak již název napovídá, jedná se o aritmetický průměr (nebo jiný typ průměru) dané hodnoty a nejbližších předchozích hodnot.

Lineární dynamické modely

Nejčastěji se jedná o příčinné, neboli takzvané kauzální modely. Tyto modely se zabývají hledáním proměnné a_t (v podstatě predikcí vývoje číselné řady) pomocí rozpoznání šumu v řadě, a hledání závislostí jednotlivých hodnot (k tomu je použita převážně korelační analýza). Dále se pro zpřesnění výsledků hodí pracovat s řadou příčinných faktorů. [32] Ty by nad cenovými řadami burzy představovaly míru ovlivnění řady pomocí fundamentální analýzy, což jak je vysvětleno v kapitole 2.6 Analýzy komodit je prakticky těžce realizovatelné.

Spektrální analýza časových řad

Na rozdíl od lineární dynamické metody, tento postup zkoumá číselnou řadu jako směs sinusových a kosinusových křivek, které mají různé amplitudy a frekvence. Pomocí statistických nástrojů se zjišťuje tzv. spektrum řady, čili její složení z hlediska poskládání sinusových a kosinusových funkcí. [35] Po získání daného spektra funkcí se metoda pokouší predikovat vývoj řady čistě dopočítáním budoucího stavu hodnot. Pro řady, které obsahují minimální šum, může tato metoda

rozkladu na sinusové a kosinusové funkce dávat dobré výsledky. Protože se využívá rozklad na funkce sinů a kosinů, někdy tato metoda bývá nazývána Fourierova analýza. [32]

Kalmanův filtr

Kalmanův filtr se pokouší predikovat polohu příštího prvku číselné řady a_t na základě funkce trasování objektů. Dle znalosti o současných hodnotách, tedy odhaduje trajektorie vývoje a tím umožnuje predikovat pravděpodobné následující hodnoty. [36] Metoda se řadí mezi dynamickou filtraci a provádí optimální odhady (optimal estimator), což znamená, že pokud řada nemá Gausovo rozložení, tak Kalmanův filtr minimalizuje průměrnou čtvereční odchylku od odhadovaných parametrů. Jedná se o rekurzivní metodu, která v každém běhu zpřesňuje míru své predikce.

Filtr je dvoufázoví a využívá predikci a filtraci. První fází je filtrace vzorku, která vytvoří bodový odhad pro střední hodnoty jednotlivých dat z daného vzorku číselné řady. Tento vektor středních hodnot se značí $X_{n|n}$ kde první *n* představuje n-té kolo odhadu, a druhé *n* znamená, že vychází z n-tého vzorku dat. [37] Po filtračním procesu se provádí prediktivní zpracování. Predikce pracuje se středními odhady a využívá matici očekáváného šumu pro zpřesnění predikce (například pokud je známa odchylka čidla, může matice šumu výrazně pomoci správnému odhadu hodnot). Pokud není šum znám, je využita matice s neúměrně velkými hodnotami, což znamená, že predikovaným hodnotám v úvodní fázi filtru není možné věřit. Predikce pak spočte nové hodnoty $X_{n+1|n}$ a novou matici šumu (nad stále stejným vzorkem dat). Tyto hodnoty se pak ověřují s predikcí z hodnot $X_{n|n-1}$. Pokud je predikce příliš nepřesná, využije se další iterace filtru s tím, že vzorek dat *n* může být rozšířen. [38]

Regresní analýza

Pod tímto názvem se skrývá celá kolekce metod, která slouží k odhalování nějaké náhodné veličiny (také někdy nazývané jako cílová proměnná) na základě znalosti jiných (častokrát předcházejících) veličin, kterým se říká regresory. [25] Lidé regresní analýzu využívají každý den, například při předpovědi počasí. Lidé zde využívají svoji znalost počasí za předchozí týden (tato znalost tvoří skupinu vstupních regresorů), také pak zohledňují informace o počasí v okolních zemích (tvořící druhou skupinu regresorů). Na základě těchto regresorů lidé určí, jaké počasí daný den nejpravděpodobněji bude, a podle toho pak volí oblečení.

Pokud by byla uvažována pouze aktuální hodnota řady, odpovídal by obecný zápis regresní funkce následujícímu vztahu: [13]

$$E(Y|X_1, X_2, \dots, X_n) = f(X_1, X_2, \dots, X_n) \qquad (2.2.1)$$

Kde E je symbolem střední hodnoty, Y je pak hledaným skalárem nebo vektorem (pokud by bylo hledáno více hodnot). *Y* je hledáno na základě znalosti X_1 až X_n. Tyto hodnoty vyjadřují předchozí hodnoty číselné řady. Funkce f je regresní funkce, která vytváří odhad hodnoty *Y* dle vstupních hodnot *X*. Podrobněji je daná metoda probrána v kapitole 3.

2.3 Další typy řad

Číselné řady existují v mnoha obměnách a rozšířeních, a proto nelze čtenáře seznámit se všemi možnostmi. Následující podkapitoly se věnují pouze takovým typům řad, které rozšiřují časové řady o vlastnosti, které se uplatňují na komoditních trzích.

Časové řady

V této práci je možné definici číselné řady zpřesnit, protože se jedná o hodnoty, které byly získány (například naměřeny, nebo vyžádány od brokera) v nějakém čase t. Řada tedy může být definována jako: [32]

$$a_{1t_1}, a_{2t_2}, \dots, a_{nt_n}; \text{kde } n \in \mathbb{N} \text{ a } t \in \mathbb{N}$$
$$\text{Neboli} \qquad (2.3.1)$$
$$a_t, \text{taková, že } t \in 1, 2, \dots, n$$
$$\text{Hodnota } t \text{ znamená že } a \text{ bylo zaznamenáno v čase } t$$

Takováto řada se nazývá řadou časovou.
Oproti obecným číselným řadám, mají časové řady některé další vlastnosti:

Řady mohou být buď: [32]

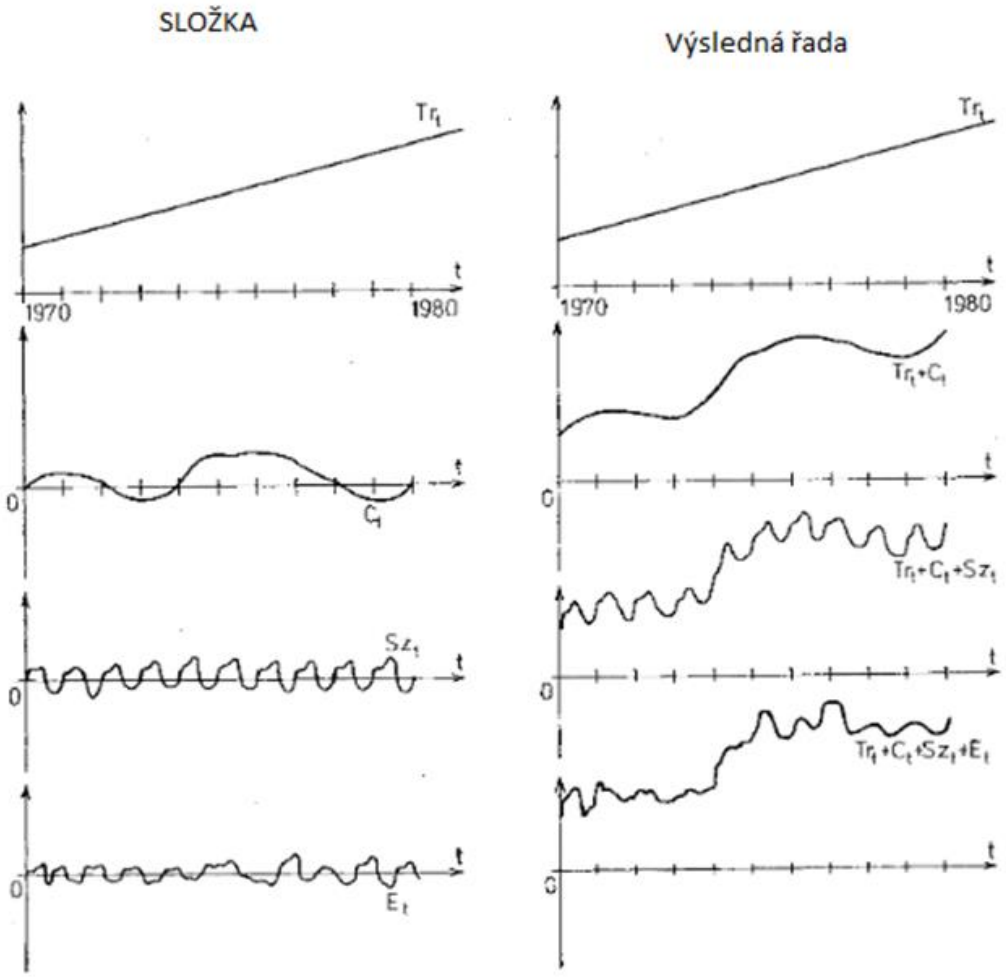
- **Okamžikové** – jedná se o řady znázorňující hodnotu ukazatele, právě v daném okamžiku *t*. Tyto řady by například odpovídaly aktuálním hodnotám ceny určité komodity, jiným příkladem může být aktuální rychlost vozidla.
- **Intervalové** – Tyto řady nepracují pouze s aktuální hodnotou v čase t, ale jejich hodnota je vázána k danému sledovanému intervalu (například aktuální spotřeba vozidla, což je hodnota závislá na velmi malém Δt). [31]

Časová řada obsahuje 4 složky, které mohou být získány dekompozicí časové řady: [32]

1. **Trendová složka (Tr_t)** – neboli zkráceně jen trend. Jedná se o obecný vývoj řady za určité období. Trend může být rostoucí nebo klesající. Časová řada může být i bez trendu.
2. **Sezónní složka (Sz_t)** – tato složka představuje pravidelně se opakující odchylku od trendu řady. Perioda opakování musí být menší, než je sledované období řady. Sezónní složka se nejvýrazněji projevuje v časových řadách, které jsou závislé na ročním obdobím. [35]
3. **Cyklická složka (C_t)** – na rozdíl od sezónní složky má cyklická složka periodu delší než je sledované období, zpravidla delší než jeden rok. Jednotlivé cykly pokaždé ovlivní trend řady, ale nemusí jej ovlivňovat stejně. V extrémních případech může cyklická složka připomínat náhodně přidanou hodnotu jedenkrát za více než rok. [34]
4. **Náhodná složka (E_t)** – neboli stochastická, představuje nahodilé a nesystematické výkyvy hodnot. Do jisté míry se dá náhodná odchylka charakterizovat velikostí rozptylu nad klouzavým průměrem dané řady.

Jednotlivé složky pak mohou být zapsány v aditivním, nebo multiplikativním tvaru podle typu dekompozice: [32]

- Aditivní dekompozice: $y_t = Tr_t + Sz_t + C_t + E_t$
- Multiplikativní dekompozice: $y_t = Tr_t * Sz_t * C_t * E_t$



Obrázek 2.3.1: Adaptivní dekompozice časové řady[1]

Predikce časových řad

Při pokusech o predikci časových řad je možné oproti číselným řadám využít některé další postupy. Mezi nejznámější pak patří dekompoziční metoda a sezónní diference.

Dekompoziční metoda

Pomocí dekompozice je časová řada rozdělena na trendovou, sezónní, cyklickou a náhodnou složku. [33] To slouží k lepšímu pochopení řady jako takové. K dekompozici se mohou využít diference, sezónní diference, kumulativní součty, a jiné prvky, které jsou schopné oddělit jednotlivé složky. [32]

Sezónní diference

Při aplikaci této metody, se analytik snaží z řady odstranit sezónní vlivy. Například v burzovních datech se sezónní vlivy vyskytují hlavně u farmářských plodin, ale i u akcií (zájem o získání akcií se mění vzhledem k blížícím se kvartálům a jiným firemním milníkům). [7] Při sezónní diferenci se získává rozdíl mezi okamžiky vzdálenými o celistvý násobek délky sezonní periody. Takže z hodnoty ze září aktuálního roku, by byla odečtena hodnota ze stejného dne roku minulého. Diference tak

[1] Převzato z [32]

vyjadřuje velikost změny, která nastala mezi dvěma okamžiky. [32] Je-li změna kladná, řada v časovém okamžiku roste a naopak při záporné je řada klesající.

K odhalení, či odstranění sezonních vlivů mohou být využity principy vyhlazování řad. Klouzavý průměr by pak měl velikost sezónní periody.

Vektorové řady

Jedná se o specifický případ číselných řad, kdy je každá hodnota řady doplněna o n-tici (nebo vektor konstantní délky). První hodnota takovéto n-tice (či vektoru o konstantní délce) odpovídá původní hodnotě číselné řady a další hodnoty představují doplňující informace, které pomáhají řadu specifikovat. [7] Vektorové řady jsou dány jako:

$$(a_1, b_{11}, \dots, b_{1m}), (a_2, b_{21}, \dots, b_{2m}), \dots, (a_n, b_{n1}, \dots, b_{nm}); \text{ kde} \quad (2.3.2)$$
$n \in \mathbb{N}; m \in \mathbb{N};$
a_i je hodnotou číselné čady pro i od 0 po n
$b_1, \dots, b_m$ jsou doplnující hodnoty vektoru číselné řady

S vektorovými řadami je možné provádět veškeré operace jako s číselnými řadami. Při těchto operacích je jako hodnota číselné řady využita hodnota vektoru a_i. Operace však mohou být rozšířeny o práci i doplňujícími hodnotami, což se hodí obzvláště pro lepší pochopení a predikci daných řad.

Cenové řady komoditních trhů

Řady komoditních trhů kombinují vlastnosti všech doposud zmíněných řad. Jedná se tedy o časové řady, které se zaznamenávají od brokera v určitý časový okamžik a zároveň je každá hodnota této řady reprezentována vektorem několika dalších hodnot. [7] Mezi tyto hodnoty mohou patřit například počty obchodů, které byly na dané cenové hladině provedeny (tzv. volume), nebo pozice nejbližších hodnot bit a ask, atd. Cenové řady komoditních trhů jsou tedy definovány jako:

$$(a_t, b_1, \dots, b_m)_t \text{ taková, že } t \in 1, 2, \dots, n \text{ a } m \in \mathbb{N} \quad (2.3.3)$$

Jelikož se autor práce rozhodl metody predikce aplikovat nad burzovními daty, bude se práce zaměřovat na cenové řady. Pod tímto pojmem jsou myšleny cenové řady komoditních trhů neboli cenové řady vývoje burzovní ceny.

Specifické vlastnosti cenových řad

Veškeré řady, které zobrazují historii vývoje ceny nějaké komodity, jsou intervalové, protože burzovní domy poskytují data pro svíčkové grafy (vysvětleno později v podkapitole 2.4 Vizualizace číselných řad), které agregují vývoj ceny za období Δt, čili za dobu mezi jednotlivými dotazy na aktuální cenu

Pro zpracování cenových řad je důležitá sezónní diference a nejvíce získání trendové složky.[1] Pokud časové řady obsahují trend, jsou spekulanty vyhledávány pro jejich snadnější možnost předpovědi vývoje a možný profit. Cenové řady s rostoucím trendem jsou spekulanty označovány jako býčí trhy (z anglického bull), naopak klesající jsou spekulanty označovány jako medvědí trhy

[1] Dle serveru Finančník, <http://www.financnik.cz>

(z anglického bear). Dále je možné, že je řada bez trendu. Pro burzovní spekulanty se tyto řady projevují tzv. chop pohyb a jsou velmi nezajímavé, neboť neposkytují možný prostor pro profit.[1]

Pro predikci cenových řad se využívají metody fundamentální a technické analýzy (popsané v kapitole 2.6 Analýzy komodit), zohledňující znalosti a specifika obchodů.

2.4 Vizualizace číselných řad

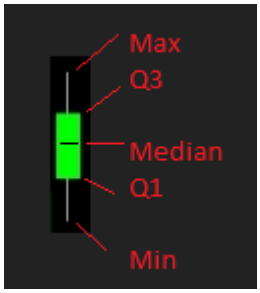
Nejpoužívanějším způsobem pro vizualizaci hodnot číselné řady jsou grafy. [13] Řady se ale mohou velmi rychle měnit a kolísat a pak standartní grafy nemusí být vhodné. Aby se s grafy analytikům dobře pracovalo, musí být jejich vizualizace přehledná. Je vhodné nějakým způsobem dané kolísání „*učesat*". K tomu by se daly využít například klouzavé průměry, je ovšem vhodné zobrazit daná data vždy se stejným časovým měřítkem, neboť někdy se hodnota během jedné minuty změní pětkrát, jindy jen jednou (záleží na frekvenci a pravidelnosti zaznamenávání). Kdyby byla časová osa grafu pokaždé jinak dlouhá, byly by grafy pro člověka velmi nepřehledné (navíc by tyto grafy byly velmi široké, například s časovou osou po vteřinách, a špatně by se analytikům zobrazovaly). [4] Je nutné data v dané minutě (nebo jiné šířce časového okna) nějakým způsobem agregovat. Samotný průměr hodnot v rámci intervalu by skryl maxima, minima a i nejvyšší koncentraci hodnot a proto se nevyužívá. Místo něj je vhodné využívat tzv. krabicové grafy. [13]

Krabicový graf

Tento typ grafu je velmi vhodný, protože zachovává informaci o tom, kde se hodnoty řady nejvíce pohybovaly a právě tyto informace jsou pro analytiky velmi cenné. K tomu se hodí využití kvantilů a z nich se využívají právě kvartily. [13]

Krabicový graf (tzv. boxplot) je pětice K=(min, Q_1, M, Q_3, max), kde:

- *Min* je minimální hodnota rozptylu.
- Q_1 je první kvartil
- *M* je medián daného rozptylu
- Q_3 je třetí kvartil
- *Max* je maximální hodnota rozptylu.



Obrázek 2.4.1: Krabicový graf

Je vhodné si uvědomit, že rozdíl $Q_1 - Q_3$ značí jak široký je interval, ve kterém se nachází 50% všech hodnot. Tento interval se nazývá mezikvartilová vzdálenost a označuje se IQR. [13]

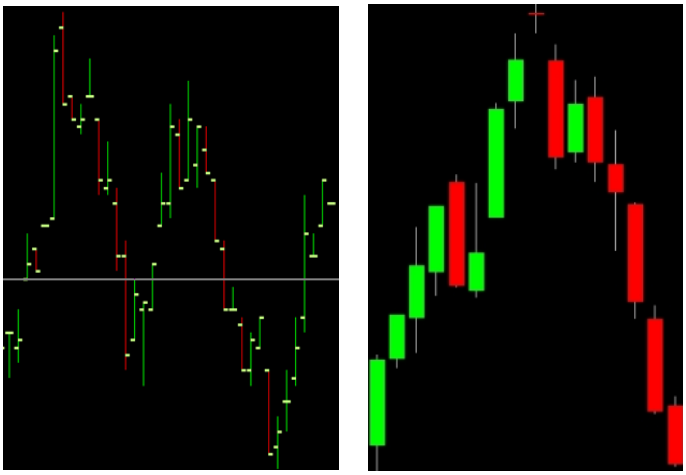
Právě využití mediánu pro získání střední hodnoty může být problematické, neboť medián je holistickou funkcí. Při využití příliš podrobných dat (se záznamem několikrát do vteřiny) a zároveň velmi širokým časovým měřítkem (například jeden krabicový dat popisující celý týden), mohly by počet prvku pro zpracování nepříjemně ovlivnit dobu výpočtu a zobrazování dat by nemuselo být uživatelsky plynulé.

Je také možné krabicové grafy nejrůzněji modifikovat, aby více sloužili potřebám analytika. Jednou z častých možností je místo mediánu využít jinou funkci. Typické burzovní grafy také modifikují krabicové grafy na takzvané svíčkové grafy. [4] [7]

[1] Dle serveru FXstreet, <http://www.fxstreet.cz/technicka-analyza-indikatory-sledujici-trend.html>

Svíčkový graf

Jedná se o nejčastěji využívanou verzi krabicového grafu na burze. Svíčkové grafy již neobsahují Q_1 a Q_3, ale aktuální cena, kterou komodita měla v době zahájení sledovaného okna (například pro 5 minutový graf). [7] Této hodnotě se říká open, a značí, na jaké hodnotě cena komodity začínala. Druhou hodnotou je close, což je cena při které bylo časové okno uzavřeno. Pokud je hodnota close vyšší než hodnota open, je interval považován za rostoucí (bez ohledu na min./max. vývoj). Je-li close nižší, pak je interval považován za klesající. [4] Dle programu a vlastních preferencí (většina programů se dá plně přizpůsobit) se pak rostoucí intervaly zobrazují zeleně a klesající červeně.[1] Takto zobrazené grafy se nazývají v burzovním světě grafy svíčkovými. [7] Jejich zobrazování se v průběhu času mění a programy je umožnují přizpůsobovat (některé svíčkové grafy znázorňují i střední hodnotu (mediánem, nebo některým průměrem) jiné ne.



Obrázek 2.4.2: Starý (vlevo) a nový (vpravo) vzhled svíčkového grafu

Vlevo se nachází starší vzhled svíčkového grafu, který je standartním pro program Ninja Trader[2] a vpravo novější vzhled svíčkového grafu, který je standartním pro program Sierra Chart.

2.5 Cenové řady a obchodní systém

Pokud by se práce soustředila jen na izolovanou cenovou řadu, unikal by kontext, jak tato řada vznikla a proč se nějak vyvíjí. Nebylo by pak možné využívat principy, které umožnují predikci zpřesnit. Výsledný program by sice byl obecný (pracoval by jen s řadou bez ohledu na sémantickou informaci), ale jeho výsledky by byly nepoužitelné. Zvyšující se tendence číselné řady, má nějaký důvod, a po pochopení prostředí, které se vyskytuje v pozadí tvorby ceny, bude snazší najít i logický důsledek jednotlivých změn.

Existují metody na predikce vývoje číselných řad, ale teprve s pochopením ostatních náležitostí, mohou být využity heuristiky, které dané metody zpřesní. Jelikož budou zkoumány cenové řady na burze, je vhodné pochopit jací účastníci se na ni vyskytují a jaké jsou jejich záměry a strategie. Proto práce vysvětluje některé prvky, které jsou na burze důležité a usnadní tak pochopit pozdější předpoklady.

[1] Sierra Chart, <http://www.sierrachart.com>
[2] Ninja Trader, <http://www.ninjatrader.com>

Cenové řady na burze sice někomu mohou připadat zcela náhodné, avšak nejsou. I burzovní svět má svá pravidla, kterými se řídí (problém však je dopředu odhalit, jaké pravidlo, či jev jej zrovna ovlivňuje, a proto je daná předpověď, tak nestálá a plná nečekaných chyb). [1] Právě nejrůznější postupy, jak odhalit, co se v řadě v budoucnu objeví, vedou k mnoha a mnoha vzorům, které s určitou pravděpodobností vývoj predikují. Jednotlivé vzory a ukazatele (například náhlá změna počtu obchodů za krátký čas) pak slouží k vytvoření základu pro obchodní systém. [11] Kompletní obchodní systém, však neobsahuje pouze pravidla, dle jakých se predikuje cenový vývoj, ale i mnohá další doporučení (například kolik transakcí má být provedeno za den, jaká je maximální únosná denní ztráta, nebo cílové zisky). [12] I když jednotlivé vzory tvoří nezbytný základ, celistvý a použitelný obchodní systém musí zohledňovat i psychologii a finanční možnosti toho, kdo jej bude používat.

Protože tato práce nemá sloužit jako příručka, jak se stát obchodníkem, ale má nastínit pohled, jakým způsobem odhalovat ony vzory pro predikci číselné řady, bude pojmem obchodní systém myšlen právě soubor vzorů, které se k predikci používají (bez dalších doprovodných pravidel jak obchodovat). Množství provedených obchodů a řízení rizika, ať si každý obchodník, spekulant, či analytik nastaví podle vlastních možností.

Tato práce rozhodně nemůže nahradit rozsáhlé knihy za tisíce korun, ani nejrůznější kurzy, které stojí ještě více. Nejedná se tedy o žádnou příručku obsahující manuál, jehož následování dokáže nahradit cit a léta zkušeností. Není v silách autora ani v rozsahu této práce, aby po jejím přečtení, byl ze čtenáře zkušený burzovní spekulant, který si bez obtíží mnohonásobně zvýší svůj vklad (ostatně tento cíl, nedokáží naplnit ani velmi drahé publikace a kurzy, neboť většina účastníků končí ve ztrátě). Po přečtení všech kapitol se ze čtenáře žádný obchodník nestane! I mnoho erudovanějších publikací raději tvrdí, že slouží primárně k teoretickým studijním účelům.[1] Zejména proto, že bez ohledu na kvalitu vybudovaného obchodního systému, finančního managementu, či predikci vývoje ceny, stále se jedná o ekonomickou činnost, která velmi blízce připomíná hazard.[2] Z tohoto důvodu není vhodné jakéhokoliv čtenáře v hazardu podporovat, aniž by předtím nebyl varován.

Dění na burze se neustále zrychluje. Zmiňme dobu, kdy se akcie držely po několik týdnů, protože nebylo možné najít dostatek nakupujících/prodávajících protistran. Přes doby, kdy díky novinám, rychlejšímu šíření informací, růstu obchodního světa a jeho centralizaci, došlo na obchodování v rámci jednotek dní. Později se začaly objevovat takzvaní intra denní obchodníci, což byli lidé, kteří obchodovali jen v rámci jednoho dne na hodinových grafech. Vývoj burzy pokračoval, až do dnešní doby kdy intra denní obchodníci využívají minutové až pětiminutové grafy, či obchodní systémy, které už spolu soupeří v jednotkách menších než sekundy. [10] Jak se se vyvíjejí možnosti tohoto odvětví, nutně se musí vyvíjet i obchodní systémy a strategie lidí, kteří se v něm pohybují. Všechny tyto systémy (ať už obchodní nebo strojové) musejí pracovat s velkým množstvím dat, které se snaží analyzovat, aby dokázali předpovědět budoucí trend trhu a dokázaly tak své rozhodnutí proměnit v profit.

2.6 Analýzy komodit

Ve světě ekonomiky je vše provázáno se vším. Příchod ničivé katastrofy rozpoutá velké množství řetězových reakcí, které ve svém důsledku ovlivňují výslednou cenu. Po přírodní katastrofě, bude poptávka po surovinách, které slouží k rychlému obnovení obydlí a životního minima. Tato vyšší poptávka způsobí zdražení stavebního materiálu nejen v dané oblasti. V poškozené oblasti bude poptávka po stavebním materiálu a stavebních pracích. Na rozdíl tomu v ostatních oblastech zvýšená cena materiálu posune stavební práce do útlumu a majitelé stavebních firem, budou muset provádět

[1] Dle Investopedia, <http://www.investopedia.com>
[2] Dle Finančník, <http://www.financnik.cz>

nejrůznější opatření, aby dané období přežili. Každá událost ve světě a reakce lidí na ni, vyvolává další a další řetězec událostí, které pak ovlivňují poptávku nebo nabídku surovin a služeb a tím i jejich cenu. Tyto fundamenty, jejich hledání a pochopení tvoří fundamentální analýzu. [6]

I v burzovním světě se setkáme s fundamentální analýzou. Jejím protikladem je analýza technická. Některé publikace pak ještě uvádějí analýzu racionální, která vychází z propojení obou předchozích analýz.

Fundamentální analýza

Jedná se o analýzu prostředí jako takového. Fundamentální analýza se snaží zachytit nejrůznější informace, které ovlivňují dané odvětví a podle toho pak určit směr vývoje ceny. Pokud se má například otevřít nová farma se skotem, dá se v dané oblasti očekávat zvýšení ceny krmiva (například kukuřičné granule), naopak se pravděpodobně sníží cena hovězího masa. Dané proměnné je nutné nalézt a k tomu jsou důležité komplexní znalosti a bohatá praxe. [9]

Pokud se obchodník snaží využívat fundamentální analýzy, je nutné, aby využíval znalostí daného odvětví. Pokud by si nedokázal stanovit, které události ovlivňují cenu zboží, byla by mu veškerá práce spojená s danou analýzou zbytečná. Při fundamentální analýze, je nejcennějším prvkem informace. Je nutné sledovat světová média a mít dostatek kontaktů s informovanými lidmi. Když se podaří obchodníkovi dostat k informacím dříve než ostatním, má značný náskok a může podle toho upravovat svoji strategii, aby utržil zisk. [7]

Některé komodity jsou tak propletené, že se růst či pád přímo odráží v trhu komodity jiné. Jindy se jedná o trh příliš obecný a fundamentální analýza na něm nemá příliš dobré výsledky. Burzovní svět je velmi rozsáhlý, a poskytuje mnoho možností. Fundamentální analýza se vždy specifikuje na jeden typ komodity, a proto nepřináší žádná obecná řešení. Asi nejobecnějším prvkem fundamentální analýzy (u kterého navíc jde velmi pěkně poznat, co obchodníkovi analýza vlastně přináší) je roční cyklus komodity. U některý komodit, jako je například ropa, roční cyklus neexistuje, protože ropa se těží a spotřebovává stále. Naproti tomu zemědělské plodiny, jsou přímo ovlivňovány ročním obdobím a předpovědí počasí. Pokud obchodník například ví, kdy se sklidí nové obilí, ví také, kdy se má zbavit zásob starého zrna. [3]

Fundamentální analýza byla první způsobem, který se v obchodním světě využíval. Základní prvky fundamentální analýzy, pak byly využívány ještě dříve, než burza vůbec vznikla. [9]

Při fundamentální analýze, je důležité rozumět informacím, které se nějak k danému odvětví váží. [5] To je pro strojové zpracování velmi náročné a ještě dnes je vyžadována korekce od člověka. S rozvojem počítačové schopností porozumění přirozeného jazyka, se však otevírají nové možnosti pro strojové zpracování. Zatím je však pro počítače složité analyzovat celý článek, některé slovo, může například sémanticky měnit smysl celé myšlenky. Hrozí tak, že by si stroj myslel, že továrna se staví, a přitom se v článku mohlo o postavení teprve uvažovat, nebo hůř, článek by se mohl zmiňovat o tom, že se továrna rozhodně nepostaví. Automatické obchodní systémy, nad fundamentální analýzou zatím moc nevznikají. [1] Pokud ano, jedná se pouze o doprovodné systémy, které mají pro obchodníka jen poradní hlas.

Tento druh analýzy má další velmi výrazný problém. Pro fundamentální analýzu je velmi složité sehnat veškerá potřebná data, která aktuálně trh ovlivňují. V případě historického vývoje fundamentálních prvků, je daný úkol téměř neřešitelný (je možné získat hlavní události, které byly ve zprávách, jedná se ale jen o zlomek všech používaných a potřebných dat). [9] Navíc je jejich zpracování velmi složité, a jdou vyžadovány komplexní znalosti z daného odvětví, které je velmi obtížné získat. [1] Vzhledem k rozsáhlosti fundamentální analýzy, nedostatku dat a vzhledem k velmi nepřívětivému prostředí pro strojové učení, nebude fundamentální analýza v dané práci využita. Fundamentální

analýza by mohla být využívána maximálně jako rozšíření systému o zpřesňující faktor v pozdějších úpravách systému.

Technická analýza

Zatímco fundamentální analýza vychází ze sběru informací, které ovlivňují dané odvětví, technická analýza pracuje výhradně s grafem cenového vývoje. Při technické analýze je hlavním cílem, najít v grafech určité vzory (z anglického paterns), které budou s určitou pravděpodobností předpovídat následující vývoj. [4]

Výhodou technické analýzy je její jednoduchost. K provedení analýzy stačí pouze vývoj cenového grafu (v posledních dobách se přidalo několik dalších ukazatelů), což je mnohem snazší, než sledovat všechna světová media v různých jazycích. Technická analýza tak otevřela burzovní svět mnoha účastníkům, kteří neměli prostor na získání rozsáhlých znalostí pro fundamentální analýzu.

Další nespornou výhodou je možnost pracovat s historií. Protože si stačí zaznamenávat pouze aktuální cenu, existuje mnoho společností, které se zabývají jen sběrem a distribucí těchto dat (například IQFeed).[1]

Technická analýza začala vznikat v době vytvoření prvních pitových místností (zde se odehrávaly veškeré obchody na burze, viz příloha). Jelikož informace byly jednoduše dostupné (pro toho kdo se na pitu nacházel), bylo možné sledovat vývoj ceny a zaznamenávat jej. [1] Vznikli odborníci na grafy nejrůznějších komodit, tzv. Chartists. Tito lidé hledaly v grafech vizuální vzory. Vznikly tak vzory jako například „hlava a ramena“ (z anglického head and shoulders), „dvojité dno“ (double bottom), „dvojitý vrchol“ (double top) a nejrůznější „námořní vlajky“ (wedges). [5] Dané vzory byly z počátku dostatečně efektivní, i když jejich efektivitu mnozí fundamentální analytici velmi zpochybňovali. Problémem daných vizuálních vzorů je fakt, že vznikají i na grafech, které nemají nic společného s vývojem ceny. [8] Pokud by se vytvořil graf počtu úmrtí při autonehodách (nebo hodu mincí), vzory jako head and shoulders by v nich vznikaly také, ovšem jen stěží by dokázaly předpovědět vývoj následujících dní.

Obrázek 2.6.1: Head and shoulders graph patern

Obrázek 2.6.2: Double bottom graph patern

[1] IQ Feed, <http://www.iqfeed.com>

S rozvojem fundamentální analýzy se přidávaly další ukazatele, jako například volume. Tím se zpřesňovaly jednotlivé analýzy a jejich predikce. Nad cenami se začali využívat nejrůznější funkce (například klouzavý průměr, exponenciální průměr, low či high dne, R&S rezistence, CCI) a tím se technická analýza vyrovnala analýze fundamentální. [8]

Analytici cenových grafů, nejčastěji pracují právě s klouzavým průměrem z předchozích hodnot, kdy určují vhodné obchodní situace dle křížení klouzavých průměrů různé šířky. [7] Například jeden klouzavý průměr o šířce jednoho měsíce, pro určení hlavního trendu trhu. Druhý denní klouzavý průměr pro potvrzení že aktuální trh odpovídá hlavnímu trendu a pak poslední hodinový klouzavý průměr, který indikuje signály pro otevření kontraktu.

Velmi oblíbený je pak indikátor EMA (Exponential Moving Average), který odpovídá klouzavému průměru s tím, že novější data mají vyšší váhu (nezaměňovat s klouzavým Exponenciální průměr z kapitoly 2.4), EMA o šířce N (velikost zpracovaného okna) je dán vztahem:[1]

$$EMA_t = Price_t * k + EMA_{t-1} * (1 - k) \qquad (2.6.1)$$

$$k = \frac{2}{N + 1}$$

Dalším typickým příkladem technické analýzy je pak hledání dostatečné míry korelace. Vzájemná podobnost se pak nejvýrazněji objevuje u komodit, jež se nějakým způsobem ovlivňují. Například ocel a výrobky z ní, pokud se zvedne hodnota oceli, musí podražit i cena produktů. [3]

Korelace mezi jednotlivými řadami, je velmi užitečnou informací. Jednak se podle ní dají dostavět chybějící hodnoty v grafech cen a zároveň přináší jisté vodítko k fundamentální závislosti. Velmi silně korelují například cenové grafy pro sójové boby a sójovou drť, nebo kukuřice a obilí. I mnoho brokerských domů nabízí takzvané spreadové obchodování. [7] Jedná se o speciální typ obchodu, kdy nejde o pohyb ceny jako takové, ale o pohyb korelujících cen. Obchodník tak jednu komoditu nakoupí, a druhou korelující komoditu prodá. [9] Pokud jdou obě komodity cenou nahoru, tak na jednom kontraktu utrží zisk a na druhém kontraktu utrží ztrátu ve stejné míře. Jeho opravdovým ziskem nebo ztrátou, tak může být pouze případ, kdy spolu korelující komodity přestanou být v korelaci. Tyto stavy se sice dějí relativně často, ale jen v řádu několika málo bodů (hodnota bývá jen pár dolarů), riskované částky jsou proto velmi nízké. [7] Při spreadovém obchodování se jedná o systém, který sází na dvě komodity zároveň, a v procesu učení této práce nebude zahrnut. Podobnost řad je sice využita pro získání chybějících hodnot, nebo pro zesílení a zpřesnění nějakého vzoru, ne však pro hledání obchodního systému nad spready.

Obrázek 2.6.3: Příklad korelujících trhů v pořadí: ES (Eversorce Energy), YM (Dow Industrial Index Mini Futures), NQ (E-mini NASDAQ 100 Futures Quotes)

[1] Dle FXstreet, <http://www.fxstreet.cz/technicka-analyza-indikatory-sledujici-trend.html>

Možná provázanost fundamentální a technické analýzy, není na první pohled vidět, ale opak je pravdou. Pokud nastane událost, která ovlivní cenu (což spadá pod fundamentální analýzu), tak z vývoje ceny je schopen analytik získat v technické analýze jistou informaci o tom, že nastala změna. [1] Technická analýza není schopná zpětně dohledat, o jakou událost se jednalo, ale pokud se jedná o opakující se fundament, je pro ni získán vzor i z technické analýzy. [6] Při podrobném rozboru výsledků technické analýzy (a pokud má analytik dostatek zkušeností v daném oboru), může být rekonstrukce událostí relativně přesná.

Vzhledem k jednoduchosti technické analýzy, v dnešních dobách většina účastníků na burze využívá právě dané techniky. Jelikož je technická analýza vhodná pro počítačové zpracování (získání dostatečného vzorku dat není překážkou) zaměřuje se tato práce právě na tento způsob analýzy.

Racionální analýza

Jak již bylo nastíněno v úvodu, jedná se o metodu, jenž kombinuje prvky předchozích analýz. Stejně jako fundamentální analýza i tato metoda vyžaduje rozsáhlé znalosti z oboru a mnoho let praxe, aby se dalo rozhodnout o tom, které prvky jsou, a které nejsou důležité. Racionální analýza tak páruje výsledky technické a fundamentální analýzy, které jsou navíc velmi provázané. Pokud data k fundamentální analýze nejsou dostupná, je nutné je rekonstruovat za použití zkušeností a znalostí daného analytika, který racionální analýzu provádí. [2]

Pomocí technické analýzy může například pro určitou komoditu vyjít středa jako nejvýnosnější den v týdnu. Kdyby se analytik, zaměřil jen na tuto skutečnost, upravil by obchodní systém tak, aby obchodoval převážně ve středu. Při racionální analýze, je však nutné odhalit fundamenty v pozadí (jednotlivé fundamentální události se projevují na vývoji ceny, a proto se projevují i v technické analýze). [9] Pokud se dané fundamenty nepodaří odhalit, je daný výsledek technické analýzy ignorován a považován jen za statistickou odchylku.

V případě, že by středa byla nejvhodnějším dnem pro obchodování za minulých 7 let, nemusela by být tato informace pravdivá za posledních 50 let. Pokud je například středeční růst způsobován tím, že ve středu vláda USA vydává nařízení omezující, či povolující růst trhu, jedná se o odhalený fundament a vzor z technické analýzy by byl přijat. [8]

Tato práce se racionální analýzou zabývat nebude, ale je vhodné ji uvést pro úplnost.

2.7 Kategorie dat

V burzovním světě se obchoduje s nejrůznějším zbožím (tzv. komodity), dále se zde vyskytují akciové balíčky (tzv. indexy), či balíčky s kurzy měn (tzv. rate) a mnohé další. [1] Všechny tyto obchodní položky jsou reprezentovány cenovými řadami.

Cenové řady jsou tím nejdůležitějším, co obchodník od svého brokera může získat, aby viděl, jaká je aktuální cena a jestli se pohybuje očekávaným směrem nebo ne. [9] Informace pro fundamentální analýzu se na burze nenacházejí a je pro ni nutné sledovat externí dění, což je velmi náročné. [8] Když se mluví o historických datech, je řeč právě o cenových řadách od brokera pro danou obchodovatelnou položku.

I obchodování s jedinou komoditou má však své strategie a jednou z odlišností těchto strategií je časové měřítko, dle kterého analytik cenovou řadu (a její graf) vyhodnocuje. Existují různí lidé s různými strategiemi a mentalitou, kteří využívají různé časové intervaly. V burzovním prostředí tak vznikly dva hlavní směry obchodování a s nimi i míra abstrakce nad cenovými řadami.

Poziční obchodování

Poziční obchodování a s ním spojená analýza se zabývá nákupem a prodejem komodit na dobu delší než jeden den. Daný kontrakt (smlouva o nákupu, či prodeji) je držen přes noc. Poziční obchodování má výhodu v tom, že pro analýzu nejsou potřeba drahá (nebo alespoň hůře sehnatelná) data s krátkými časovými intervaly. Například data s minutovými intervaly jsou pro danou analýzu nepotřebná a svou četností by zatěžovala systém. Jelikož analytik, využívá pouze jednu hodnotu za každý den (nebo jednu za týden, či měsíc), pracuje s relativně krátkými cenovými řadami.[1] Výhodou těchto dat je i skutečnost, že zatímco se minutové hodnoty mohou u různých brokerů drobně lišit, denní hodnoty komodit budou stejné. I cena historických dat je mnohonásobně nižší, protože hodnota komodity bývá zveřejňována burzou samotnou na konci každého obchodního dne.[2] Většina historických dat je tak pro analytika veřejně přístupná.

Nespornou výhodou tohoto obchodování je nízká časová náročnost pro budoucího obchodníka, jelikož mu stačí sledovat grafy jen pár minut denně (většinou při zahájení obchodních hodin a při ukončení). Protože následovníci dané strategie nejsou příliš časově vytíženi, většinou se snaží sledovat nejrůznější informační zdroje a provádějí fundamentální analýzu, aby dokázali odhadnout kdy je nejvýhodnější obchody uzavřít. [3]

Analýza však přináší rizika, neboť přes noc se kontrakt nedá uzavřít (nebo velmi špatně a s výrazným zpožděním). Pokud se cena začne vyvíjet nepříznivým směrem, není spekulant schopen omezit ztráty. Poziční obchodování je využíváno méně často, jelikož je nutné disponovat větším kapitálem, jenž umožní obchodníkovi ustát nepříznivé období ztrát.[4] Při využití pozičního obchodování se pracuje s timeframe na celé dny, týdny či dokonce měsíce. [3] Takto dlouhé držení kontraktů, dovoluje utržit vyšší zisk, který by se v rámci kratších intervalů (minuty, hodiny) nepodařilo vytvořit. Možnost vyššího zisku je ekvivalentní s možností vyšší ztráty, což může být psychicky náročné. [12] Jako příklad může být uvedeno, že kdyby se cena pohybovala vzhůru o deset dolarů každou hodinu, tak po týdnu by rozdíl činil 1 680 dolarů. Obchodník, který by obchodoval každou hodinu, by musel provést o 167 obchodů více, aby utržil stejný zisk. Poziční obchodování využívá malé množství příležitostí k uzavření obchodu, ale s mnohem vyšší efektivitou. Nespornou nevýhodou je pak případ, kdy se hodnota komodity pohybuje proti očekávání obchodníka.

Intradenní obchodování

Jedná se o strategii nákupů a prodejů kontraktů v rámci jediného dne. Kontrakt nikdy není držen přes noc.[5] Během kratší doby se cena kontraktu nedokáže výrazně posunout, a z toho důvodu následování dané strategie nevyžaduje vysoký kapitál.

Jelikož na jeden provedený kontrakt jsou zisky a ztráty relativně nízké, musí spekulanti uzavírat mnoho kontraktů. Vzhledem k množství kontraktů, mohou (na rozdíl od následovníků poziční strategie) spoléhat pouze na technickou analýzu.[6] Fundamentální analýza by se pro krátké intervaly (například 3 minut) nedala efektivně využít. Za takto krátký časový úsek by byl problém získat a zpracovat potřebný fundament. Dále je problematické odhadnout, kdy se fundament v cenové řadě projeví.

---

[1] Dle Investopedia, <http://www.investopedia.com>
[2] Dle Finančník, <http://www.financnik.cz>
[3] Dle Finančník, <http://www.financnik.cz>
[4] FXstreet, <http://www.fxstreet.cz>
[5] Dle Investopedia, <http://www.investopedia.com>
[6] Dle Finančník, <http://www.financnik.cz>

Jelikož je fundamentální analýza pro intradenní obchodování velmi neefektivní, spoléhají následovníci dané strategie na statistiku a pravděpodobnost. [3] Za předpokladu, že jim technická analýza vytvoří systém s predikcí úspěšných obchodů v 60%, je pro ně jednodušší bezmyšlenkovitě využít každou odhalenou příležitost a s jistou statistickou pravděpodobností se jim z dlouhodobého hlediska musí dařit.

Jak napovídá již předchozí odstavec, je intradenní obchodování pomocí technické analýzy tím nejvhodnějším stylem, pro automatické systémy. [10] Tak jako v rámci denních analýz mohou být rozdílné pohledy na sledovaná časová okna (dny/týdny/měsíce), tak se liší i míra abstrakce pro intradenní analýzy. Spekulanti využívají různé velikosti timeframe od několika hodin až po jednotky minut. Plně automatické systémy pak mohou využívat časové pohledy ještě mnohem kratší (v jednotkách sekund). Automatické systémy využívají výhodu rychlosti vyhledávání vzorů v cenových řadách. [7] Práce s takto krátkými intervaly má ovšem, také své nevýhody. Jednak je nutné vzory v cenové řadě odhalit co nejrychleji, což častokrát vyžaduje nákladný hardware. Vzhledem k tomu, že v rámci vteřiny se cena nedokáže téměř pohnout, není možné využívat typické smlouvy s brokery, kteří si účtují poplatek, za každý zpracovaný obchod.[1] Tyto poplatky by byly mnohem větší, než utržený zisk, a proto je nutné přejít na speciální paušální smlouvy (kdy je stanoven poplatek za měsíc, bez ohledu na počet provedených obchodů). Jelikož je potřeba informaci o otevření pozice doručit burzovnímu domu rychleji než ostatní spekulanti, je nutné spolupracovat s vhodným brokerem. Někteří brokeři příkazy zpožďují (shlukováním požadavků dohromady), aby snížili vlastní náklady za poplatky burzovnímu domu.[2] Další alternativou je založení vlastní brokerské společnosti, což bohužel vyžaduje ohromný kapitál v řádu desítek milionů dolarů. Tato částka je pro drtivou většinu spekulantů nepřekonatelný problém.

Potřebná data

Jelikož existují různé pohledy na cenové řady, jsou k odlišným analýzám vhodná různá data. Analytik, který se soustředí na denní intervaly, nepotřebuje jemná TICK data, která by zbytečně zahlcovala jeho systém. Naopak pokud je budován automatický systém, pracující nad časovým rámcem kolem několika málo sekund jsou TICK data vhodná. 1 TICK odpovídá nejmenšímu možnému pohybu ceny na burze. Odpovídá tedy těm nejjemnějším datům, která existují. [1]

Burza jako taková s analytikem komunikuje pomocí prostředníka (brokera). Broker se zavazuje, že v případě finančních problémů svého klienta ponese následky za něj, a tím ochrání burzu. [1] Burza je tak stabilní prostředí, ve kterém se nemůže stát, že by kontrakt nebyl naplněn z důvodů náhlé ztráty kapitálu některé strany. Toto je burze zaručeno tím, že každý broker musí disponovat doporučeným množstvím kapitálu, který je v řádů milionů dolarů. Pokud brokerův kapitál klesne pod minimální hranici, burza s ním přestane komunikovat, aby měla jistotu, že nenastane situace, kdy ani broker nebude schopen pokrýt závazky svých klientů.

Burzovní domy poskytují brokerům svá data o cenách za různé poplatky, a čím jemnější pohled na data broker vyžaduje, tím jsou pro něj dražší. Někteří brokeři proto vůbec nemusí být vhodní pro intradenní strategie. Jiní TICK data od burzy vyžadují, ale účtují si pak vyšší poplatky za zprostředkování práce s kontrakty, což se zase nemusí hodit obchodníkům, kteří tyto data nevyžadují. Brokeři si pak mohou a nemusí udržovat vlastní historii těchto dat. Každý broker může svým klientům nabízet různě vzdálený pohled do minulosti (to je ovlivněno jednak kapacitou jeho uložiště, jednak dobou jak dlouho broker působí na trhu). Samotný rozsah historie není nejdůležitějším ukazatelem, protože mnoho brokerů historická data agreguje, aby jim zabírali méně místa na uložišti (broker tak

[1] Dle Interactive Brokers, <https://www.interactivebrokers.com/en/home.php>
[2] Dle Investopedia, <http://www.investopedia.com>

například odebírá od burzy TICK data, ale po týdnu je převádí na minutové (nebo větší) timeframe). Dalším ukazatelem kvality brokera je množství obchodovatelných položek. Někteří brokeři se specializují na nerostné suroviny, jiní naopak na zemědělské plodiny nebo na různá spektra nejžádanějších komodit. Může se proto stát, že pokud chce analytik zpracovávat široké množství komodit, musí svá data získávat od více brokerů.[1]

Existují i společnosti, které se specializují pouze na sběr dat a uchování jejich historie. Tyto společnosti se zaručují za kvalitu dat (v případě přerušení spojení s burzou, schraňují data z jiných zdrojů nebo je doplňují zpětně). Nevýhodou je vysoká cena těchto dat.

IQFeed

Jedná se o společnost s velmi kvalitními daty s téměř plnou historií, což znamená, že mají většinu dostupných komodit od doby, kdy byly na burze zaznamenány. Jimi poskytovaná data jsou velmi kvalitní a ve většině případů i s nejvyšší jemností na 1 TICK.[2]

Mnoho analytiků raději využívá distribuovaná aktuální data od IQFeed, než aby je braly od svého brokera, který může mít pomalejší linku spojení. Největší nevýhodou daných dat je cena, která pro mnoho komodit přesahuje i tisíce dolarů. Cena dat je závislá na míře agregace požadovaných dat, komoditě jako takové a rozsahem historie.

Sierra Chart

Tato společnost se specializuje na vývoj programu, pro zobrazení cenových grafů a nejrůznějších ukazatelů, jakými jsou klouzavé průměry, volume a jiné. Výhodou programu je vysoká přizpůsobivost, aby vyhovoval téměř jakékoliv strategii. Program je plně univerzální a tak je možné jej propojit i s různými zdroji dat, ať už od brokera, nebo od IQFeed.[3] Program bude nastíněn později v kapitole 2.9 Software pro zobrazení dat.

Přestože se u programu předpokládá využití jiného datového zdroje, je společnost Sierra Chart také poskytovatelem historických dat (data jsou však o několik hodin zpožděná a nevyplatí se pro intradenní obchodování). Nejedná se tedy o takzvaná živá data, ale pro technické analýzy postačují. Společnost Sierra Chart však poskytuje 7 letou minutovou historii a poslední 3 roky mají u mnohých komodit dokonce s nejjemnějšími TICK intervaly. Data jsou poskytována za paušální měsíční poplatek, který činí zhruba 500 korun. [4]

Nevýhodou daných dat jsou občasné výpadky v hlavních obchodních hodinách a naprosto nevhodný způsob agregace mimo ně (vlastní zkušenost autora, který s daty pracoval).

---

[1] Dle FXstreet, <http://www.fxstreet.cz> a specifikace mnohých brokerů:
Degiro, <https://www.degiro.cz>
Big Option, <https://www.bigoption.com>
Fidlity, <http://www.fidelity.com>
ETrade, <http://www.etrade.com>
Scot Trade, <https://www.scottrade.com>
Capital One Investing, <http://www.CapitalOneInvesting.com>
Interactive Brokers, <https://www.interactivebrokers.com>

[2] IQ Feed, <http://www.iqfeed.com>

[3] Dle Sierra Chart, <http://www.sierrachart.com>

[4] Dle Sierra Chart, <http://www.sierrachart.com>

Interactive Brokers

Jedná se o brokera komunikujícího přímo s burzou. Poskytovaná data mají jen roční historii a nevýhodou může být absence TICK intervalů. Výhodou pro aktivního spekulanta (který prování obchody) je skutečnost, že data od tohoto brokera získá zdarma.[1]

Ostatní

Je také možnost získávat data od jiných zdrojů. Některá historická data se dají najít na internetu a umožní tak analytikovi ušetřit v jeho inicializační fázi. Je však důležité zvážit, jestli neoficiální data nemohou obsahovat chyby, které by způsobily chybnou analýzu a vytvoření chybného systému.

Zatímco intradenní minutové intervaly, jsou hůře dostupné a častokrát zpoplatněné, denní vývoj hodnot je častokrát zobrazován na nejrůznějších webech zcela zdarma. Přesto je pohodlnější získat jemná data od některé společnosti a aplikovat na ně abstraktní pohled, který daný analytik sám vyžaduje.[2]

Rozhodně se nejedná o kompletní výčet všech poskytovatelů dat. Takový seznam by byl úplně mimo rozsah této práce. Každý spekulant by měl svého brokera vybrat dle svého požadovaného systému. Nemá smysl platit za drahá TICK data, pokud se analýza bude soustředit na denní intervaly a naopak.[3]

2.8 Spojování cenových grafů

Jednou z posledních věcí, které jsou pro cenové grafy na burze důležité, je jejich spojování. Výrazná většina komodit na burze není tvořena jednou cenovou řadou, jako je tomu například u akcií, ale naopak jich obsahují více. Každá komodita, má totiž určené své datum dodání. Na burze se proto vyskytuje ropa, která má termín dodání v září, a pak ropa s termínem dodání v prosinci. Byť se jedná o stejnou surovinu, její cenové grafy se mohou výrazně lišit. Z toho důvodu analytik musí specifikovat konkrétní cenovou řadu vztahující se pouze k určité komoditě a termínu dodání. [7]

Veškeré burzovní komodity mají kódové označení, které se skládá z velkých písmen a číslic. Například komodita Dow Industrial Index Mini Futures má zkratku YM. Každý měsíc v roce má pak také jednopísmennou zkratku, která se k názvu přidá. Takže zářiový termín dodání roku 2016 pro kontrakt YM bude mít zkratku YMU16.

[1] Dle Interactive Brokers, <https://www.interactivebrokers.com/en/home.php>

[2] Dle obchodni specifikace různých brokerů:
Degiro, <https://www.degiro.cz>
Big Option, <https://www.bigoption.com>
Fidlity, <http://www.fidelity.com>
ETrade, <http://www.etrade.com>
Scot Trade, <https://www.scottrade.com>
Capital One Investing, <http://www.CapitalOneInvesting.com>

[3] Dle Investopedia, <http://www.investopedia.com>

Seznam měsíců:[1]

Leden	*January*	*F*	*Červenec*	*July*	*N*
Únor	*February*	*G*	*Srpen*	*August*	*Q*
Březen	*March*	*H*	*Září*	*September*	*U*
Duben	*April*	*J*	*Říjen*	*October*	*V*
Květen	*May*	*K*	*Listopad*	*November*	*X*
Červen	*June*	*N*	*Prosinec*	*December*	*Z*

Tabulka 2.8.1: Označení kontraktních měsíců

V případě již zmíněné ropy, se o dodávku v září mohou zajímat určité společnosti, zatímco jiné mohou mít svojí výrobu nejaktivnější až v prosinci a budou vyžadovat jiný termín dodání. Ropa se však těží a dováží neustále, a proto se cenové grafy budu lišit hlavně nutností komoditu získat. Pokud je cena příliš vysoká, mohou společnosti s dostatečně velkými zásobami na skladě počkat a objednat si komoditu s pozdějším termínem dodání. Naproti tomu například takové obilí se získává sezóně. Proto jsou termíny ihned po sklizni mnohem levnější, než termíny pozdější, ve kterých se projeví náklady na skladování. [1] Jelikož je dnešní trh plně globální, je možné objednat komoditu z jiného koutu světa (kde je zrovna sklizeň), ovšem pak se zde projevují náklady na přepravu. Naopak ihned po sklizni cenu zvedá vyšší poptávka po zboží (levnější cena a možnost předzásobení). Pokud však podnik nakoupí příliš mnoho zásob, například na celý rok až do další sklizně, musí obilí skladovat sám, což mu jistě zvyšuje náklady a navíc musí mít dostatečně velké sklady. Je proto otázkou pro každého obchodníka jak vybalancovat nejlepší dobu a množství pro nákup komodity.

Většina spekulantů (kteří nechtějí danou surovinu zpracovávat, ale jen přeprodat), považují různé cenové řady za jednu. Pohyby v cenových řadách jsou nejvýraznější vždy nejblíže k termínu dodání. V tomto období s komoditou obchoduje nejvíce obchodníků a trh vytváří prostor pro spekulantův zisk. Naopak vzdálené termíny jsou příliš plytké, a zisk by byl nevýrazný.

Většina programů umožňuje nastavení pro zobrazování pouze nejaktuálnějšího termínu dodání. Ve chvíli kdy nastane čas termínu dodání je obchodování s komoditou u konce. Programy automaticky začnou zobrazovat cenové řady komodity k následujícímu termínu dodání. Brokeři dokonce poskytují tuto funkčnost pro své uživatele, kteří vlastní kontrakt přes dané období. [7] Pokud spekulant vlastní určitou komoditu ke konci obchodovatelného období, broker ji automaticky prodá. Tím chrání spekulanta, aby mu nebylo doručeno například 127 tun obilí, což je objem jednoho kontraktu. Po prodeji broker automaticky nakoupí danou komoditu z následujícího termínu dodání. Tomuto procesu se říká překlopení kontraktu (contract rollover). Uživateli je proces přeprodání skryt. Vlastník kontraktu vidí tuto událost jen jako gap v ceně, který se běžně stává (například přes noc). [1] Tento gap naštěstí nebývá příliš velký, protože s blížícím se koncem obchodování většina obchodníků uzavře své obchody a začnou obchodovat v následujícím období. S přesunem většiny obchodníků se i velmi rychle srovnají ceny následujícího a daného období.

Pro učení a vytváření obchodních systémů se vyplatí pracovat s jednou celistvou cenovou řadou, která odpovídá právě nejaktuálnějším dodacím obdobím. Například roční graf obilí je sestaven z 5 cenových grafů, pro různá dodací období, která je nezbytné pospojovat.

[1] Dle Investopedia, <http://www.investopedia.com>

2.9 Software pro zobrazení dat

Práce již zmínila většinu základních problémů, které při zpracování cenových řad na burze vznikají. Od získání dat přes nezbytné úpravy až po vytvoření jednotné cenové řady. Pro každého analytika je však důležité tuto řadu přehledně zobrazit. V kapitole 2.4 Vizualizace číselných řad byly představeny svíčkové grafy. Existuje mnoho způsobů, jak tyto grafy zobrazovat. Například i program Excel od společnosti Microsoft umí tyto grafy vykreslit. Jejich praktické použití je však pro obchodníky poněkud nepohodlné.[1]

V oblasti práce s cenovými řadami by bylo vhodné zmínit dva nejvýznamnější systémy, které jsou velmi používané. Jedním z nich je program Sierra Chart[2]. Tento program není možné zakoupit, ale pouze pronajmout, je tak nutné každý měsíc platit za využívání programu určitý poplatek. Druhým příkladem je program Ninja Trader[3], který je nutné zakoupit a poskytuje doživotní licenci pro danou verzi programu. Nové verze, které by vyžadovali další investici, příliš často nevycházejí a nutné úpravy se řeší jen aktualizacemi systému (zdarma). Zakoupení tohoto programu vyjde zhruba na 20 tisíc, jelikož se autorovi podařilo se s daným programem setkat, poskytuje tato práce základní srovnání.

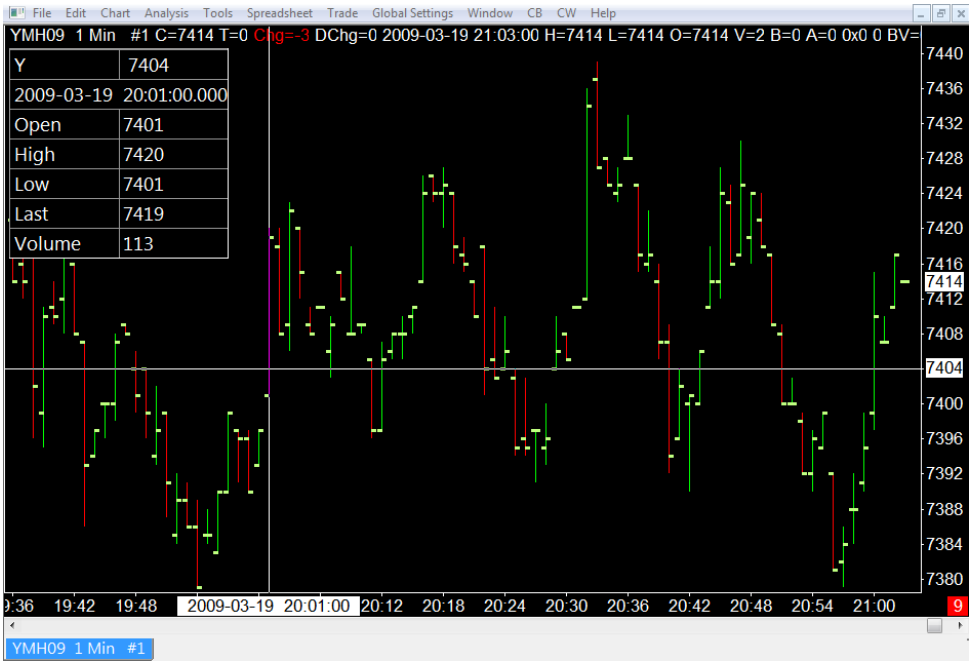
Sierra Chart

Sierra Chart je plně modifikovatelný, což z praktického hlediska znamená, že umožňuje nastavit a ovlivnit téměř všechny vlastnost programu. Nevýhodou však je složité ovládání, kde začínající uživatelé stráví mnoho hodin nastavováním vhodných parametrů. Program naštěstí nabízí exportování celého aktuálního nastavení (včetně toho jak se mají zobrazovat grafy a jejich rozložení na obrazovce) a importovat je do programu na jiném stroji. Spekulanti, kteří poskytují kurzy ke svým obchodním systémům, pak většinou šíří své myšlenky i s tímto nastavením, aby každý uživatel měl práci připravenou tak, jak se s ní setkal na kurzu daného spekulanta.

Jak je vidět na obrázku 2.9.1 základní nastavení programu zobrazuje rostoucí svíčky zeleně a klesající červeně (standartní vžité zobrazení většiny programů pro zobrazování burzovních dat). V záhlaví je stručný popis grafu, zkratka sledovaného trhu a šířka agregovaného časového okna. Kurzor je zobrazen křížem pro snadnou navigaci na časové měřítko a měřítko hodnoty dané komodity. Je nutné ještě dodat, že hodnota cenové řady je v bodech. Každý komodita má pak jiný přepočet bodů na dolary a je nutné brát danou informaci v potaz. V tabulce jsou pak zobrazeny nejčastější hodnoty pro definici aktuálně zvolené svíčky.

Jelikož je program plně modifikovatelný, je možné v něm zobrazovat více pohledů na data zároveň. Příkladem může být vykreslení sloupového grafu pro volume, či nejrůznější indikátory. Program podporuje velké množství klouzavých průměrů a jiných hodnot, podle kterých se na burze za poslední desetiletí vyhledávaly vhodná místa pro otevření pozice. Obrázek 2.9.2 pak ukazuje upravené rozložení. Cenové grafy jsou obohaceny o exponenciální klouzavý průměr EMA204 (červený) a EMA34 (modrý), dále o zelené šipky ukazující otevření obchodní pozice a její uzavření. Pod tímto grafem jsou zobrazeny další dva grafy a to indikátory CCI50 a CCI25. Poslední čtvrtý graf vykresluje volume. Více o indikátorech zmiňuje kapitola 2.10 Obchodní systémy, zde je jen předvedeno možné upravení programu Sierra Chart.

[1] Dle Investopedia, <http://www.investopedia.com>
[2] Sierra Chart, <http://www.sierrachart.com>
[3] Ninja Trader, <http://www.ninjatrader.com>



Obrázek 2.9.1: Standartní rozložení programu Sierra Chart

Obrázek 2.9.2: Modifikované rozložení programu Sierra Chart

Ninja Trader

Program Ninja Trader má mnoho stejných vlastností jako program Sierra Chart. Sice defaultně nastavené schéma je jiné, ale i zde může být modifikováno. Výhodou programu je snadné prvotní použití, protože program se snaží zůstat jednoduchým (nestandartní nastavování, kterému se uživatel Sierry Chart nevyhne, je zde častokrát skryto). Nevýhodu představuje práce s konfiguračními soubory, pokud je nutné nestandartní hodnoty měnit. Ninja Trader je lépe přizpůsoben pro komunikaci a obchodováním s brokerem. Obsahuje tlačítka pro nákup a prodej.[1] Tato funkčnost je možná i v programu Sierra Chart přidáním externích modulů (což může být komplikované).

Obrázek 2.9.3: Rozložení programu Ninja Trader

Jak jde vidět, samotné zobrazení grafu a indikátorů, je nastaveno pro program Sierra Chart a Ninja Trader velmi podobně, což umožnuje snadný přechod z jednoho programu do druhého. Největšími rozdíly tedy zůstávají cena (respektive způsob platby) a očekávané využití. Zatímco program Sierra Chart se specializuje na práci s historickými daty (má jednoduší vyhledávání v historii a může obsahovat i indikátory, které pro svojí složitost nemohou být zobrazeny v reálném čase). Oproti tomu Ninja Trader je optimalizován pro obchodování (tlačítka pro komunikaci s brokerem a přítomnost jednodušších v reálném čase spočítatelných indikátorů). Tyto programy nemusí znamenat vzájemnou konkurenci. Profesionální obchodníci mohou využívat oba (a proto se dá nastavit zobrazení velmi podobně). [7]

Tato práce rozhodně nemůže představovat manuál na ovládání některého z programů, ani výčet všech možných programů pro zobrazování cenových grafů. Podkapitola tedy slouží jako možný příklad, který pouze uceluje obecné znalosti, potřebné pro možnosti efektivně s burzovním světem pracovat.

[1] Dle Ninja Trader, <http://www.ninjatrader.com>

2.10 Obchodní systémy

Programy jako Sierra Chart a Ninja Trader slouží k přehlednému pohledu na data a jejich nejrůznější indikátory. Tyto programy umožňují snadnější vyhledávání vzorů, podle kterých se otevírají a zavírají pozice. Vzory jsou jednou ze základních složek obchodního systému. Jak již bylo naznačeno, tak ke kompletnímu obchodnímu systému, je nutné vytvořit i pravidla money managementu, řízení rizika a zvládnutí psychiky obchodování. [12] Tato práce, se však těmito složkami nezabývá a soustředí svoji pozornost pouze na vzory vzniklé technickou analýzou.

Před samotným vysvětlením principů učení je vhodné, aby se čtenář seznámil s některými vzory nebo indikátory, které obchodníci běžně používají. Tato práce nemůže shrnout všechny indikátory už jen z toho důvodu, že vznikají neustále nové. Na mnoha diskuzních fórech se denně objevuje nějaká kombinace vlastností cenové řady (o které její autor tvrdí, že se jedná o dokonalý vzor). Dokonalým vzorem je myšlen takový, který by efektivně fungoval na všech komoditách, a zároveň by se jeho pravděpodobnost správné predikce blížila ke stu procentům. [12] Takové vzory však neexistují a většinou se ukáže, že zmiňovaný vzor je jen překombinovaným a otestovaným na příliš malém vzorku dat. [3] Mnoho dnešních obchodníků neumí programovat a pracovat s automatickými systémy, a proto nedokáží svůj nalezený vzor efektivně ověřit. Při ručním zpracování je takřka nemožné porovnat vzor napříč dvaceti lety historie a desítkám různých komodit.

Přesto by bylo vhodné popsat pár vzorů, aby měl čtenář konkrétní představu toho, jak výsledné vzory mohou vypadat. Jednotlivé vzory pracují s různou skupinou nejrůznějších indikátorů. Pro ruční obchodování, je však vhodné, aby kombinace všech nutných indikátorů, nebyla příliš vysoká, jinak by se v nich nedokázal spekulant vyznat. Pokud by pracoval s například 10 vzory, a každý by využíval 3 různé indikátory, bylo by v cenovém grafu zobrazeno přes 30 dalších hodnot. Toto množství by spekulant již nezvládl efektivně kontrolovat. Následující vzory (0/V, 2V a BigV) pracují s indikátory EMA a CCI.

Indikátor EMA

Exponencial Moving Avarage jedná se o typ klouzavého průměru. Tento klouzavý průměr nepracuje s hodnotami, ale jejich poměrnou částí. Míra, o kterou se hodnota sníží, závisí na stáří záznamu, čímž se do grafu lépe promítají aktuální hodnoty.[1] Pokud by byl využit pouze klouzavý průměr tak by se po rostoucím trendu náhlý propad objevil až za delší čas. Jenže v EMA se nejaktuálnější hodnota nejsilněji projeví (přesto je propad zmírněn ostatními hodnotami průměru), a tak se náhlé propady ihned zobrazují i v grafu. Při analýzách cenových grafů se využívá EMA nejrůznějších délek (počet svíček zahrnutých do výpočtu). Server finančník.cz ve svém obchodním systému využívá EMA204, která bude na následujících grafech zobrazena červeně, a EMA34 která je zobrazena modře.[2] Ze znalosti cenových řad lze odvodit, že EMA204 představuje dlouhodobý trend, který se v cenové řadě vyskytuje, naproti tomu EMA34 reaguje na změny mnohem svižněji a zobrazuje aktuálnější hodnoty (například trend dané hodiny).

[1] Dle FXstreet, <http://www.fxstreet.cz>
[2] Dle Finančník, <http://www.financnik.cz>

Indikátor CCI

Commodity Channel Index jedná se o takzvaný momentum indikátor. Měří a zobrazuje sílu a rychlost trendu. [8] Pokud jsou hodnoty tohoto indikátoru výrazné (+100 pro kladný růst a -100 pro záporný), znamená to, že aktivita obchodníků je vysoká a jejich touha nakupovat také (CCI tedy zpracovává volume). Tento stav pak napomáhá k udržení daného trendu i jeho zesílení. Naopak pokud jsou CCI hodnoty nízké (okolo 0), je vůle obchodníků slabá a trh ztrácí svůj trend, což vede k pohybu trhu do strany (chop). Indikátor CCI se většinou využívá pro intradenní obchodování, protože na dlouhodobých intervalech bývá nepřesný.[1] Tento indikátor se využívá například v obchodním systému WoodyCCI a FinWin. Hodnota za zkratkou indikátoru, pak představuje množství zpracovávaných svíček. Čím je vyšší, tím pomaleji indikátor reaguje

I přes příznivou kombinaci indikátorů nemusí se cenová řada vydat očekávaným směrem. Například se muže stát, že CCI indikátor bude u hodnoty +100, EMA204 bude ukazovat rostoucí trend, EMA34 se bude nacházet nad EMA204, čímž potvrdí, že aktuální trend odpovídá dlouhodobému trendu a i cena komodity poroste. Podle indikátorů se jedná o ideální místo pro otevření pozice, neboť vše nasvědčuje tomu, že trh ještě dlouho bude pokračovat ve svém růstu. A přesto ihned po vstupu do pozice, se vše otočí a spekulant ztratí svůj kapitál. Jistě existuje nějaké pravidlo, které by dokázalo danou situaci vysvětlit a predikovat, ale daný obchodní systém jej nebral v potaz. [11] Vzory v obchodním systému pracují jen s jistou pravděpodobností, že je jejich predikce správná, a tak je nutné, aby byl spekulant připraven na ztrátu, i když vše vypadá naprosto ideálně. [1]

K omezení ztráty slouží stop-loss. [7] Jedná se o techniku z money managementu a snižování rizika, a proto zde nejsou popsány správné postupy použití. Pro tuto práci plně stačí vysvětlení, že se jedná o jakousi míru ztráty, kterou je spekulant ochoten podstoupit. Pokud predikce nedopadne dobře, a trh se vydá nepředpokládaným směrem, tak po dosažení stop-lossu je pozice automaticky uzavřena (stop-loss je v grafech zaznačen zelenou horizontální linkou).

1. **BigV –** Jedná se o vzor, který nepracuje jen se CCI14. Pokud trh roste (spekulant chce provést nákup) tak by EMA34 měla být nad EMA204 a v grafu CCI50 se musí vytvořit útvar připomínající V či W, ne však U (Jde o to, že útvar musí mít ostrou špičku). Dále musí být hodnota CCI50 kladná a vrchol špičky se musí nacházet mezi hodnotou 0-35, zatímco nejvyšší hrana útvaru V či W musí být nad hodnotou 100. Potom byl vykreslen vzor BigV, který predikuje pokračující růst a je vhodné provést otevření pozice nákupem kontraktu.[2] Pro prodej musí být situace přesně opačná: EMA34 pod EMA204 a CCI50 se musí nacházet v záporných hodnotách 0 až -35 a pod hodnotou -100. Právě tyto nepřesně definované útvary V/W ne však U, jsou pro automatické zpracování velmi složité, protože nikde není řečen úhel, který by měl vzniknout, ani jak moc může a nemůže být hrana útvaru lámaná.

[1] Dle Investopedia, <http://www.investopedia.com>
[2] Dle Finančník, <http://www.financnik.cz>

Obrázek 2.10.1: BigV vzor

2. **0/v –** U tohoto vzoru musí CCI50 udělat V či W. Je vhodné aby EMA34 byla ve stejném trendu jako EMA204 (viz BigV) a indikátor CCI14 se přenese do stejných hodnot jako CCI50. Pokud se V u CCI50 vytvořilo v kladných hodnotách, potom musí CCI14 přejít ze záporných hodnot do kladných.[1]
3. **2V –** Oba indikátory CCI udělají útvar V či W ve stejném směru. Pokud jsou v kladných hodnotách, jedná se o indikaci k nákupu. Pokud jsou v záporných, je vhodné otevřít prodejní pozici. U tohoto vzoru nehraje roli postavení indikátorů EMA.[2]
4. **EC –** V tomto vzoru se CCI50 musí dostat přes 0 do stejné oblasti, jako se nachází CCI14. CCI14 musí být maximálně „pár úseček“ nad oblastí 100. Pokud se tedy CCI14 nachází v oblasti -100 a méně, musí CCI50 přejít z oblasti kladné do záporné a naopak. Opět jde vidět, že definice „pár úseček“, může být pro automatické zpracování velmi složitá, nicméně obecně se doporučuje, že by těch úseček nemělo být více jak 4.[3]

[1] Dle Finančník, <http://www.financnik.cz>
[2] Dle Finančník, <http://www.financnik.cz>
[3] Dle Finančník, <http://www.financnik.cz>

Obrázek 2.10.2: Zmiňované vzory

Obchodní systém bývá tvořen několika podobnými vzory. Žádný ze vzorů však není 100% a cílem obchodního systému je sestavit takovou paletu vzorů, aby z dlouhodobého hlediska dávaly kladné výsledky a přinášely profit.[1]

2.11 Vývojové nástroje

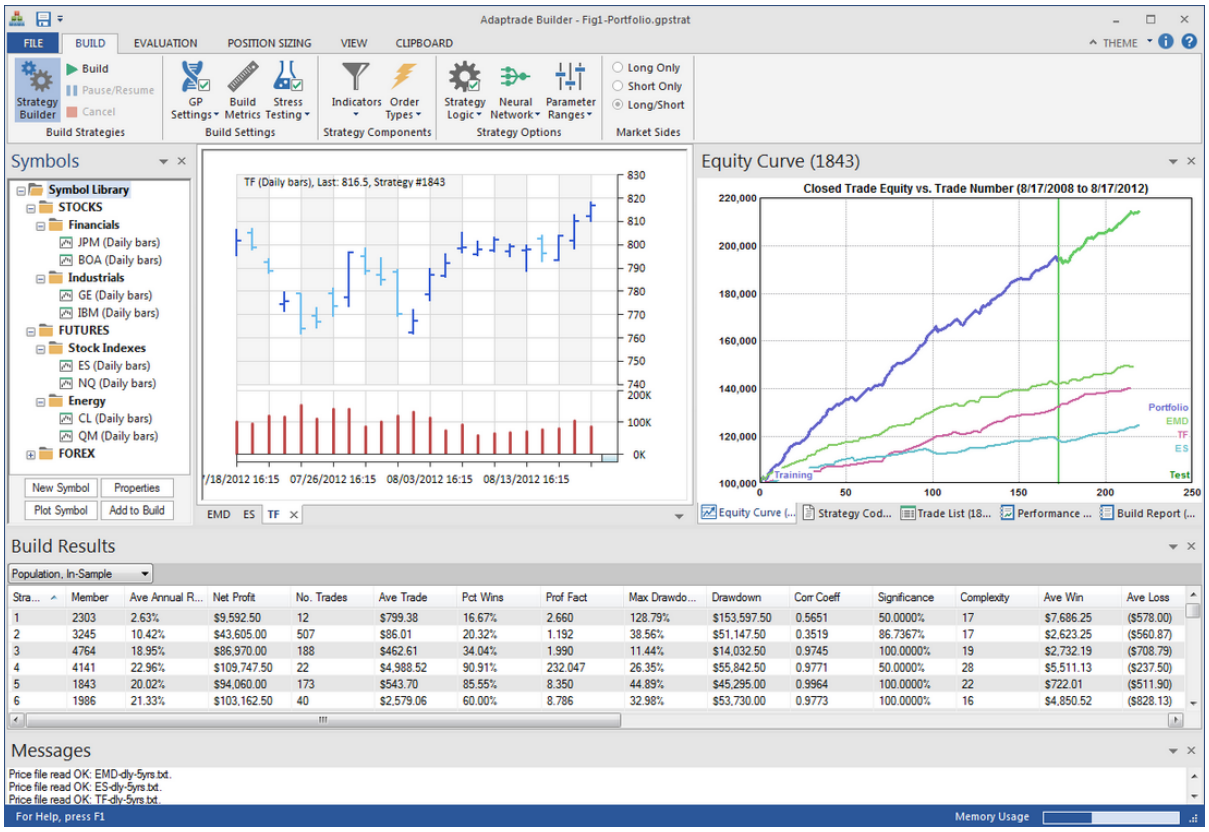
Jelikož se obchodní systémy neustále vyvíjí, je vhodné, aby proces hledání byl nějakým způsobem automatizovaný. Stejně tak je nutné, ověřit zdali nové vzory jsou dostatečně obecné, a k porovnání mnohaleté historie se opět hodí automatický nástroj.

V dnešní době již mnoho automatických systémů existuje, problém však může být s věrohodností nebo přesností vzorů. Nicméně existují programy, které dávají velmi kvalitní, přesné a věrohodné výsledky. Takovéto systémy jsou tvořeny desítkami analytiků a developerů a jejich přesnost je rozhodně obdivuhodná. Příkladem těchto programů může být TradeStation[2] nebo WelthLab[3]. Dalším pozitivem je, že tyto společnosti dodávají vlastní systém pro vizualizaci trhu a indikátorů (podobný jako Sierra Chart). Jediným problémem je v tomto případě cena (více než 10 000 dolarů ročně). Jedná se tedy o systémy pro profesionální obchodníky, kteří operují s tak ohromným obratem, že jsou pro ně poplatky 50 tisíc dolarů ročně (poplatek za data, automatické upozorňování na vhodné obchodní příležitosti, odebírání analýz od externích stran, atd.), jen zanedbatelnou částkou. Pro klasického začínajícího spekulanta jsou však tyto programy cenově nedostupné. I kdyby dokázal obyčejný spekulant zdvojnásobit celý svůj kapitál každý rok (což je velmi nepravděpodobné), tak je nutné si uvědomit, že většina obchodníků začíná s 5 až 10 tisíci dolary. Poplatky za využité systémy by tak byly mnohem vyšší, než celý jejich zisk. Tito lidé, kterých je na burze výrazná většina, proto musejí volit levnější systémy a u těch již hrozí, že se nebude jednat o úplně vhodné řešení.

[1] Dle Investopedia, <http://www.investopedia.com>
[2] Trade Station, <http://www.tradestation.com>
[3] Welth Lab, <http://www.wealth-lab.com>

AdaptTrade Builder

Jedná se o program, který slouží k automatické tvorbě obchodního systému. Licence programu, je již o poznání levnější, přesto poplatek zhruba 20 tisíc korun (záleží na rozsahu funkčnosti) nemusí všem vyhovovat. Tento program již nenabízí žádné doprovodné systémy pro zobrazování grafů, cenových grafů a indikátorů (vše co program nalezne, si musí uživatel nastavit a vyhledávat v programu jako Sierra Chart sám).[1] Program pracuje na bázi genetických algoritmů, které mu prý zaručují vysokou rychlost a přesnost ve zpracování dat. [16] Dle manuálu je možné vyměňovat vstupní data a tím si alespoň částečně ověřit, jestli jsou nalezené vzory aktuální.



Obrázek 2.11.1: Program AdaptTrade Builder[2]

Vzhledem k vysokým cenám licencí autor s danými programy nepracoval a nemůže porovnat jejich výsledky.

[1] Dle Adaptive Trader, <http://www.adaptrade.com/product.php>

[2] Převzato z Adaptive Trader, <http://www.adaptrade.com>

3 Strojové učení

Strojové učení se využívá v mnoha oblastech, avšak tato práce se bude soustředit na vyhledávání použitelných vzorů pro predikci vývoje cenové řady. Postupů strojového učení je mnoho a z důvodu omezeného rozsahu práce, jsou vybrány, pouze příklady, které nejčastěji souvisejí s predikcí cenových řad.

Díky strojovému učení mohou analytici mnohonásobně zvýšit rychlost, s jakou své obchodní systémy obnovují a přizpůsobují. Budou tak schopni provádět jednotlivé procesy opakovaně a hledání vhodného obchodního systému jim již nezabere dlouhé týdny. [5]

Jak již bylo naznačeno v úvodní kapitole, strojové učení je netriviálním úkolem. Pro strojovou predikci je důležité naučit program rozeznat očekávaný výsledek. Už samotná definice výsledku může být velmi problematickou úlohou, jejíž náročnost závisí na specifičnosti úlohy. Ve světě obchodu je odpověď většinou jednoduchá: zisk. [3] Poté co se podaří definovat výsledek, je nutné zpracovat veškerá data a najít v nich specifické vzorce. [16] Tyto vzorce pak slouží ke statistické predikci chování nebo vývoje.

Následující podkapitoly se pokusí shrnout často užívané principy, které jsou vhodné pro strojové učení nad cenovými řadami. Kompletní rozbor obecného strojového učení, by překročil rozsah práce. Jsou tedy nastíněny postupy od očištění dat, přes samotné učení a vyhledávání vzorů až po testování a vyhodnocení.

3.1 Získávání dat a jejich analýza

Jelikož se práce zabývá predikcí číselných řad, bylo by vhodné nastínit celý proces od získání dat až po výsledné testování systému. Jednotlivé oblasti budou podrobněji rozvedeny v následujících kapitolách, jenž čtenáři umožní získat ucelený pohled na celou problematiku.

Strojové učení podléhá několika fázím, od získání dat přes klasifikaci a predikci, až po závěrečnou interpretaci výsledků. Po samotném získání dat je nutná fáze předzpracování těchto dat. [13] Údaje, se kterými se pracuje, častokrát vznikly v průběhu několika let. Během té doby se mohly měnit způsoby, jakými byly data zaznamenávána i samotné zaznamenávané hodnoty. Pro strojové učení je vhodné data unifikovat, tedy převést na stejný způsob zápisu a se stejnými atributy.

Poté, co se podaří získat strojově zpracovatelná data, je vhodné nad nimi provést analýzu. Touto analýzou se autor, pokusí o datech zjistit co nejvíce informací, které mu pomohou vybrat vhodné metody pro učení. Je možné využít více způsobů, jejichž výběr závisí na typu dat samotných. Mezi nejznámější způsoby zobrazení rozložení dat patří histogram hodnot. Například histogram, kdy byla data zaznamenávána, umožňuje odhadnout množství výpadků a chyb v datech. Tato hodnota pak může indikovat nepoužitelnost některých metod strojového učení. Dále je možno vyloučit dny, které mají až nepřirozeně málo záznamů oproti ostatním a jen by zanesly chybu do celkových výpočtů. Tomuto procesu se říká čištění dat a odstranění šumu. Na bázi nejrůznějších statistik jsou odstraněny hodnoty, které by do učení přinášely chyby. Stroj by se nemohl naučit se systémem dobře pracovat, protože v samotné definici systému (číselné řadě), by se nacházelo mnoho chyb. Když by pak počítač pracoval s aktuálními hodnotami, představovaly by pro něj úplně jinou řadu. [19] Například většina poskytovatelů dat vývoje ceny komodit nezaznamenávala vývoj ceny o víkendu příliš pečlivě, protože

dříve se o víkendu neobchodovalo. Tyto výpadky však v dnešní době již nevznikají, a kdyby byl použit takto naučený systém, nemusel by se o víkendech chovat spolehlivě.[1]

Některé postupy vyžadují ještě integraci dat z různých zdrojů. Je nutné datové vstupy opět unifikovat a stanovit jeden formát (historická data jedné společnosti mohou být zaznamenávána jinak než u jiné společnosti). Je nutné sjednotit intervaly dat a vhodně vyřešit překryvy. [20] Opět se jedná o nelehký úkol, kdy autor programu musí rozhodnout, jak se bude s překryvy pracovat. Pokud jeden zdroj hlásí jiné hodnoty než druhý, musí být stanoveno, které údaje se použijí nebo jakým způsobem se získá výsledná hodnota (například aritmetickým průměrem). Dané rozhodnutí většinou vychází z komplexnější znalosti oblasti, nad kterou program provádí učení.

Poté co jsou data vyčištěna a již je stanoven jen jeden datový vstup, je pro některé metody vhodná i jistá redukce dat. [1] Při historii několika let se záznamem pro každou vteřinu, by program musel zpracovat enormní množství záznamů. Vhodnost tohoto zpracování závisí převážně na tom, co uživatel v daném systému hledá. Přesto může být vhodné využít jisté redukce, například shlukováním dat do minutových intervalů. Při reprezentaci cenových řad se využívají svíčkové grafy s různou velikostí timeframe. [4]

Ať už jsou data redukována či nikoliv, předají se jedné z metod, která je vhodná pro strojové učení. Pro zpracování číselných řad se častokrát využívají neuronové sítě, které se snaží simulovat způsob učení, jenž provádí i lidský mozek. Další možné metody jdou například násobná regrese nebo Baerovská klasifikace, Kalmanův filtr a některé další. [13] Rozličné metody v datech vyhledávají různé podobnosti, vzory nebo jen statistické odpovědi na aktuální vstup. Výsledkem však je, že by stroj měl být schopen klasifikovat či rozpoznat vstupní vzorek dat a na jejich bázi odhadnout data výstupní nebo následující. Metody strojového učení se nevyužívají jen na číselné řady, ale i například na rozpoznávání předmětů v obrázcích, nebo klasifikaci zákazníků, aby bylo možné odhadovat, jaké položky si přijdou koupit (ty se pak většinou dávají dál od sebe, aby zákazník musel projít celým obchodem).

Pokud metoda strojového učení dokončí svůj běh úspěšně, je nutné otestovat ji na dalších datech. Většinou se využívá část historických dat k učení a druhá část je pak použita pro validaci programu. Zde opět musí zasáhnout tvůrce programu, který by měl mít komplexnější znalosti o dané problematice, aby posoudil, zdali jsou odhalené vzory odpovídající realitě a jsou pravděpodobné.

Po schválení všech výsledků je strojové učení ukončeno a může být používáno. Protože se však vyvíjí lidská společnost i svět, ve kterém žijeme, nebudou odhalené zákonitosti platit vždy a proto je nutné postupy pravidelně opakovat a nalezené vzory zpřesnit. Některé metody pak využívají principu učení i v průběhu svého používání, aby se tím neustále kalibrovaly a zpřesňovaly. [26] Přestože se některé systémy optimalizují v průběhu svého běhu, je někdy nutné aplikovat jiné metody učení při skokové změně chování systému.

Pro účely práce jsou data vnímána jako vstup celého procesu učení. Jelikož se práce zaměřuje na predikci číselných řad reflektující vývoj na burze, budou vstupními daty právě cenové řady a jejich historický vývoj. Stejným způsobem je možné specifikovat výstup celého strojového učení. Výstupem bude množina vzorů, která slouží k predikci vývoje nebo ke klasifikaci vzorku do skupiny (pak se nejedná o vzory ale třídy). V burzovním prostředí, je vzor reprezentován postavením nejrůznějších indikátorů a hodnot cenové řady. Vzor také říká, jak se má spekulant zachovat, když dané uskupení nastane. Možné akce na výskyt vzoru jsou otevření obchodní pozice a to buď nákupem kontraktu nebo naopak prodejem.

[1] Dle Investopedia, <http://www.investopedia.com>

Pokud by mělo být shrnuto strojové učení v několika málo krocích, tak jak je vhodné je aplikovat nad cenovými řadami, jednalo by se pravděpodobně o tyto body [13]:

1. Získání dat
2. Získání charakteristiky dat
3. Vyčištění dat
4. Integrace dat
5. Strojové učení
6. Testování
7. Provoz

3.2 Učení s učitelem a bez učitele

Před samotným procesem učení je vhodné veškerá data rozdělit obvykle do svou skupit, a to data určená pro učení tzv. trénovací data a data pro ověření jestli se program naučil správě tzv. testovací data. [17] Program je nad testovacími daty spuštěn teprve, až dokončí proces učení nad trénovacími daty. Na testovacích datech, se pak vyhodnotí statistická správnost systému, aby se dalo předpokládat, jaká bude úspěšnost i nad ostatními daty.

Učení s učitelem

Takzvané supervized learning, nebo supervized machine learning. Jedná se o učení, kdy jsou ke vstupním datům známá i data výstupní. [17] Tyto výsledky, nemusí nutně zadávat člověk, čímž je myšleno, že není důležité, aby člověk prováděl kontrolu, jestli byla data užita správně nebo ne.

Příkladem takového učení může být hledání vazeb mezi slovy. Předpokládejme, že existují určitá pravidla, kterými se z podstatných jmen tvoří jména přídavná nebo slovesa. Pokud počítač na vstupu získá nějaké podstatné jméno, a nebude existovat způsob, jakým by vybral správné pravidlo pro tvorbu přídavného jména, musel by aplikovat všechna pravidla, která se pro tuto morfologickou derivaci používají. Vygeneruje tak seznam potenciálních přídavných jmen, učitelem mu zde bude český slovník, který vyloučí všechna špatně vygenerovaná přídavná jména. Počítač se tak nad jednotlivými podstatnými jmény, bude učit rozpoznat jaké pravidlo pravděpodobně použít. Pro kompletní rozbor příkladu je možné použít předchozí práci autora Morfologické derivace českého jazyka. [15]

Učení se s učitelem, bývá mnohem rychlejší, než bez něj. Po každém vytvořeném výstupu, se program rychle dozví, jestli uspěl nebo nikoliv a může tak své rozhodování rychle poupravit, aby nad daty již nechyboval.

Při učení se s učitelem, je velké riziko tzv. přeučení. Jedná se o stav, kdy je program špatně citlivý na jednotlivé signály, které na trénovacích datech dostává. [14] Program se pak sice naučí správně reagovat na trénovací data, ale jeho úspěšnost nad testovacími daty je velmi slabá. Pokud dojde k přeučení, je program pro budoucí použití nevhodný a je nutné učit jej znovu s jinými parametry na citlivost výstupních signálů.

Učení bez učitele

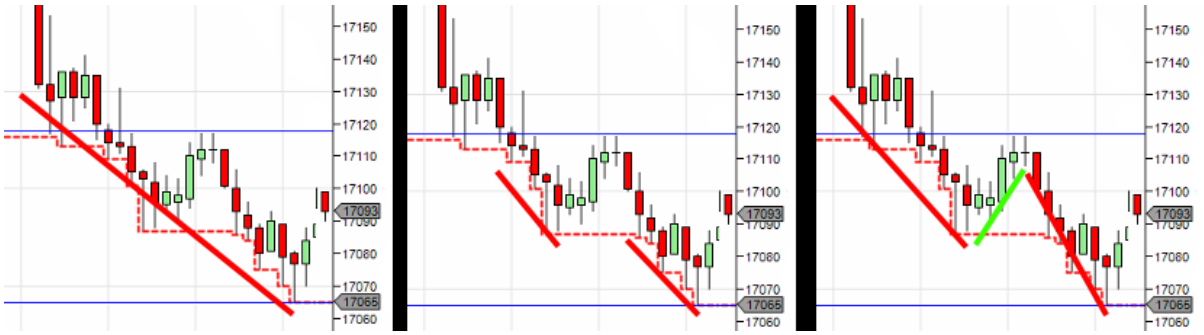
Jedná se o tzv. unsupervizited learning. Při učení bez učitele nejsou známy výstupy, jakými má program reagovat na vstupní data. Proces učení bez učitele se využívá hlavně ke klasifikaci či uspořádání dat. Hledá se optimální rozložení dat nebo zcela nové vazby. [14] Při hledání nových vazeb mezi daty, je pak velmi důležitá kontrola daných výsledků a logické posouzení výstupních dat. V cenových burzovních grafech by daný proces odpovídal racionální analýze, protože je nutné pomocí

odhalení fundamentálních prvků odfiltrovat chybné výsledky technického zpracování. Tento princip učení trvá výrazně delší dobu a častokrát i ve více iteracích. [17]

Cenová funkce

Jelikož učení se bez učitele je časově náročnější, je proto vhodné využívat učení s učitelem. Tato metoda může být využita i při zpracování cenových řad. Existuje funkce, která říká, jestli program vyhodnotil situaci správně či nikoliv. Takováto funkce vlastně stanovuje očekávané výsledky a v obchodním světě je tímto očekávaným výsledkem profit. Protože nad cenovými řadami je možné spočítat výši profitu mezi dvěma body, je možné využít cenovou funkci, zdali je odezva programu správná či nikoliv. V burzovním světě se cenové funkci typicky říká funkce profitová, neboť slouží k výpočtu možného profitu. Právě funkce vyhodnocující profit bude simulovat učitele, podle kterého se musí program při učení korigovat.

Přestože se může zdát, že tato funkce stanovující výši profitu, bude velmi jednoduchá (pouhý rozdíl hodnot grafu v bodě A a B) opak je pravdou. Pro odpověď jestli je míra profitu správná, musí být stanoveno, jakým způsobem může vypadat cenový graf mezi body A a B. [7] Na následujícím obrázku lze vidět, že míra profitu je závislá na tom jak se budoucí řada zachová. Pokud se směr křivky mění, musí existovat hranice, podle které bude určeno, jestli je čekání na vrácení se ještě akceptovatelné nebo nikoliv a problém by měl být řešen dalším vzorem.



Obrázek 3.2.1: Stejná cenová řada, ale různé profity

Každý ze znázorněných cenových grafů může být považován za jednu zápornou míru profitu, anebo za několik posloupností. Je proto velmi důležité správně nastavit jak se má program zachovat a toto chování musí profitová funkce respektovat.

Pro některé spekulanty by totiž mohl být velmi nepříjemný případ B nebo C a obchodní systém by pak pro ně byl neobchodovatelný. Někteří lidé nedokáží psychicky unést, že by ztráceli již nabytý profit, a mohli by potom provést chybu, která by narušovala jejich obchodní systém [11] (je nutné poznamenat, že některé publikace zabývající se obchodováním na burze přidělují váze psychiky až 60%, proto pokud nebude spekulant v klidu, a klesající graf jej rozruší, může to jeho obchodní strategii velmi narušovat). Je proto možné, že by si spekulanti raději přáli, aby se systém naučil vyhodnotit dané posloupnost jako dva oddělené profity. [12]

Když bude cenová funkce schopna odpovídat tak, jak by odpovídal sám spekulant. Může být proces považován za učení se s učitelem a výsledky budou mnohem více odpovídat potřebám každého analytika.

3.3 Proces získávání znalostí

V této podkapitole, budou shrnuty jednotlivé kroky, které musí být vykonány k tomu, aby mohlo být strojové učení a získání znalostí z dat úspěšné. Pod pojmem získání znalostí je v tomto případě myšleno nalezení vzorů, které by byly schopné s určitou pravděpodobností předpovídat vývoj cenové řady neboli možný utržen profit. Obecné kroky pro získávání znalostí z dat jsou definovány takto: [13]

1. **Čistění dat** – v této fázi se programátor musí vypořádat s chybami, které datový zdroj obsahuje. Jedná se hlavně o chybějící data, odstranění šumu v datech a vyřešení nekonzistence. Pro komodity je typické, že se dříve (před rozmachem internetového obchodování) v určitých hodinách (mimo pracovní domu burzovního domu) a o víkendech neobchodovalo a data nebyla zaznamenávána nebo velmi náhodně (broker zjišťoval hodnotu komodity například jen v poledne). Tyta chybějící data by měla být odstraněna, neboť by při učení zanesla mnoho chyb. Jak je zmíněno v kapitole 5, při výběru nevhodného zdroje dat, může tato fáze být velmi problematická a zdlouhavá.
2. **Integrace dat** – v některých případech nejsou zpracovávána všechna data jen od jednoho zdroje. Je tak nutné, převést jednotlivá data takovým způsobem, aby byla vzájemně kompatibilní a bylo možné s nimi pracovat, jako kdyby pocházeli pouze z jednoho zdroje. [21] Tato fáze je velmi důležitá, protože jinak by bylo vytváření obecného učícího se algoritmu velmi složité. Velmi často se od různých poskytovatelů nacházejí různé položky na různých místech. Hodnoty open, hight, low, close, sice bývají často ve správném pořadí, protože se jedná o základní veličiny, pro vykreslení svíčkového grafu. [7] Ostatní doplňující hodnoty, jako například volume, ask, bid, už každý poskytovatel dat má na jiném místě (pokud hodnoty vůbec poskytuje). Také je nutné sjednotit oddělovače jednotlivých hodnot.
3. **Výběr dat** – v této fázi musí analytik specifikovat, s jakými daty se vůbec bude pracovat. Získávání znalostí z databází se neprovádí jen nad cenovými řadami, ale i například nad relačními databázemi, kde jednotlivé tabulky obsahují mnoho atributů, z nichž většina může být pro danou úlohu zbytečná. Systém by se pak mohl snažit v klasifikacích shlukovat prvky podle atributů, které by jen snižovaly shodu. Pro cenové grafy může být důležité specifikovat, zdali bude brán důraz i na pomocné hodnoty jako volume, nabídku, poptávku, či výběr časového měřítka.
4. **Transformace dat** – cílem této fáze je převedení dat do podoby vhodné pro zpracování danou metodou učení. Častokrát se provádí nejrůznější agregace. U cenových řad je nejčastější agregací právě sestavení tzv. svíčky o velikosti zvoleného časového měřítka. Pokud jsou data s přesností na jeden TICK je hodnota open, hight, low i close častokrát stejná, neboť se za tak krátký časový interval nemusí vůbec změnit. [7] Při práci s hodinovým intervalem, by bylo učení zahlceno množstvím stejných hodnot (nebo lišících se právě o jednu obchodovatelnou jednotku změny (např. desetina, ale i 1/25 atd.)). Při vhodné agregaci se výrazně sníží počet dat, se kterými se musí pracovat. To je velmi výhodné pro rychlost a efektivitu procesu učení.
5. **Dolování dat / proces učení** – jedná se o nejdůležitější část procesu získávání znalostí. S využitím zvolené metody se systém pokusí získat co nejlepší výsledky. S ohledem na typ metody, která je použita (s ohledem na očekávaný výsledek) se rozlišují dva základní přístupy:
    a. **Deskriptivní** – jedná se o přístup, který se snaží charakterizovat vlastnosti analyzovaných dat. Hledají se skryté vazby nebo optimální rozložení. [14] Příkladem může být hledání preferencí zákazníků cestovních společností s ohledem na nejrůznější aspekty, které se podařilo během historie společnosti zaznamenat.

Tato práce se daným typem úloh zabývat nebude a tento typ je uveden pouze pro kompletní doplnění.

b. **Prediktivní** – tento typ úloh se snaží vstupní data analyzovat za účelem vytvoření vzorů nebo kategorií. Výsledné vzory by pak měly být použitelné pro odhadování budoucího chování. Příkladem může být odhalování délky života. Vstupem takových odhadů častokrát bývá životní styl a choroby daného člověka a jeho rodičů. Systém pomocí klasifikací vytvoří několik skupin (vzorů), které by pak měly být schopné odhadovat délku života jakékoliv osoby. Dalším typickým příkladem jsou vývoje číselných řad, což přímo souvisí s touto prací.

6. **Hodnocení modelů a vzorů** – cílem je identifikovat zajímavé a vhodné vzory a naopak ignorovat vzory, které nejsou v praxi použitelné nebo nepřinášejí žádnou novou užitnou hodnotu.
7. **Prezentace znalostí** – poté co jsou vzory nalezeny, je nutné je prezentovat, aby mohly být získané znalosti využity v praxi. [16] Nemusí se jednat o veřejnou prezentaci, ale je důležité výsledky zaznamenat i se všemi závěry v pochopitelné podobě.

3.4 Předzpracování dat

V předchozí podkapitole 3.3 o procesu získávání znalostí, jsou vysvětleny jednotlivé kroky, které musí být pro získání znalostí provedeny. Body 1 až 4 (čili čištění dat, integrace dat, výběr dat a transformace dat) by se daly souhrnně nazvat jako fáze předzpracování dat. Tato kapitola podrobně rozvádí jednotlivé části a obecné problémy, které se v datech objevují.

Fáze předzpracování dat je velice důležitá, protože pomáhá data „očistit“, aby při procesu učení nedocházelo k chybám ani jiným zavádějícím výsledkům. Málokdy se podaří získat data zcela připravená pro strojové učení. Je mnohem pravděpodobnější, že analytik musí zpracovat data „špinavá“ (z anglického slova dirty). Pomocí nejrůznějších technik, které jsou vysvětleny později, se pak tvůrce snaží vytvořit čistá data. Za čistá data se obecně považují taková, která mají následující vlastnosti: [19]

- **Přesnost** – data by měla co nejvěrněji reprezentovat realitu.
- **Úplnost** – jedná se o vlastnost co do šířky definice dat, čímž je myšlen počet atributů, tak i do hloubky, čili vhodného množství takovýchto dat.
- **Konzistence** – data musí být konzistentní, neměly by například odporovat sami sobě.
- **Aktuálnost** – data musí být aktuální, pro danou znalost, která má být z dat získána. Jen s aktuálními daty je možné rozhodovat o budoucnosti.

U cenových řad obchodních komodit je vhodné také zvážit následující vlastnosti: [7]

- **Důvěryhodnost** – data musí být z důvěryhodného zdroje, protože jinak by analytik neměl jistotu, zdali může výsledky použít.
- **Přidaná hodnota** – jedná se o míru prospěšnosti dat pro řešení zvolené úlohy nebo pro získání hledané znalosti (vzoru).
- **Interpretovatelnost** – hodnoty v datech by měly být snadno interpretovatelné, možným příkladem je nevhodná reprezentace enumerátorů.
- **Dostupnost** – jedná se o vlastnost, která udává, jak snadno jsou data dostupná.

Vlastnosti čistých dat platí obecně nejen pro číselné řady. Při procesu čištění dat se analytik nejčastěji setká s následujícími problémy: v datech chybí hodnota nebo je uvedena ale nedává smysl či odporuje ostatním hodnotám.

Nekompletnost dat

Nekompletním datům chybí některé atributy, které by pro získávání znalostí byly vhodné. [18] Důvodů, proč data chybí, může být více. Je pravděpodobné, že se jednalo o nepovinnou položku ve formuláři a tak ji většina uživatelů nevyplnila nebo se při vytvoření aplikace a databáze nevědělo, že budou nějaká další data potřebná. [20] Typickým příkladem pro cenové grafy je nedostatek hodnot pro starší data o víkendu a mimo hlavní obchodní hodiny. Dříve se v této době neobchodovalo, a tak se data nezaznamenávala. Mnoho poskytovatelů dat, pak historii z těchto období neudržuje, protože by jim zbytečně zabírala místo na serverech.

Dalším typickým projevem nekonzistence dat je jejich agregace. Položky sice v datech existují, ale již jsou agregovaná, a tak již není možné se dostat k původním hodnotám. Agregace navíc nemusela být zvolena vhodně (byl například použit obyčejný průměr místo vhodnějšího váhového nebo mediánu) a tím může být vazba na původní hodnoty zcela zničena. Většina poskytovatelů dat například převádí starší (třeba už týden stará) TICK data na minutové intervaly, což jim výrazně šetří jejich úložnou kapacitu.

Šum v datech

Pokud data obsahují šum, jedná se o tzv. noisy data. Znamená to, že některé atributy obsahují nesprávné nebo velmi odlehlé hodnoty. Tyto hodnoty jsou v reálném světě chybné nebo opravdu velmi nepravděpodobné. [18] Příkladem takové hodnoty by mohla být tělesná teplota při horečce nad 45 °C. V případě cenových grafů, by se mohlo jednat o příliš vysoký nárůst za velmi krátkou dobu. Pokud se v datech vyskytují takové případy, je nezbytné je vhodným způsobem odstranit.

Většinu šumu, který by mohl vznikat, se snaží odstranit už samotná aplikace pro sběr dat. [20] Například nejrůznějšími kontrolami, jestli zadaný email obsahuje znak „@“ a tečku.

Dalším typickým problémem šumu, je změna nějaké konvence v datech. Mezi nejčastější příklady rozhodně patří změna kódování (například známky na vysoké škole dřív byly číselné hodnoty, dnes se jedná o písmena) a změna formátu dat (především u dat je pořadí roku, měsíce a dne v různých aplikacích reprezentováno jinak). Tyto změny v konvencích se většinou dají transformovat a tím data sjednotit a odstranit z nich daný šum.

Nekonzistence dat

Obecně je potenciálním zdrojem každé nekonzistence redundance atributů. Pokud se v databázi vyskytuje stejný atribut na více místech, hrozí, že některé hodnoty nebudou aktualizovány. [18] Ať už je databáze navržena nevhodně (nebo byla redundance využita schválně, aby se nemuselo spojovat mnoho tabulek při běhu programu) nebo se jedná o stejná data z různých zdrojů, vždy hrozí, že dva stejné atributy budou mít různou hodnotu. Je nutné pak určit věrohodnější zdroj a nekonzistenci opravit. U cenových grafů je pak možné využít váhového průměru, nebo jiných technik popsaných v kapitole 2.1 Číselné řady. [39]

Charakteristika dat

Aby mohly být předchozí problémy odstraněny, musí být nejprve odhaleny. K tomu může sloužit mnoho způsobů jak získat charakteristiku dat a jak ji vhodně zobrazit, aby tvůrce strojového učení nemusel ručně procházet miliony záznamů. [10]

Protože číselné řady obsahují velké množství hodnot, je vhodné vzít v úvahu i vlastnosti funkcí, které se na daná data aplikují. Využití distributivní nebo algebraické míry oproti holistické může mít na běh programu velmi silný vliv. Pomocí daných metod může být získán rozptyl hodnot, jejich

vzájemná poloha či počet kategorií. Všechny tyto výsledky je vhodné zobrazit, aby bylo jednoduché rozhodnout, které hodnoty vyčnívají, a aby se na ně analytik mohl soustředit.

Vizualizace výsledků, je nejčastěji prováděna pomocí grafu. Typické grafy pro zobrazení charakteristik jsou: [22]

- Koláčové grafy
- Krabicové grafy (vysvětleny v kapitole 2.4 Vizualizace číselných řad)
- Čárové grafy
- Bodové grafy
- Sloupcové grafy a histogramy – pokud se zobrazují kategorické atributy (například barvy) je v grafu pro každou hodnotu atributu vlastní sloupec a pak se jedná o sloupcový graf. [20] Naopak pro kvantitativní atributy (například příjem) je využito intervalů (zpravidla o stejné šířce), které představují jistý rozptyl, a poté se graf nazývá histogramem.

Čistění dat

Poté co se podaří získat charakteristiky dat, a tím odhalit jednotlivé problémy, která data obsahují, je nutné tyto problémy odstranit (vyčistit). Proces čištění dat se zabývá postupy, jak vyřešit nejznámější problémy zmiňované dříve. Jednotlivé postupy je možné aplikovat pro řešení neúplnosti, šumu i konfliktů v datech. U řešení Konfliktu v datech z více různých zdrojů, však většinou stačí následovat věrohodnější zdroj dat. [21]

Nejjednodušším řešením problému v datech je ignorace těchto chybných dat. Avšak tímto způsobem vznikají v datovém vzorku mezery, které mohou znemožnit způsob učení. Další možností je manuální opravení vzorku dat, ale pokud je chybných více záznamů, může být manuální oprava velmi nepraktická. Nejvyužívanější metodou je automatické opravení dat. Při automatické opravě může tvůrce programu následovat mnoho různých metodik. Práce se pokusí nastínit alespoň nejčastěji používané: [13]

- **Globální konstanta** – jedná se o nejjednodušší řešení. V databázových systémech se jako globální konstanta používá hodnota NULL. Pokud už bude zvolena nějaká reálně použitelná hodnota, mělo by se jednat o neutrální hodnotu, která ovlivní výsledky pozdějšího zpracování jen minimálně.
- **Průměrná hodnota** – doplní chybějící atribut dle průměrné hodnoty daného atributu u ostatních záznamů. Při hledání průměru je vhodné zvážit využití mediánu nebo váhového průměru.
- **Průměrná hodnota z vzorků stejné třídy** – jedná se vlastně o specifický případ váhového průměru. Vzorky jsou roztřízeny do tříd dle jiných atributů a průměr se pak hledá jen mezi vzorky dané třídy (ostatní vzorky budou mít váhu 0). Bylo by tak možné hledat odhadovanou cenu podle hodiny a dne nebo například podle odhadované nabídky a poptávky v blízkém okolí. [20]
- **Průměrná hodnota dle trendu** – jedná se o doplnění hodnoty velmi podobné předchozímu případu. Třídou však je klasifikovaný probíhající trend. Dle vlastností číselné řady se tak dá určit jak silný je trend, jaké přináší průměrné stoupání nebo klesání a jak často a s jakým rozptylem kolísá. [32]
- **Hodnota dle korelace** – pokud je k dispozici více číselných řad, z nichž jedna je neúplná, je možné doplnit chybějící hodnoty vzhledem ke korelaci jednotlivých řad. V případě, že řady silně korelují, dá se z úplné řady doplnit hodnoty, které chybí. Naopak pokud silně nekorelují, budou chybějící hodnoty opakem úplné řady. [32]

- **Nejpravděpodobnější hodnota** – jedná se o mechanizmus, který se pokouší sám všechny vzorky klasifikovat a následně predikovat jaká hodnota by na daném místě měla být. Jinými slovy se využije strojového učení k nachystání datového vstupu pro strojové učení. Tato metoda sice dává nejkomplexnější výsledky, ale je velmi složitá oproti metodám předchozím. V tomto procesu učení je vhodné zahrnout i analýzu trendu dané řady a pokusit se najít korelace s jinými číselnými řadami. [13]
- **Plnění** – tato technika vyžaduje, aby se data dala setřídit tak, aby bylo možné určit blízké okolí. Poté se bude vyhlazovat právě toto okolí. Prostor bývá rozčleněn do košů (z anglického bins), které by měli mít stejnou frekvenci, čili by měly obsahovat zhruba stejný počet prvků. Hodnoty v koši jsou pak nahrazeny průměrem (nebo mediánem, či jinou zvolenou technikou) všech hodnot v koši. Při využití těchto principů jsou pak všechny hodnoty v koši stejné, což umožnuje redukci dat. Další možností je zachovat minima a maxima košů, aby byly přechody mezi koši plynulejší. [13]
- **Regrese** – data se při této metodě vyhladí tak, aby odpovídala regresní křivce, která je nalezena regresní funkcí (ta je vysvětlena později). Opět se jedná o metodu, která využívá strojové predikce (tentokrát se neurčuje chybějící hodnota, ale jak by hodnota měla zhruba vypadat) [23]
- **Shlukování** – jedná se o metodu, která se snaží najít odlehlé hodnoty, slouží tedy hlavně k detekci hodnot s pravděpodobným šumem. Tyto hodnoty se obvykle nepodaří zařadit do žádného shluku (nebo jeho velikost je v porovnání s ostatními zanedbatelná), tyto odlehlé shluky jsou pak přímými adepty na zbavení se šumu. [13]

Všechny tyto typy automatické náhrady však do dat přinášejí určité zkreslení (anglicky bias), se kterým je nutné počítat. Pokud by zkreslení bylo příliš výrazné (vstupní vzorek dat by byl příliš nepřesný a obsahoval by mnoho chyb), mohly by být výsledky strojového učení natolik zavádějící, že by pro praktické využití neměly být použity. V takovýchto případech je vhodné hledat kvalitnější zdroj datového vstupu.

Integrace a transformace dat

Při práci s více zdroji se data musí do sebe nějakým způsobem integrovat. Také musí být transformována na jednotný typ zápisu a všechny hodnoty musí být sjednoceny. Například teplota může být zaznamenávána v C nebo F, ale také může být zaznamenávána absolutně, nebo přírůstkově oproti minulému měření. [18] Tyto rozdíly je třeba odstranit, protože jinak by data jako taková nedávala smysl. Mezi nejznámější a nejčastěji řešené problémy při integraci dat patří: [21]

- **Konflikty schématu** – jedná se o problém především sloučení atributů. Stejné atributy mohly být u různých zdrojů pojmenovány jinak. V případě, že se nejedná o relační databázi, ale například o data uložená v souboru, je nutné, aby byly hodnoty převedeny i do správného pořadí.
- **Konflikty hodnot** – jedná se o problém, když stejné hodnoty jsou reprezentovány různě. U jednoho zdroje může být rok uveden celý (2016), zatímco u druhého je zkrácený (16). Dalším typickým příkladem jsou údaje zapsané absolutně nebo relativně.
- **Konflikty identifikace** – jedná se o problém, kdy je naprosto stejná položka u každého zdroje reprezentována jinak. Hlavním problémem identifikace je, že nemusí (a pravděpodobně ani nebudou) sedět ID jednotlivých záznamů. Je tak nutné vyhledávat jednotlivé záznamy a vzájemně je mapovat k sobě, což může být velmi zdlouhavé.
- **Redundance** – pokud se podaří data z různých zdrojů sloučit, mohou vzniknout redundance v rámci stejných atributů nebo atributů, které jsou vzájemně odvoditelné (věk a rok narození).

Jak jde vidět, tak většina problémů integrace dat je odstranitelná bez zanesení nějakého zkreslení do procesu učení. [20] Přesto integrace vyžaduje podrobnější nastudování datových zdrojů, aby mohl analytik rozhodnout, jakým způsobem se mají konflikty řešit.

Redukce dat

Přestože některé metody jsou uzpůsobené i pro zpracování extrémně velkého počtu dat, jakákoliv redukce je vždy vítaná. Výpočet trvá kratší čas a vyžaduje i méně prostředků (hlavně paměti). Proto je vhodné minimalizovat počet zpracovávaných dat na nezbytné minimum. [18]

Při redukci se snižuje počet atributů, například pokud jsou redundantní či se hodnota jednoho atributu dá odvodit z druhého (například datum narození a věk). Redukce atributů, úzce souvisí s výběrem nutných atributů pro zpracování. Je zbytečné klasifikovat data, která budou mít více rozměrů, jen protože analytik neodstranil atributy, které na očekávané znalosti nemají vliv.

Dalším typickým příkladem je redukce množství dat, se kterými se pracuje. [21] Některé vzorky mohou být pro výpočet zbytečné (nebo byly vyloučeny, protože se je nepodařilo opravit). Je však možné množství vzorků redukovat nejrůznějšími způsoby agregace. Využití plnění může mít pozitivní vliv na celkové množství zpracovávaných dat. [13]

Příprava pro strojové učení

Poté, co se analytikovi podaří získat čistá data, přichází na řadu fáze strojového učení. Existuje spousta metod, které slouží ke strojovému učení a rozpoznávání či predikci. Vzhledem k omezenému rozsahu práce není možné, aby byly všechny metody podrobně probrány. Přesto je vhodné, aby byly zmíněny postupy, které se pro analýzu nejčastěji používají. Těmito postupy mohou být regresní analýza, rozpoznávání neuronovými sítěmi, evoluční/generické algoritmy nebo jiné případy klasifikace. Jednotlivé metody se pak mohou částečně kombinovat pro zpřesnění výsledků.

3.5 Regresní analýza

Pod pojmem regresní analýza se skrývá celá škála metod, které slouží k předpovědi vhodné hodnoty Y na základě skupiny regresorů. Regresní analýzu je vhodné využít, pokud jsou na sebe hledané hodnoty závislé. [25] Příkladem takových hodnot může být počet autonehod za jeden rok na 10 000 obyvatel a vyspělost dané země. Je pravděpodobné, že vyspělejší země mohou více investovat do prostředků, jak autonehodám zabránit (například povinností technické kontroly na vozidle nebo kvalitnějšími a udržovanějšími silnicemi). Tyto hodnoty jsou na sebe závislé, a proto je vhodná regresní analýza.

Obecně je regresní analýza dána vztahem: [13]

$$E(Y|X_1, X_2, \dots, X_n) = f(X_1, X_2, \dots, X_n) \qquad (3.5.1)$$

Přesto je však vhodné pokusit se problém převést na konkrétnější typ regresní analýzy. Například lineární regresní analýza přináší mnohonásobně snadněji získatelné výsledky. [23] Metod pro regresní analýzu existuje mnoho, ale pro dodržení rozsahu práce, budou představeny jen některé příklady.

Lineární regresní analýza

Asi nejjednodušší regresní analýzou je lineární regresní analýza. Tato metoda se danými daty snaží proložit přímku. Pro připomenutí je rovnice přímky definována jako:

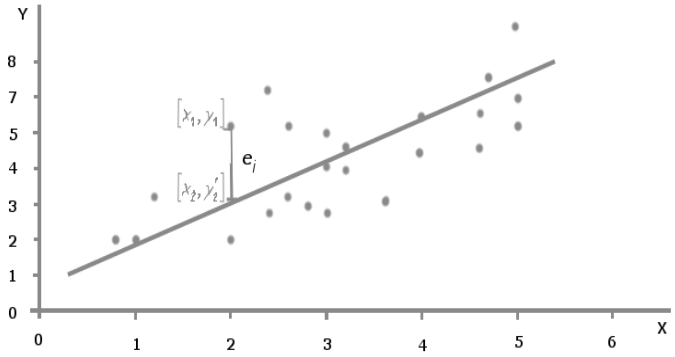
$$y_i' = a + bx_i \tag{3.5.2}$$

Cílem lineární regresní analýzy je, aby byl součet všech odchylek co nejnižší. Odchylka e_i se získá jako vzdálenost mezi hodnotou y_i a proloženou přímkou y_i':

$$e_i = y_i - y_i' \tag{3.5.3}$$

Lineární regresní analýza má tedy tvar: [23]

$$S_{rez} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_i')^2 \text{ ; takové že } S_{rez} \text{ je minimální} \tag{3.5.4}$$



Obrázek 3.5.1: Proložení přímky lineární regresní analýzou

Poté co je přímka proložena je vhodné zjistit, do jaké míry je přímka vhodná pro predikci dané řady. K tomu se využívá koeficient determinace: [23]

$$R^2 = \frac{S_{reg}}{S_{yy}} = 1 - \frac{S_{rez}}{S_{yy}} \tag{3.5.6}$$

Kde S_{reg} je regresní součet čtverců odchylek predikcí od průměru řady, čili:

$$S_{reg} = \sum_{i=1}^{n} (y_i' - \bar{y})^2 \tag{3.5.7}$$

A S_{yy} je součet čtverců datových hodnot od průměru řady:

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.5.8}$$

Koeficient determinace je tedy vyjádřen vztahem: [25]

$$R^2 = \frac{\sum_{i=1}^{n}(y_i' - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.5.9}$$

Je vhodné, aby se koeficient determinace blížil hodnotě 1, což naznačuje, že zvolená přímka je pro extrapolaci řady vhodná. Pokud se ovšem koeficient determinace blíží hodnotě 0, je vhodné použít jinou aproximující křivku.

Vícerozměrná regresní analýza

Jedná se o rozšíření lineární regresní analýzy o možnost pracovat s více než jednou předvídatelnou hodnotou, čili s více regresory. Vícerozměrná regresní analýza se dá také provést pomocí metody nejmenších čtverců (stejně jako lineární regresní analýza), ale výpočet je již časově náročnější (v závislosti na množství regresorů). Získání hodnoty Y pro n-rozměrnou regresní analýzu je dáno vztahem: [24]

$$Y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n; \text{kde } n \in \mathbb{N} \tag{3.5.10}$$

Nelineární regresní analýza

Ne vždy je potřeba danými body proložit právě přímku. Nelineární regresní analýza se tedy snaží danými body proložit křivku. Celý problém nelineární regresní analýzy pak spočívá v hledání takové křivky, jejíž koeficient determinace by se blížil hodnotě 1. [25] Křivka je pak obecně složitým polynomem a nalezení takového polynomu představuje velmi náročný výpočet. Hodnota Y pro n-tý stupeň polynomu se získá pomocí: [24]

$$Y = a + b_1x^1 + b_2x^2 + \cdots + b_nx^n; \text{kde } n \in \mathbb{N} \tag{3.5.11}$$

Některé nelineární polynomy se však dají transponovat na mnohonásobnou regresi (přesto je však problém odhalit potřebný stupeň polinomu):

$$\begin{gathered} Y = a + b_1x^1 + b_2x^2 + \cdots + b_nx^n; \text{kde} \\ n \in \mathbb{N}; \text{trans.}: x_1 = x, x_2 = x^2, \ldots, x_n = x^n \end{gathered} \tag{3.5.12}$$

$$\begin{gathered} Y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n; \text{kde} \\ n \in \mathbb{N}, \text{což je rovnice vícenásobné regrese} \end{gathered}$$

3.6 Neuronová síť

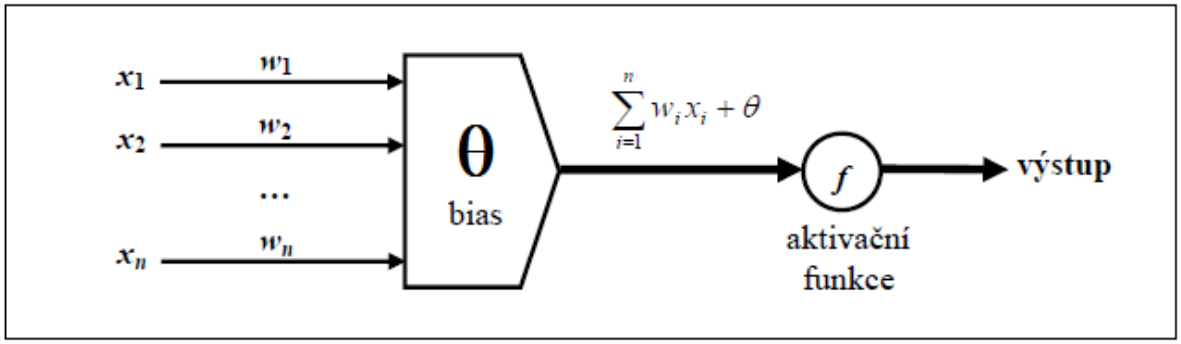
Neuronové sítě se velmi využívají v oblasti rozpoznávání a klasifikace. Pro snadné pochopení, bude rozebrán jeden příklad často užívané neuronové sítě, neboť kompletní výčet všech způsobů zapojení a nastavení neuronových sítí by překračoval rozsah práce. Jejich princip učení vychází z reálných neuronů nacházejících se v pokročilejších živých organizmech. [13] Neuronová síť je tvořena soustavou různě propojených neuronů (stejně jako neurony například v lidském mozku). Každý prvek této sítě se snaží simulovat reálný neuron.

Reálný biologický neuron je tvořen: [14]

- Soma – neboli tělo neuronu
- Dendritů – které představují vstupní rozhraní neuronu. Jedná se o krátké výběžky z těla neuronu, kterých může být od 10na2 až po 20na5
- Axonu – což je jediný výstupní výběžek z těla neuronu

Jednotlivé dendrity se s axony ostatních neuronů stýkají prostřednictvím synapsí. Vzájemně si přenášejí elektrické impulsy (pomocí uvolňování chemických látek). Tyto impulsy mohou fungovat jako excitátory (podporují, aby neuron vyslal axonem signál) nebo naopak jako inhibitory (snižují schopnost generovat na axonu signál). Jestli neuron vyšle signál, je pak dáno sumou energie všech synapsí, které má. Pokud celková energie překročí určitou hranici, emituje neuron na svém axonu elektrický signál. [14]

Biologický neuron se stal inspirací pro vznik umělého neuronu. Umělý neuron má N vstupů, které modifikuje určitou váhou (jestli budou extitátory nebo inhibitory neuronu). Pokud je suma všech vstupních hodnot vyšší než práh, tak umělý neuron bude emitovat výstupní hodnotu. Práh vlastně představuje posunutí (bias) od nulové hodnoty neuronu. [13] Pokud je bias nulový, neuron emituje signál, pokud je suma kladná, jinak nikoliv. Proces učení daného neuronu spočívá ve hledání ohodnocení vektoru vah w_i. [14]

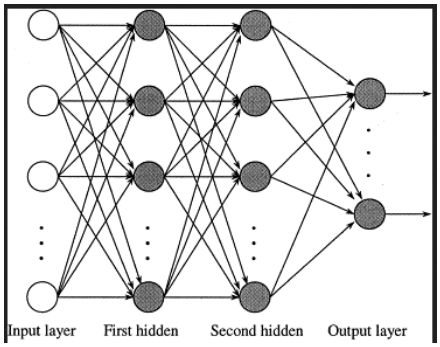


Obrázek 3.6.1: Umělý neuron[1]

Takovéto umělé neurony jsou většinou vystavěny do neuronové sítě. Obecně má neuronová sít vrstvu vstupní, skrytou vrstvu a výstupní. Na vstupní vrstvu je přiveden vzorek dat, která jsou určena ke klasifikaci. Vstupní data mohou být například skupinou svíček z cenové řady, přičemž každá svíčka obsahuje spoustu atributů. Skrytých vrstev může být v neuronové síti libovolný počet. Nakonec výstupní vrstva představuje odezvu systému (neuronové sítě) na daný vstupní vzorek dat. Protože se neuronové sítě využívají hlavně ke klasifikaci a rozpoznávání, odpovídá výstupní vzorek dané třídě, do které je klasifikován. [27] Pro klasifikaci burzovních číselných řad tak mohou klasifikované třídy být například: chop, způsobující profit, velký profit nebo ztrátu atd.

Je dobré si uvědomit, že jeden neuron je schopen rozeznávat pouze dva stavy (kdy emituje signál a kdy nikoliv). Prostor všech možných řešení je tak jedním neuronem rozdělen na dva podprostory. Právě pro zpřesnění dělení (a získání více kategorií) je nutné využít větší počet neuronů, které se dají propojit nejrůznějšími způsoby. Různé propojení například dovoluje rychlejší šíření informace do výstupní vrstvy. Najít nejvhodnější uspořádání a propojení neuronů může být specifické pro každou úlohu.

[1] Převzato z [13]



Obrázek 3.6.2: Příklad zapojení neuronové sítě[1]

Přestože možných způsobů propojení neuronové sítě (aktivačních funkcí, vah a prahů) může být téměř neomezené množství, je vhodné, aby se čtenář seznámil s konkrétnější verzí neuronových sítí. Čímž získá představu o způsobu fungování neuronových sítí:

Nejznámějším typem učení s využitím umělých neuronů je neuronová sít Backpropagation, která po procesu klasifikování vzorku dat šíří zpětně na všechny neurony chybu, která byla při klasifikaci vykonána. [27] Tato síť se využívá velmi často (asi kolem 80% všech případů strojového učení neuronovou sítí) a řeší se s ní nejrůznější problémy.

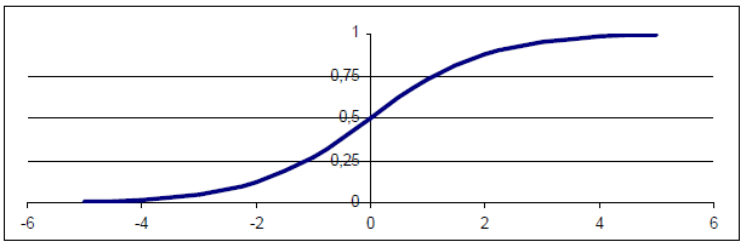
Proces učení neuronové sítě bývá inicializován náhodnými hodnotami w_i a biasy na všech neuronech. Poté učení probíhá ve dvou fázích, klasifikace a propagace chyby. Na vstupní vrstvu neuronové sítě je přiveden vzorek dat a neuronová sít provede klasifikaci daného vzorku. Na výstupní vrstvě se tak objeví výstupní vektor (který odpovídá klasifikované třídě dat). Tento výstupní vektor je porovnán s očekávaným vektorem. Pokud se vektory liší, dopustila se síť chyby a je nutné poupravit její váhy w_i. Míra, s jakou sít pozmění své aktuální váhy, se nazývá koeficient učení. Jedná se o reálné číslo, které se nachází v intervalu <0, 1>. [26] Pokud je koeficient učení příliš nízké číslo, síť se již téměř neučí, neboť není ochotná pozměnit své váhy (při koeficientu 0 nikdy nedojde ke změně váhy w_i). Naopak příliš vysoké číslo způsobí, že neuronová síť na svoji chybu reaguje příliš přehnaně. Sít tak pracuje v podstatě jen s aktuálním vzorkem dat a mnohem více ignoruje předchozí zpracování. Je proto důležité volit koeficient učení velmi pečlivě a jeho hledání bývá nesnadným problémem. [14] Je také vhodné zvážit možnost koeficient učení dynamicky zmenšovat, aby se síť z počátku rychle učila, ale ke konci učení, aby již byla velmi perzistentní. Příkladem dynamicky zmenšovaného koeficientu učení může být $1/t$. [13]

Kromě koeficientu učení je nutné ještě určit aktivační funkci neuronu. Existují dva základní typy aktivační funkce a to funkce spojitá a funkce skoková. Skoková aktivační funkce je taková, která popisuje, kdy ještě zůstane v klidu, a kdy již bude emitovat nějaký výstup. Dochází tedy ke skokové změně z žádného výstupu a nějaký. Naopak spojitá aktivační funkce může výstup emitovat s různou intenzitou, která je rovna výstupu dané aktivační funkce. U sítí Backpropagation se nejčastěji volí spojitá aktivační funkce: [13]

$$y = \frac{1}{1 + e^{-x}} \tag{3.6.1}$$

[1] Převzato z [13]

Jejíž graf odezvy vypadá následovně:



Obrázek 3.6.3: Odezva aktivační funkce y=1/(1+e^(-x))[1]

Neuronová síť tak může být použita k rozpoznání určité posloupnosti v cenové řadě a kategorizaci, kterým směrem se graf s danou posloupností vydá. Ona posloupnost pak představuje obchodní vzor. Poté, co neuronová sít provede klasifikaci aktuálního okna, se k ověření výstupní kategorie využije profitová funkce následujícího vývoje grafu. Profitová funkce tak představuje učitele, který koriguje odezvu neuronové sítě a napomáhá přiřazování hledaných kategorií profitu (velký profit, profit, chop, ztráta, atd.).

3.7 Evoluční algoritmy

Podobně jako neuronové sítě i evoluční neboli genetické algoritmy jsou velmi inspirovány děním v přírodě. Evoluční algoritmy simulují vývoj společnosti jedinců. Simulace vývoje společnosti je možné provádět mnoha způsoby, ale práce se musí omezit na několik příkladů, vhodných pro predikci číselných řad. Každý jedinec má dva rodiče a jeho genetická informace nějakým způsobem reflektuje jejich genetickou informaci. Jedinci, kteří jsou danému prostředí lépe přizpůsobeni, mají v přírodě větší šanci na přežití a tím i větší šanci na plození dalších potomků. [13]

Evoluční algoritmy fungují velmi podobně. Každý algoritmický problém má jistou množinu všech možných řešení (nebo přesněji kombinaci všech možných vstupů). Při inicializaci je vytvořena populace jedinců. Každý jedinec může například představovat jednu kombinaci vstupů daného problému. Populace není úplná, jedná se pouze o určitý (většinou náhodný) vzorek možných kombinací. [28]

Ohodnocení

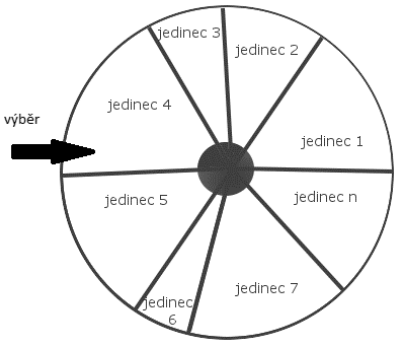
Každému jedinci je přiřazeno tzv. fitness ohodnocení, které říká jak moc kvalitní daný jedinec je. [29] Hodnotící funkce fitness je specifická pro každou úlohu a analytik ji musí vhodně odhalit. Podle ohodnocení fitness se totiž vybírají vhodní jedinci pro další křížení a rozmnožení populace, takže kdyby byla funkce zvolena špatně, algoritmus by nebyl schopen vracet kvalitní výsledky.

Selekce

Po ohodnocení celé populace, jsou vybrány nejvhodnější jedinci pro křížení a tvorbu nové populace. Samotných principů výběru vhodných jedinců existuje mnoho a je na tvůrci programu, kterou strategii zvolí. Práce se pokusí nastínit pouze několik nejčastějších postupů:

- **Roulette wheel selection** – proces výběru jedince odpovídá principu rulety. Každý jedinec zabírá určitou část rulety dle svého fitness ohodnocení. [28] Daná metoda je velmi vhodná, pokud je ohodnocení jedinců velmi podobné. [29]

[1] Převzato z [13]

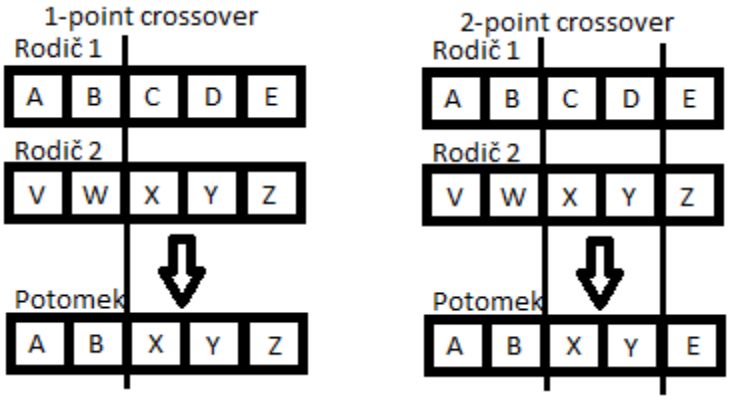


Obrázek 3.7.1: Vizualizace roulette wheel selection

- **Rank selection** – předchozí metoda rulety má velkou nevýhodu, pokud se v populaci nachází velmi malé procento jedinců, kteří ale mají mnohonásobně vyšší fitness ohodnocení než jedinci ostatní. V takovém případě jsou ke křížení vždy vybrání právě tito jedinci a zbytek populace vyhyne. V této metodě se jedinci postaví do pomyslné řady a pak se pro křížení vybírá náhodná pozice v řadě. Nevýhodou této metody je ale jisté nerespektování fitness ohodnocení. [28]
- **Tournament selection** – tato metoda řeší problém rank selection způsobu, protože se z populace vybere určitý počet náhodných jedinců a z nich je ke křížení vybrán ten s nejvyšším ohodnocením fitness. Jedinci s vysokým fitness tak vždy vyhrají, ale slabší jedinci mají mnohem vyšší šanci k rozmnožení (když nebyly nejsilnější jedinci do tournament skupiny vybráni. [28]

Křížení

Nová populace vzniká kombinací genů (parametrů vstupu) svých rodičů. Opět je mnoho způsobů, jakým mohou být geny zkombinovány. Aritmetický průměr hodnot jednotlivých genů, není vhodnou metodou, protože by se rozmanitost populačního vzorku neustále přibližovala, až by nakonec vznikl jediný prvek. [29] Evoluce by tak mohla velmi snadno dosáhnout nějakého lokálního maxima. Mnohem častěji se využívají různé poměry křížení jednotlivých genů (část od jednoho rodiče a část od druhého). Tato metoda se jmenuje crossover a ilustruje ji následující obrázek: [28]



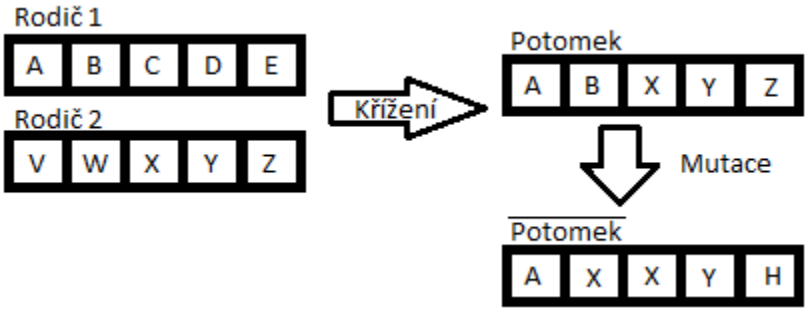
Obrázek 3.7.2: Příklad crossover křížení genetické informace jedinců

Mutace

Mutace zavádí do procesu křížení nejrůznější odchylky a napomáhá tak k vytvoření geneticky zcela nových jedinců, kteří by teoreticky mohly být vhodnější než samotná kombinace vlastností jejich

rodičů. [29] Míra mutace velmi výrazně ovlivňuje celý proces evoluce. Pokud by jednotlivé geny mutovaly příliš často, celý algoritmus by se mohl proměnit v náhodné hledání, neboť by při křížení vznikaly potomci, kteří by vůbec nebyly podobní svým rodičům. [30]

Přesto je vhodné jistou míru mutace umožnovat a to proto, aby nebyl výsledek algoritmu závislí na kvalitě prvotního rozložení populace.



Obrázek 3.7.3: Příklad mutace

Při křížení rodičů $Rodič_1$ a $Rodič_2$ měl vzniknout potomek $Potomek$, ale díky mutaci vznikl mírně vzdálený potomek $\overline{Potomek}$. Vzdálenost mezi $Potomek$ a $\overline{Potomek}$ závisí na vlastnostech mutace. Ta může pozměnit různý počet genů a může gen oddálit různou měrou od původního dle své síly.

Je možné využívat proměnlivou mutaci, kdy se pravděpodobnost a rozsah mutace zvýší, pokud si jsou jednotlivý jedinci geneticky velmi podobní. Takovéto řešení je velmi vhodné, protože může pomoci evolučnímu algoritmu před uváznutím v lokálním maximu. [28] Většina populace si v lokálním maximu bude velmi příbuzná, a proto je vhodné zesílit mutaci, aby zde byla šance nalezení lepších fitness jedinců.

Nalezení vhodné pravděpodobnosti a síly mutace je nesnadný úkol a pro různé řešení mohou optimální hodnoty ležet jinde.

Eliminace

Po vzniku nové populace následuje poslední část evoluce a tou je vyhynutí slabých jedinců. [30] Některé postupy nechají vyhynout celou starou generaci, jiné jen určitou procentuální část, aby bylo možné provést křížení i se starými jedinci (toto řešení ovšem vyžaduje vhodně zvolenou metodu pro výběr jedinců pro křížení, aby jedinci s nižší fitness hodnotou měli šanci k vytvoření nové populace). [28]

Takto se celý proces neustále opakuje. Evoluce může probíhat určitý počet cyklů, nebo dokud není nalezeno nejlepší řešení či do uváznutí v nějakém lokálním maximu.

Pro zpracování cenových řad na burze by mohli jednotlivý jedinci představovat vzory pracující s nějakým oknem svíček a indikátorů a jejich fitness funkcí by byl celkový profit, kterého jsou schopny dosáhnout. Evolucí by tak postupně vznikly pouze zajímavé vzory dosahující kvalitních profitů.

3.8 Vyhodnocení

Poté co je proces učení dokončen je vhodné otestovat správnost systému ještě před reálným nasazením. Postupy, jak vyhodnotit správnost strojového učení, závisí na specifikaci problému, ke kterému bylo strojové učení využito. Je nutné, aby se tvůrce programu zamyslel, nad specifikací problému a podle ní zvolit vhodné metody. Pro vyhodnocení cenových řad je vhodné rozdělit vstupní

množiny dat na data určená pro učení (tzv. trénovací množina) a na množinu testovacích dat (testovací množina). Strojové učení probíhá pouze nad trénovacími daty. Když je učení dokončeno, je systém spuštěn nad testovacími daty, u kterých jsou ještě známi správné výsledky. Tím může být srovnána odezva systému s očekávanými daty. [17] Je získána jistá charakteristika systému, která analytikovi udává, jak je systém přesný, neboli s jakou pravděpodobností se na výsledky systému může spoléhat.

Pro jednotlivé výsledky strojového učení jsou pomocí testovací množiny získány pravděpodobnosti úspěšnosti. Jednou ze závěrečných činností, kterou by měl analytik vykonat před nasazením systému je zhodnocení statistiky jednotlivých výsledků. Dle vlastních zkušeností dané problematiky zhodnotit, zdali jsou výsledky použitelné. Je vhodné stanovit nějaké metriky, které budou sloužit k rozpoznání kvalitních výsledků strojového učení.

Při využití strojového učení nad cenovými řadami se analytik snaží najít vzory, které by mu signalizovaly, jak se zachovat na určitou posloupnost svíček. I k těmto vzorům je možné už v průběhu učení získat spoustu statistik, které pomohou odhalit vlastnosti budoucího systému.

Mezi nejdůležitější charakteristiky vzorů patří: potenciální celkový profit, průměrný profit na jeden kontrakt, průměrný profit za určité časové období, celkový počet výskytů vzoru a průměrný počet výskytů za určité období (pokud se vzor vyskytuje pouze několikrát, je vhodné jej ignorovat) a pravděpodobnost správného odhadu (v kolika případech vzor predikoval správnou akci na danou posloupnost cenové řady).

Jednou z dalších sledovaných vlastností je pak rozmanitost a počet vzorů. Obchodní systém nepotřebuje obsahovat stovky různých pravidel, neboť by s nimi spekulanti nedokázali pracovat. [7] Je proto vhodné vybrat pouze zlomek těch nejlepších vzorů, které se vyskytují dostatečně často a přinesou očekávaný profit. Stejně tak pokud by každý vzor pracoval s jinými indikátory, byly by cenové grafy tak nepřehledné, že by v nich vzory spekulant nestíhal rozpoznávat.

Jelikož každý spekulant může mít jinou strategii a každý má různě odolnou psychiku, je v podstatě nemožné definovat jednotné metriky. Z tohoto důvodu nelze stanovit, jaké vzory jsou kvalitní a které ne. Jistě se všichni spekulanti shodnou, že vzor by měl generovat profit, ale jeho výše už může být těžce specifikovatelná. Pro někoho může být dostatečný zisk již 10$ na kontrakt (což může představovat posun o půl bodu), jinému tak nízký profit nestojí za námahu a zajímají jej pouze třímístné částky. Stejné konflikty vznikají i při samotné definici predikce. Většina lidí by se jistě shodla, že predikce s úspěšností 40% znamená, že program spíše nepredikuje. Jenže je nutné se dívat na všechny vlastnosti vzoru a posuzovat jej jako celek. Pokud vzor v 40% dokáže správně predikovat zisk 100$ a v 60% nesprávnou predikcí způsobí ztrátu 10$ tak na sto kontraktech je celkový profit daného vzoru 3400$, což už naopak většina lidí za správnou predikci považovat bude. Jednotlivé vlastnosti vzorů spolu velmi úzce souvisí, a tak by vzory neměly být posuzovány jen podle jediné. [12] Je tedy nutné, aby se každý budoucí uživatel takového systému zamyslel nad tím, jestli je pro něj nalezený seznam vzorů vhodný či nikoliv.

4 Analýza a návrh řešení

V této kapitole je čtenáři představeno shrnutí současného stavu predikce cenových řad. Po prostudování mnohých obchodních diskuzí (a díky přátelskému vztahu některých reálně obchodujících spekulantů) se autor pokusí specifikovat, jaké problémy spekulanti nejčastěji řeší, a jak by jim autorův program mohl usnadnit práci. Na základě zmíněných postřehů, pak bude specifikován program pro tvorbu obchodních systémů, který by současným a začínajícím spekulantům mohl v jejich práci pomoci.

4.1 Vyhodnocení současného stavu

V dnešní době existuje mnoho nejrůznějších programů, které se chlubí schopností nalézt ten nejvhodnější obchodní systém.

Hlavním problémem těchto programů je velmi vysoká cena a uzavřený kód. Uživatel programu nemá jak si ověřit, jestli je vytvořený systém dostatečně obecný. Programy naleznou nějaké vzory, hrozí ovšem, že budou příliš přetrénované. Uživatel nemá jistotu, jestli vůbec byla využita nějaká testovací množina, a potom hrozí, že vzory byly programem ušity přesně na míru daným datům a nad reálnými daty již nebudou fungovat. Internet je plný nejrůznějších prográmků, které se chlubí 99% úspěšností, přesto o ně zkušení obchodníci nemají zájem (názor autora práce, který vychází z mnoha diskuzí na toto téma). Zkušený obchodník ví, že zázračné systémy neexistují. Nepodařilo se je sestavit za 70 let, a nyní není situace o moc lepší.

Nicméně i pokud se spekulantovi narazit na sytém, jehož parametry již vypadají reálně, nemůže si být plně jist, zdali je obecnost dostatečná. Pokud pak dojde k několika ztrátovým obchodům, jsou pochyby největším problémem. Pokud spekulant svému systému nevěří, častokrát jej ve ztrátovém období změní, a to vede jen k větším ztrátám. [11] Víra majitele systému ve svůj sytém musí být velmi vysoká, aby pomohla překonat ztrátová období (existují obchodní systémy, které mají statisticky přijatelné ztrátové období i 25 obchodů v řadě. Takovou míru proher, většina spekulantů psychicky nedokáže ustát. Každý obchod může představovat ztrátu vyšší než 1000 korun, a při 25 obchodech za sebou, již může být obchodníkův kapitál nepříjemně poškozen). [12]

Protože je tak důležité, aby spekulant mohl věřit systému, který používá, je vhodné, aby systém měl otevřený zdrojový kód, aby si jej mohl každý prohlédnout a ověřit, že se například využívají testovací množiny. Tato práce se snaží přispět právě k této množině systémů, které vysvětlují veškeré základy a postupy a mohou být využity za odrazový můstek pro všechny analytiky, spekulanty nebo obchodníky.

Práce se také snaží podpořit začínající analytiky a spekulanty, neboť nebudou muset investovat do velmi nákladných systémů v době, kdy ještě nejsou schopní generovat zisk. Text umožňuje, aby se čtenář seznámil s použitelnými a důležitými principy, které se pro analýzu na burze využívají, a zároveň dodává nástroj, který implementuje základní vyhledávač vhodných vzorů, které jsou schopny generovat zisk.

Právě hledání vzorů, je neustále se opakující činnost, neboť mnoho prvků, které dříve fungovaly, dnes již nefungují (nebo profit je tak malý, že se nepoužívají). Jak se vyvíjí znalosti a strategie jednotlivých účastníků, musí se vyvíjet a měnit i obchodní systémy. [5] Proto neexistuje jeden univerzální systém. I kdyby se jej podařilo stvořit, jeho efektivita by byla jen velmi krátká.

Je proto nutné obchodní systémy a vzory pro predikci měnit a zpřesňovat a to podle toho, jak se trh vyvíjí (tyto aktualizace většinou bývají prováděny ročně, nebo i s větším intervalem). Je tedy

vhodné, aby spekulanti měli možnost tuto změnu učinit pomocí nějakého automatického programu. Nebudou tak muset každé tři roky zdlouhavě přezkoumávat, jak se vyvinula strategie trhu jako takového. Proto je pro ně výhodné využívat strojového učení. Při vytvoření programu, který bude historická data analyzovat sám, a sám bude vyhledávat nejvýhodnější strategii obchodování, bude možné využívat vždy dostatečně aktuální obchodní systém.

Jelikož práce představila mnoho způsobů pro predikci cenových řad, pokusí se autor jednotlivé postupy přehledně shrnout. Jednotlivé vlastnosti budou vztaženy k cenovým řadám, které se na burze vyskytují a potřebám spekulantů, kteří by s metodami pracovali.

	rozsáhlá data	data se šumem	krátká doba učení	trendující řada	sezoní složka	výrazná nahodilá složka	predikce mnoha kategorií	simulace reálných postupů spekulanty	vhodná pro cennové řady
Vyhlazování řad	x	x	x	x				x	
spektrální analýza					x				
kalmanův filtr	x	x	x	x					
dekompozice řady		x		x	x	x			
sezónní diference	x	x			x			x	
Lineární regresní analýza	x	x	x	x					
Vícerozměrná regresní analýza	x	x	x	x	x				
Nelineární regresní analýza		x		x	x				
Neuronová síť	x			x	x	x		x	x
Evoluční algoritmy	x	x		x	x	x	x	x	x

Tabulka 4.1.1: Srovnání efektivity různých metod pro predikci cenových řad

Daná tabulka nemůže být brána zcela dogmaticky, protože jednotlivé vlastnosti je nutné vztáhnout k cenovým řadám na burze a jejich charakteristice. Například spektrální analýza může být vhodná pro rozsáhlá data, ale pokud jsou data zatížena náhodnou složkou, stane se spektrální analýza téměř nepoužitelná. Náhodná složka je však u kratších timeframe velmi výrazná. Neuronová síť je sice schopna rozeznávat velké množství tříd, ale s přibývajícími třídami musí přibývat i počet neuronů a celkové učení se může zpomalovat. Pokud například data vykazují v určitém svém úseku nestandartní chování, znalost neuronové síť se pro daný úsek bude zhoršovat, zatímco evoluční algoritmy vytvoří nový vzor, který ale bude mít špatné vlastnosti. Kalmanův filtr může být vhodný pro predikci cenové řady, problém však je rozsah predikce. Spekulant nepotřebuje znát jen následující svíčku (poplatek za zprostředkování obchodu by byl vyšší než zisk), ale systém mu musí zaručovat dlouhodobější predikci pro následující svíčky. Jednotlivé metody mohou být pro predikci číselných řad vhodné, ale pro potřeby spekulantů nad cenovými řadami z burzy již být vhodné nemusí.

Na základě vyjmenovaných vlastností, se autor rozhodl pro implementaci genetického algoritmu. K ohodnocení vzorů se využije fitness funkce představující profit. K selekci vhodných jedinců pak strategie tournamentu.

4.2 Návrh systému

Jelikož předchozí kapitoly objasnily veškerou potřebnou problematiku, je vhodné specifikovat, jak přesně budou implementační práce probíhat. Autorovy znalosti pro programovací jazyk python, výrazně převyšují znalosti ostatních jazyků (autor se již několik let v jazyce aktivně vyvíjí), je i výsledný systém navržen a naprogramován v jazyce Python.

Protože autor nezískal žádný grant, a musel se spolehnout výhradně na své zdroje, nebylo pro něj možné získat kvalitní data od IQFeed (tím se nepodařilo sehnat ani dostatečně dlouhou historii TICK dat), bude celý program pracovat s minimálně minutovými časovými intervaly. Program by měl být pro strojové učení obecný, a tak až bude plně dokončen, bude možné vyměnit datový vstup za kvalitnější a přesnější, nastavit interval například na 1 sekundu a program by měl být schopen dodat nalezené vzory pro daný časový rámec.

Jelikož mnoho spekulantů a analytiků pro obchodování na burze potřebuje využívat nějaký obchodní systém, pokusí se program vyhledat z historických dat takové vzory, které jsou pro obchodní systém použitelné, aby si všichni mohli sestavit takový systém, který jim bude vyhovovat. Někteří lidé, mají raději agresivnější, jiní defenzivnější strategii, pomocí nejrůznějších parametrů, by měl být výsledek natolik modifikovatelný, aby vyhovoval právě tomu, kdo jej bude používat.

Specifikace pro implementaci:

- Program bude napsán v jazyce Python.
- Bude se jednat o konzolový skript (grafické rozhraní není v prvotní fázi důležité).
- Ovládání bude uskutečněno přes parametry programu
- Zpracovávání se musí vypořádat s mnoha vstupními cenovými řadami (komodita, kontaktní měsíc, rok). Bude je zpracovávat hromadně anebo je sloučí.
- Program by měl být schopen zpracovat zhruba 20 let vteřinové historie (cca 0,5 miliardy záznamů)
- Učení musí být parametrizovatelné, aby odpovídalo strategii spekulanta, který jej chce používat (timeframe, stop loss, profit target).
- Bude využit genetický algoritmus učení
- Jedna iterace algoritmu učení, by měla mít lineární složitost (projít datový vstup jen jednou)
- Program musí mít parametr, kterým se zapne vizualizace zpracovávání, aby uživatel (většinou ne-technicky vzdělaný spekulant) viděl, že průběh výpočtu.
- Program bude obsahovat validační jednotku, která sdělí, jaké jsou statistické vlastnosti navrhovaného systému, aby se mohl uživatel rozhodnout, jestli chce s daným systémem pracovat.
- Bude umožnovat hledat vhodné nastavení parametrů (timeframe, stop loss, profit target).
- Výstupní nalezené vzory systému budou vypsány do souboru, aby s nimi mohla být provedena pozdější simulace obchodování (či jiné optimalizace)

V úvodní části, neboli v prvotní verzi, se bude jednat o konzolovou aplikaci. Důležité je stvořit program, který bude schopen dodat vše potřebné pro stvoření obchodního systému. Přestože by aplikaci měli používat netechnicky vzdělaní uživatelé, tak grafické rozhraní bude z důvodu omezeného času odloženo až na budoucí rozvoj. Ostatní zmíněné vlastnosti pomohou systému, aby pracoval dle preferencí daného uživatele.

5 Implementace

Tato část práce se věnuje samotné implementaci. Po nastudování veškeré teorie, jak o strojovém učení, tak o burze samotné, je nutné daný program sestavit. Jednotlivé podkapitoly vysvětlují celý proces od získání dat až po výsledný program a jeho otestování. Při práci bylo využito všech znalostí, které byly zmiňovány v předchozích kapitolách.

5.1 Strojové učení

Celý program je sestaven z několika menších skriptů, které zodpovídají za různé části zpracování. Od úvodní analýzy dat, jejich transformaci a čištění až po proces učení a samotného otestování vzorů.

Jelikož pro vyhledávání vzorů, není nutné využívat grafického rozhraní, je celý program spustitelný v terminálu počítače a veškeré výstupy jsou textového charakteru. Aby se uživateli se sadou skriptů pohodlně pracovalo, byl vytvořen skript, který slouží k postupnému provázání jednotlivých skriptů.

```
$ bash run_proccess.sh INPUT_DIRECTORY OUTPUT_DIRECTORY \
KOMODITY TIMEFRAME PROFIT STOPLOSS WINDOW WAIT BACK VISUAL
```

Vstupem pro skript je složka data, ve které se nacházejí jednotlivé cenové řady pro různé komodity od programu Sierra Chart. Tyto cenové řady, nejsou sjednoceny, a tak existuje jedna řada pro každý kontraktní měsíc daného roku.

```
$ ls data
CH09  CH12  CK11  CN10  CU09  CU12  CZ11  YMH10  YMM10  YMU09  YMU12  YMZ11
CH10  CK09  CK12  CN11  CU10  CZ09  CZ12  YMH11  YMM11  YMU10  YMZ09  YMZ12
CH11  CK10  CN09  CN12  CU11  CZ10  YMH09  YMM09  YMM12  YMU11  YMZ10
```

Cenové řady pro kukuřici a indexový trh v letech 2009 až 2012 s různými kontaktními měsíci

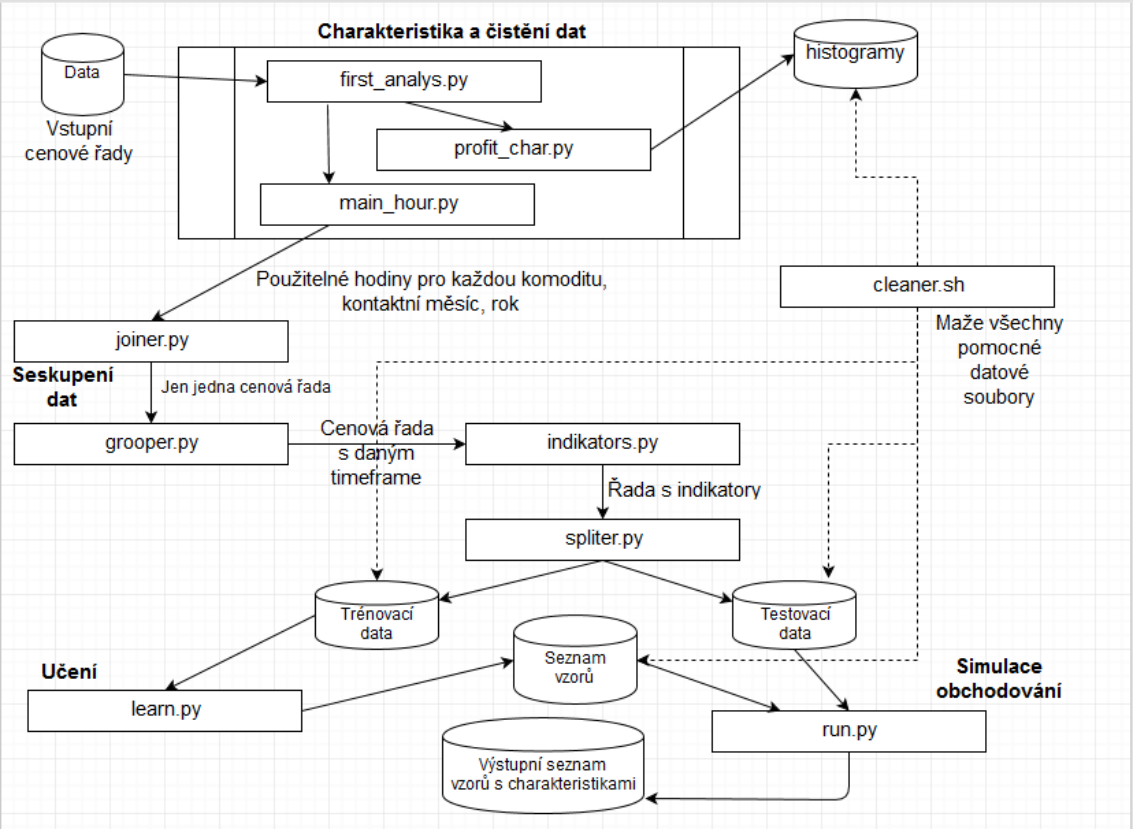
Skript provede všechny zmíněné úkony strojového učení a výstupem programu je seznam použitelných vzorů i s jejich charakteristikami. Jednotlivé záznamy pak jsou:

- **Akce** – udává směr, kterým vzor predikuje vývoj řady. Podle směru může spekulant reagovat otevřením pozice buď pro nákup (znak U), nebo pro prodej (znak D).
- **kódové označení vzoru** – označuje pozice jednotlivých indikátorů a svíček.
- **celkový profit** – kterého vzor na vstupních datech při simulaci dosáhl.
- **úspěšnost** – celková procentuální úspěšnost, kdy vzor predikoval vývoj správným směrem. Pokud se řada vyvíjela správným směrem až se zpožděním, a bylo dosaženo stop lossu, je predikce považována za neúspěšnou, byť by s vyšším stop losem úspěšná byla.
- **četnost** – Jedná se o počet výskytů daného vzoru v testovacích datech. Tento údaj slouží k odfiltrování statistické odchylky, neboť při nízké četnosti je vzor zanedbán

Akce-vzor	Profit	Četnost	Úspěšnost	Prům.profit
D-2,2,5,2,2,2,2,2,1,2	653.0	77	72 %	8
U-4,3,5,3,5,3,5,3,2,3	646.0	63	69 %	10
U-5,2,2,2,3,2,5,2,2,2	595.0	17	82 %	35
D-2,3,5,3,5,3,5,3,5,3	505.0	40	81 %	13

Příklad vzorů nad komoditou YM.
U/D znamená směr, kód vzoru (dle postavení indikátorů), celkový profit, četnost výskytu, úspěšnost a průměrný profit

Tento skript postupně spouští jednotlivé části dle následujícího schématu:



Obrázek 5.1.1: Schéma celého programu

Jak naznačuje schéma, nejdříve je nad vstupem provedena charakteristika dat. Poté jsou vybrány pouze obchodní hodiny s dostatečnou frekvencí. Vybrané cenové řady jsou sjednoceny, aby neustále reflektovaly pouze nejbližší kontraktní měsíc. Spojená cenová řada je pak agregována do svíček o určitém timeframe. Následně jsou k ní přidány indikátory a je proveden proces učení a vyhledávání vhodných vzorů. Na závěr je s danými vzory provedena simulace, aby se získala charakteristika těchto vzorů. Následující text, vysvětlí jednotlivé úkony podrobněji.

Charakteristika dat

Jedná se o sérii skriptů, které z dat získají hlavní charakteristiky a vykreslí histogramy, které pomáhají k pochopení jednotlivých vlastností dat. Pomocí těchto vlastností pak mohou být nastaveny parametry, se kterými obchodními hodinami se má vlastně pracovat.

Jako první je nutné spustit skript first_analys.py, který převede specifický datový vstup programu Sierra Chart na obecný. V případě, že by byl využit jiný datový zdroj s jiným formátem, musel by tento skript být nahrazen (nebo rozšířen). Spuštění:

```
$ python first_analys.py -xk data/YM, -xf 1990 -xt 2016 -xo _c
```

Protože datový vstup ještě nebyl sjednocen (je zde mnoho cenových řad, k různým kontaktním měsícům) je nutné zadat jaká komodita má být použita, a od kterého do kterého roku. Program pak sám zpracuje všechny datové vstupy, které se dané komodity a období týkají. K výstupním souborům je přidána koncovka _c. Program pak zpracuje všechny soubory a ke každému vytvoří histogram frekvence obchodů, ze kterého se dají odhadnout obchodní hodiny.

Dalším pomocným skriptem je profit_char.py, který slouží k odhadnutí, jestli je daná komodita vůbec vhodná pro obchodní zpracování. Daný skript není pro strojové učení nutný, ale slouží k odhalení komodit se slabým potenciálem, takové pak vůbec nemusí být zpracovávány.

```
$ python profit_char.py -a -xk data/YM,_c -xf 1990 -xt 2016 -o 13 -c 21 -d data
```

Čištění dat

Poté co jsou o komoditě získány hlavní charakteristiky a je rozhodnuto, jestli se s ní bude pracovat. Je nutné z datového vstupu odstranit všechna chybějící data (automatickým doplněním pravděpodobných hodnot) a očistit vstup tak, aby byl vhodný pro zpracování.

```
$ python clearer.py -xk data/YM,_c -xf 1990 -xt 2016 -mo 13 -mc 21 -o 8 -c 22 -t 100 -cd -oh
```

Jak je vidět i proces čištění dat, je prováděn nad všemi možnými vstupními soubory. Při čištění jsou odstraněny například víkendy, a tak je cenová řada pro sloučení o něco menší

Sjednocení cenových řad

Skript joiner.py sjednotí veškeré cenové řady dané komodity do jednoho datového vstupu, se kterým se pak mnohem lépe pracuje. Výstupní cenová řada pracuje pouze s nejaktivnějšími měsíci a proto je optimální pro strojové učení (je v ní výrazně méně chop pohybů).

```
$ python joiner.py -k data/YM,_c_fill -f 1990 -t 2016 -o data/YM_all -p 60 -dt sierra
```

Agregace dle timeframe

Různé strategie probíhají nad různým timerame. Je vhodné, aby byly jednotlivé svíčkové grafy agregovány a vytvořeny ještě před procesem učení, který se bude soustředit pouze na vyhledávání vzorů v preferovaném timeframe (nemělo by smysl provádět učení nad hodinovými grafy, když by je spekulant nechtěl obchodovat).

Další skript, následně přidá nejčastěji používané indikátory, aby se při strojovém učení nemuseli neustále přepočítávat (například výpočet rekurzivně volané EMA205 by se na každé svíčce již projevil).

```
$ TIMEFRAME=5m
$ python grooper.py -f data/YM_all -o data/YM_allG_$TIMEFRAME -t $TIMEFRAME
$ python indicators.py -f data/YM_allG_$TIMEFRAME -o data/YM_allI_$TIMEFRAME
```

Vytvoření trénovací a testovací množiny

Jedná se vskutku o jednoduchý skript, který pouze rozdělí datový vstup v daném poměru na trénovací a testovací množinu. Poměr však není získán jen dle počtu záznamů, ale dle počtu dní. Je totiž velmi vhodné, aby data nebyla rozdělena uprostřed dne. Tento uměle vytvořený gap by zbytečně znehodnocoval datový vstup. Nejdůležitější je parametr *–p*, který stanovuje, v jakém poměru se mají data rozdělit. *–p 60* říká, že 60% dní z datového vstupu připadne do trénovací množiny a 40% do testovací.

```
$ python spliter.py -f data/YM_allI_$TIMEFRAME  -p 60 -t indi \
-om data/YM_allI_${TIMEFRAME}_learn -ot data/YM_allI_${TIMEFRAME}_test
```

Učení

Pro jednotlivé timeframy je spuštěn algoritmus učení. Skript aplikuje evoluční algoritmus, který se snaží najít ty nejefektivnější vzory nad danými daty. Pro každý datový vstup, který odpovídá určitému počtu svíček z grafu, je programem rozpoznána kategorie vzoru. K danému vstupu se pak programem určí potenciální profit daného vstupu a o jeho hodnotu se rozšíří statistika dané kategorie

vzoru. Program v každé fázi získá populaci možných vzorů, které mají fitness ohodnocení dle svých potenciálních profitů.

Jelikož se vzory mohou vzájemně překrývat (pracují s podobným časovým oknem) může být provedena simulace obchodování, která překryté vzory eliminuje (problémem je pak vyšší časová náročnost běhu učení).

Slabá populace jedinců v následujícím kroku zaniká. Jedná se o jedince, kteří nedosáhli dostatečného výskytu nebo požadovaného profitu či úspěšnosti. Ze zbývající populace pak vzniká populace nových potenciálních vzorů.

Pro vznik nové populace jsou využity dva principy. Jedním je samotné crossover křížení vlastností daných vzorů. U tohoto křížení se pak využívá i přítomnosti mutace, která přináší vyšší stupeň rozmanitosti daných vzorů.

Druhou metodou je zobecnění daných vzorů. Pokud jsou dva jedinci podobní, bude vytvořen nový jedinec, který bude benevolentní k rozdílným vlastnostem rodičů. Vznikají tak obecnější vzory, které nemusí kontrolovat tolik indikátorů (nebo je nemusí kontrolovat tak striktně) jako jejich rodiče. Pokud bylo sloučení vzorů dobrým krokem a vzor získá v příštím běhu vysoké fitness ohodnocení, rodiče zanikají, protože jsou již zbyteční (jejich vlastnosti pokrývá potomek). Naopak při vzniku příliš obecného vzoru hrozí, že jeho prediktivní vlastnosti klesnou (v extrémním případě by vzor již nemusel kontrolovat nic). Při poklesu schopnosti predikovat, je zrušen vytvořený potomek a jeho rodiče jsou zachování (jejich rozdílnost je tedy pro predikci důležitá). Samotný vzor potomka je pak přidán na listinu tzv. eliminovaných vzorů, aby tento vzor nevznikal v dalším běhu programu.

Program nechává v každém běhu vyhynout více jedinců, než jich je vytvořeno. Počet potenciálních vzorů tak neustále klesá, až zbyde jen populace těch nejsilnějších a nejlepších vzorů. S těmito nejlepšími vzory je pak provedena simulace reálného obchodování, aby se získaly pro jednotlivé vzory úplné statistiky (je získána statistika profitu daného vzoru, celková úspěšnost predikce a počet výskytů vzoru). Tím program vytvoří vzory pro daný timeframe a tyto vzory se mohou stát základem pro obchodní systém spekulanta.

Samotný proces ohodnocení (simulace potenciálního profitu) se dá parametrizovat dle potřeb každého spekulanta. Jsou zde hodnoty jako stop-loss, profit-target, waiting-time atd. Pro plný rozsah parametrů a jejich funkčnosti je možné využít přiloženého RAEADME souboru, který je dodán společně s celým programem.

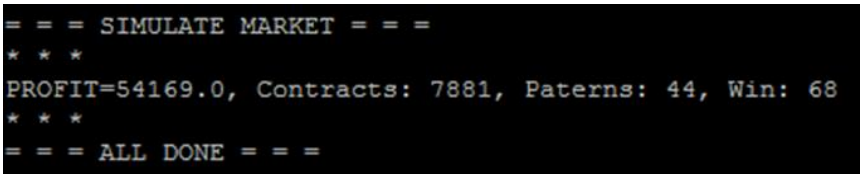
```
$ python learn.py -f data/YM_allI_${TIMEFRAME}_learn -o data/YM_allI_${TIMEFRAME}_learn_rule
  -wt WAIT_TIME -sl STOP_LOSS -ws WINDOW_SIZE -b BACK -p PROFIT -mp MIN_PROFIT
```

Simulace obchodování

Tento skript provádí simulaci reálného obchodování. Při simulaci již nejsou uvažovány žádné budoucí hodnoty cenové řady. Skript se musí spustit se stejným nastavením jednotlivých proměnných, jako proces učení:

```
$ python trade_symulation.py  -f data/YM_allI_${TIMEFRAME}_learn
   -r data/YM_allI_${TIMEFRAME}_learn_rule  -o data/YM_allI_${TIMEFRAME}_market_rule
   -wt WAIT_TIME -sl STOP_LOSS -ws WINDOW_SIZE -b BACK -p PROFIT -mp MIN_PROFIT
```

Po provedení simulace nad datovým vstupem skript vypíše charakteristiku získaného systému. Je vypsán celkový profit (pokud je systém vůbec schopen dosáhnout profitu), dále pak celková úspěšnost všech vzorů a kolik kontraktů bylo při simulaci celkem uzavřeno.



Výstup simulace obchodování. Zobrazí celkový profit v bodech, kolik bylo uzavřeno kontraktů, s kolika vzory a jaká je celková úspěšnost pro predikci

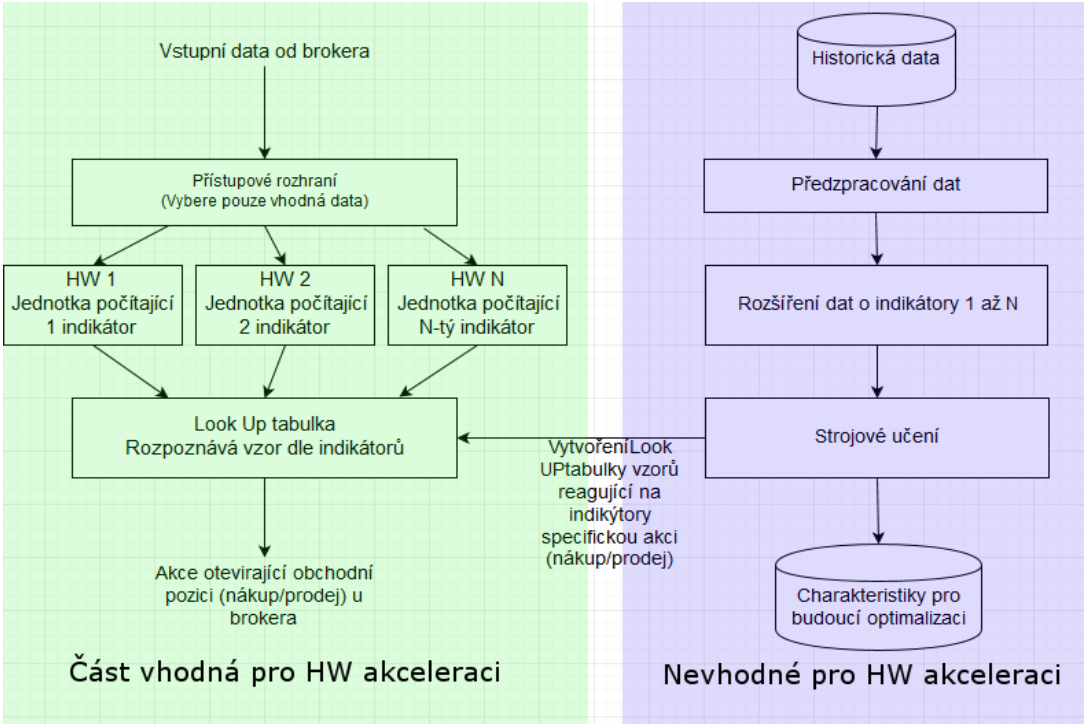
Skript také vytvoří výstupní soubor (*dle parametru –o*), který obsahuje všechny použité vzory s jejich podrobnou charakteristikou. Každý vzor je opět charakterizován počtem provedených kontraktů, pravděpodobností správné predikce, jestli se jedná o akci prodeje nebo nákupu a samozřejmě celkovým profitem. Analytik tak může vyloučit vzory, které se mu zdají příliš nevýrazné nebo riskantní. Jednotlivé vzory jsou vyznačeny svojí charakteristikou. Jedná se o kódování, které reflektuje vzájemné postavení jednotlivých svíček a indikátorů.

5.2 Možná hardwarová akcelerace

Jak bylo naznačeno již v úvodní kapitole, rychlost zpracování informací je neustále důležitější. Burzovní domy zpracovávají jednotlivé obchodní kontrakty v takovém pořadí, v jakém dorazili. Pokud je spekulantova reakce rychlá, vyvaruje se skluzu v plnění.

Přestože je rychlost důležitá pro rozpoznání vzoru a komunikaci s burzovním domem, samotný proces učení již může trvat delší čas, neboť probíhá nad historickými daty. Strojové učení je časově i paměťově náročné a nebylo by vhodné jej implementovat do nějakého hardwarového microkontroleru, který musí pracovat s omezenými prostředky.

Když je proces učení dokončen, vytvoří program jednoduchý seznam kódovaných vzorů, které indikují jakým způsobem se zachovat po rozpoznání daného vzoru. Jak znázorňuje následující obrázek, hardwarová akcelerace by umožnila okamžitou odezvu na vstupní data. Rozpoznání vzoru sice pracuje s výpočtem nejrůznějších indikátorů, většina si však vystačí se základními matematickými operacemi (není nutné rekurzivně počítat exponent hodnoty, což by bylo výpočetně náročné). Daná jednotka by tedy mohla být urychlena specializovaným hardwarem. Po příchodu nových dat od brokera (nebo burzy) by byly ve zlomcích vteřiny aktualizovány hodnoty indikátorů, a pokud by byl rozeznán vzor, systém by ihned reagoval zprávou pro otevření pozice. Naopak samotné vyhledávání vzorů a sestavení tabulky vzorů za pomoci strojového učení je proces komplikovaný a zdlouhavý. Jelikož se při něm pracuje s historickými daty (a není nutné generovat odezvu v řádu milisekund), je hardwarová akcelerace zcela zbytečná.



Obrázek 5.2.1: Návrh hardwarové akcelerace programu

Daná hardwarová akcelerace je velmi vhodná v případě, když systém zpracovává TICK data. Tato data se mění velmi rychle, zato jen velmi jemně. Případný skluz v plnění by znamenal promrhání obchodní situace. U takového systému by tedy bylo žádoucí, aby reagoval v podstatě okamžitě. Delší doba výpočtu by totiž mohla znamenat reakci na neaktuální data (protože se rychle mění), a proto by tato reakce přinášela špatné výsledky. Pro práci s TICK daty, by musel být program rozšířen o rozhraní automaticky komunikující přímo s brokerem, neboť lidský faktor by zpomaloval efektivitu programu.

Hardwarová akcelerace je naopak velmi nevhodná u timeframe překračující desítky minut. Například u hodinových časových oken je již úspora několika milisekund za zpracování obchodu zcela zbytečná a zanedbatelná. Pokud systém odhalí vhodný vzor, který pravděpodobně zaručuje růst po dobu pěti hodin, není třeba snažit se odchytit první milisekundu tohoto růstu.

Jelikož většina začínajících spekulantů nedisponuje kapitálem, aby mohla zaplatit paušální poplatek u brokera (většinou 10 tisíc dolarů) práce se podrobnou analýzou TICK vstupu nezabývá. Nicméně algoritmus učení je obecný, a kdyby byl vyměněn datový vstup za TICK data (program by jen pracoval výrazně delší dobu, neboť agregace na minutové intervaly v průměru zmenší datový vstup stokrát) a zároveň by byla hardwarově akcelerována jednotka simulující (nebo provádějící) obchodování, byl by systém použitelný i pro obchodování nad TICK daty.

5.3 Získání dat

Datový vstup jako takový je velmi důležitý. Při použití nevhodných dat, nemůže být strojové učení efektivní a získané vzory jsou v praxi nepoužitelné. Je nutné pracovat s věrohodnými zdroji a počítat s následky, které přijdou, pokud se pracuje s nečistými daty.

V úvodní části celé práce se autorovi podařilo získat data od jednoho obchodníka. Jednalo se o patnáctiletou historii čtyř nejpoužívanějších indexů, se kterými se na burze obchoduje. Jen pro zmínku se jednalo o indexy SP mini, Russel 2000, DAX a YM. Patnáctiletá historie představovala velmi dobrý datový vzorek, obzvláště když byl zaznamenáván v denním, hodinovém, čtvrthodinovém, pětiminutovém a minutovém intervalu. Data byla snadno dostupná (zprostředkovaná přímo od

obchodníka) a tak nebyl důvod je nevyužít. Teprve až charakteristika odhalila v datech mnoho znepokojujících hodnot.

Při hlubší analýze se podařilo zjistit, že mnoho dní nebyla cenová řada zaznamenávána s dostatečnou frekvencí. Některé hodnoty nedávaly vůbec smysl (například na místě uzavírací hodnoty indexu DAX se mnohokrát vyskytovalo záporné číslo, což nebylo možné. I kdyby DAX čelil naprostému propadu, mohl by se dostat pouze na hodnotu 0 bodů, nikoliv však pod ni). Při převedení minutových intervalů na čtvrthodinové, se chybovost dat opět potvrdila. Jednotlivé agregace totiž nebyly shodné, buď byla data zaznamenána z různých zdrojů, nebo chybně. Ukázalo se, že z patnáctileté historie je nejstarších pět let naprosto nepoužitelných (z celého roku bylo zaznamenáno zhruba jen 50 dní). Data nemohla být použita pro věrohodnou analýzu a musela být nahrazena kvalitnějším zdrojem.

Samotný původ dat se v průběhu práce ukázal být velmi těžce dohledatelný. Obchodník, který data poskytl, s nimi nikdy nepracoval pro strojové učení a ani si nebyl jist, od koho je před lety získal. Poskytnutá data tedy byla naprosto nedůvěryhodná a velmi nečistá.

Jelikož bylo nutné sehnat datový zdroj, provedl autor analýzu dostupných poskytovatelů dat. Data od společnost IQFeed byla cenově nedostupná. Z mnoha dalších společností, byla zvolena společnost Sierra Chart, která nabízela velmi dlouhou historii napříč všemi komoditami. To umožnilo získat opravdu velmi dlouhé cenové řady a zároveň vyzkoušet univerzálnost systému napříč různými trhy (cenové řady obilí se vyvíjejí zcela jinak, než cenové řady ocele).

Podařilo se tak získat 7 let historie dat, pro většinu komodit (dostatek pro výběr hlavních reprezentativních komodit z každé oblasti). Data měla agregaci na minutové intervaly.

Přesto práce s těmito daty přinesla mnoho problémů. Sierra Chart se zaměřuje na vývoj nástroje pro grafické zobrazování cenových grafů (jak je znázorněno v podkapitole 2.4 Vizualizace číselných řad). Poskytovaná data jsou až vedlejším produktem, který má sloužit spíše k ilustrativnímu použití programu, než se uživatel dostane ke spolehlivějším datům. Přestože se v datech mohou nacházet chyby, dle serveru financnik.cz (zabývajícího se obchodováním na burze a dle mnohých fór, sdružující české obchodníky) jsou data pro analýzy použitelná, jelikož v hlavních obchodních hodinách je většina dat správně zaznamenáno.

5.4 Předzpracování dat

Samotná charakteristika dat od Sierra Chart vypadala lépe než prvotně získaná data. Analýza dat odhalila, že mimo hlavní obchodní dobu a víkend je frekvence záznamů velmi slabá. Společnost získávala vzorky dat zhruba jednou za hodinu (namísto každé vteřiny). Po dobu obchodních hodin existovaly údaje alespoň pro každou vteřinu. Jakmile obchodní hodiny skončili, tak v některých dnech toto frekventované zaznamenávání trvalo ještě několik hodin, pro jiné dny pak třeba vůbec. Ukázalo se však, že jakmile počet záznamů klesl pod určitou frekvenci (zhruba 1 záznam do dvou minut), byla již data nesprávně agregována (ve večerních hodinách nebyla agregovaná vůbec). Vytratily se tak z dat doplňující informace, jako kolik obchodů bylo za danou dobu uskutečněno (což je velmi často používaný ukazatel) nebo kolik transakcí čeká na bližších hodnotách. Chybně agregované hodnoty, tak sice mohly sloužit pro informativní vývoj řady, již však nebylo možné od dat očekávat odraz reálného stavu.

Byl sestaven program (clearer.py), který z dat pro jednotlivé komodity získal jen údaje pro hlavní obchodní hodiny. Zároveň se však autor pokusil z dat získat co nejvíce informací, a proto byla zachována i data s dostatečnou frekvencí záznamu (přidáním parametru *–oh*). Byly získány hlavní obchodní hodiny a zbylé použitelné hodiny (ať už před nebo po hlavních obchodních hodinách). Ukázalo se, že se v datech nacházejí nejrůznější časové ostrůvky, kdy byla frekvence zaznamenávání

opět vysoká. Přítomnost těchto mezer se nepodařilo plně vysvětlit, protože na výpadky systému byly příliš časté (navíc kdyby společnost Sierra Chart měla tak nespolehlivé zaznamenávání, nepodařilo by se jí správně zaznamenat hlavní obchodní hodiny). Výpadky by mnohem více odpovídaly údržbě. Později se ukázalo, že tyto ostrůvky do procesu učení zanášejí chyby (jelikož již byla přerušena spolehlivá informace o trhu), proto byly tyto informace vyloučeny. Výsledná získaná data tedy odpovídala hlavním obchodním hodinám a různě dlouhému intervalu před nimi i po nich.

Obrázek 5.4.1: Ukázka jaké hodiny byly vybrány pro učení

Poté co se podařilo získat rozsahy použitelné pro strojové učení, bylo nutné vyčistit drobné chyby i v těchto rozsazích. Velmi často se stávalo, že ve vybraném rozsahu chybělo až pět minut z datového záznamu. Tyto drobné chyby by neměly způsobovat významnou odchylku v nalezených vzorech, protože se cena nedokázala za tak krátký časový interval výrazně vyvinout. Proto byly tyto krátké chybějící intervaly doplněny ve směru aktuálního trendu dat. Doplněná data pak aproximovala přímku mezi částí předcházející a následující pro chybějící místo. Podobný postup byl proveden i pro všechny ostatní hodnoty, jakými bylo například volume. Zanesená odchylka do tak rozsáhlého vzorku dat (v průměru asi 3 minuty na 30 hodinách) by měla mít nulový vliv a proto může být zanedbána.

Problematické pasáže nastaly, v případě když chyběly dlouhé časové intervaly. Pokud ze záznamu chyběly například tři hodny, nebylo by vhodné data doplnit jen na bázi spojnice. Kdyby spojnicová křivka kopírovala tvar křivky z minulého dne, také by se do programu zavedla pro dané období nebezpečná odchylka. Navíc hrozilo, že algoritmus podobnost dat jednotlivých dní bude vyhodnocovat jako vzor, který by sice pravděpodobně na celkovém vzorku dat neuspěl, ale zbytečně by zatěžoval proces učení. Vyloučit ze vzorku dat celý den, by jistě nebylo nejvhodnějším řešením. Naštěstí bylo možné využít vlastností korelace, mezi některými komoditami. Například obilí a kukuřice se pohybují velmi podobně. Když klesá hodnota jedné komodity obvykle (asi z 80%) klesá i hodnota komodity druhé. Autor sice neví, jakým způsobem společnost Sierra Chart zaznamenává data, ale korelující komodity bývaly pravděpodobně zaznamenávány na různých serverech, protože pokud chyběly nějaká data jedné cenové řady, bylo možné je doplnit daty řady druhé. Takto doplněné hodnoty pak velmi věrně naznačovaly vývoj, který se v daném období pravděpodobně odehrál (a nenaznačovaly vznik falešného vzoru, využitím tvaru funkce z jiného dne).

Teprve u komodit, které nebyly v korelaci, s žádnou jinou komoditou bylo nutné využít nějaký způsob spojení. Aproximující křivka, kopírující průměrný tvar daného týdne, pak měla náhodně posunuté vrcholy, aby co nejméně souvisela s jinými daty. Takto doplněných hodnot, však bylo velmi málo a proto by měla být jakákoliv statistická odchylka, která byla tímto způsobem do dat zanesena, překryta vskutku rozsáhlým vzorkem dat sedmileté historie.

5.5 Vlastnosti systému

Podařilo se vytvořit systém, který odpovídá specifikaci. Výsledný program pracuje nad očištěnými daty. Provede nutné kroky zpracování a vyhledá v cenových řadách nejvhodnější vzory, které umožňují predikci následujícího vývoje cenové řady. Tento seznam vzorů, je pak výstupem celého programu. Může být využit při simulacím reálného obchodování, podle kterého se spekulant může s nalezeným seznamem vzorů naučit pracovat. V budoucí fázi, by program mohl automaticky brokerovi nahlašovat rozpoznané akce (na otevření nebo uzavření pozice nad komoditou).

Protože program nemá grafické rozhraní, je jeho ovládání pro běžného uživatele mírně nepohodlné. Přesto se autor pokusil všechny parametry podrobně popsat v README souboru, který je k programu přiložen a také každý skript programu obsahuje vlastní help. Autor doufá, že i netechnicky zaměřený spekulant, bude schopen po přečtení práce a README s daným programem pracovat. K tomu napomáhá i zapouzdřující skript, který (bez nutných parametrů) provede celý proces učení.

Jak již bylo řečeno, programem provádějící predikci, je série skriptů:

```
$ ls
cleaner.sh  first_analys.py  indicators.py  learn.py       main_hour.py  profit_char.py  run_all.sh      run.py      utils.py
data        grooper.py       joiner.py      log_joiner.py  paterns       README          run_process.sh  spliter.py
```

Jednotlivé části již byly vysvětleny v kapitole 5.1 Strojové učení. Program je tedy šířen jako skupina skriptů, soubor README, složka se vstupními daty a nakonec složka pro výstupní seznamy pravidel. Pro shrnutí bude naznačeno ukázkové spuštění všech částí programu.

```
#! /bin/bash

echo "= = = Get firs analysis and make same format = = = "
python first_analys.py -xk data/${KOMODITY}, -xf 1990 -xt 2016 -xo _c|| exit 1
echo "= = = Get good trade day and complete trade hour = = = "
python main_hour.py -xk data/${KOMODITY},_c -xf 1990 -xt 2016 -mo $MO -mc $MC -o $OO -c $CC -t $TRESHOLD -cd -oh  || exit 1
echo "= = = Get profit histogram data = = =  "
python profit_char.py -a -xk data/${KOMODITY}@_c_OH -xf 1990 -xt 2016 -o 13 -c 21 -d data  || exit 1
echo "= = = Joining data = = = "
python joiner.py -k data/${KOMODITY},_c_fill -f 1990 -t 2016 -o data/${KOMODITY}_all -p 60 -dt sierra || exit 1
echo "= = = Groop data $TIMEFRAME = = ="
python grooper.py -f data/${KOMODITY}_all -o data/${KOMODITY}_allG_$TIMEFRAME -t $TIMEFRAME || exit 1
echo "= = = Add some indicators $TIMEFRAME = = ="
python indicators.py -f data/${KOMODITY}_allG_$TIMEFRAME -o data/${KOMODITY}_allI_$TIMEFRAME || exit 1
echo "= = = Spliting files (learn / test) $TIMEFRAME = = ="
python spliter.py -f data/${KOMODITY}_allI_$TIMEFRAME -om data/${KOMODITY}_allI_${TIMEFRAME}_learn \
                  -ot data/${KOMODITY}_allI_${TIMEFRAME}_test -p 60 -t indi || exit 1
echo "= = = Going to learn $TIMEFRAME = = ="
python learn.py -f data/${KOMODITY}_allI_${TIMEFRAME}_learn -o data/${KOMODITY}_allI_${TIMEFRAME}_learn_rule \
                -wt $WAIT -sl $STOPLOSS -ws $WINDOW -b $BACK -p $PROFIT -mp $MPROFIT -v || exit 1
echo "= = = Simulate learning data = = ="
python run.py -i data/${KOMODITY}_allI_${TIMEFRAME}_learn -r data/${KOMODITY}_allI_${TIMEFRAME}_learn_rule \
              -o paterns/${KOMODITY}_all_${TIMEFRAME}_learning_rule -wt $WAIT -sl $STOPLOSS -ws $WINDOW -b $BACK -p $PROFIT -mp $MPROFIT -v || exit 1
echo "= = = Simulate testing data = = ="
python run.py -i data/${KOMODITY}_allI_${TIMEFRAME}_test -r data/${KOMODITY}_allI_${TIMEFRAME}_learn_rule \
              -o paterns/${KOMODITY}_all_${TIMEFRAME}_testing_rule -wt $WAIT -sl $STOPLOSS -ws $WINDOW -b $BACK -p $PROFIT -mp $MPROFIT -v || exit 1
echo "= = = SIMULATE MARKET = = ="
python run.py -i data/${KOMODITY}_allI_$TIMEFRAME -r data/${KOMODITY}_allI_${TIMEFRAME}_learn_rule \
              -o paterns/${KOMODITY}_all_${TIMEFRAME}_market_rule -wt $WAIT -sl $STOPLOSS -ws $WINDOW -b $BACK -p $PROFIT -mp $MPROFIT -v || exit 1
if [ "$CLEAR" == "YES" ]; then
  echo "clearing temp files"
  bash cleaner.sh data
fi
echo "= = = ALL DONE = = ="

```

Obrázek 5.5.1: Příklad spuštění úpravy dat, učení a simulace

Přestože sestavený program, prochází vstupní data několikrát (v průběhu předzpracování, shlukování, učení a testování), je celková složitost algoritmu nižší jak 20N (záleží, jestli je vyžadována postupná vizualizace učení, kdy se při každé iteraci musí projít vstupní data, aby se mohl zobrazit přesný výsledek aktuálního systému). Násobitel vstupu je z hlediska algoritmické složitosti zanedbatelný, takže je celková složitost programu lineární. To je velmi užitečná vlastnost, neboť při zpracovávání dlouhé TICK cenové řady by musel program procházet kolem miliardy záznamu, což by při kvadratické složitosti výpočtu vyžadovalo desítky hodin.

Program byl optimalizován nad zpracováním indexů, která mají oproti ostatním komoditám výrazně vyšší frekvenci obchodování a tím i záznamů od brokera. Indexy mají více obchodních příležitostí a tak s nimi pracuje mnoho spekulantů. Jelikož jsou indexové trhy velmi živé (je na nich prováděno přes milion obchodů denně) jsou velmi vhodné pro intradení obchodování. Plodiny již mají provedených obchodů za den již mnohem méně (ty nejvýraznější zhruba 200 tisíc). Například pomerančový extrakt, má kolem 10 tisíc obchodů denně a pro intradení obchodování se nehodí (je vhodný pro poziční obchodování). Jelikož je program zaměřen na intradenní obchodování a na práci s živými trhy (co mají přes statisíce obchodů za den), jeho schopnosti predikce u nevýrazných komodit selhávají. Program se sice dá spustit s parametrem „slow“ (který se snaží přizpůsobit učení pro slabé a pomalé trhy), ale výsledné predikce nejsou tak uspokojivé jako nad indexy (či jinými často obchodovanými trhy).

Kromě samotného nalezení vzorů, sloužících k predikci cenové řady, je vhodné program využít i k získání statistik jednotlivých řad. Histogramy počtu obchodů v jednotlivých hodinových mohou

sloužit k optimalizaci již existujících systémů. Stejně tak histogramy potenciálního profitu a velikosti trendů umožnují analytikovi získat přehled jestli je daná komodita vhodná k obchodování.

Celkový program je tak velmi univerzální a může sloužit jako pomocný rádce mnoha spekulantům bez ohledu na jejich psychologii a strategii obchodování. Přestože nemá dokonalé grafické rozhraní, může být program z hlediska schopností a ceny opravdovým přínosem.

5.6 Testování a vyhodnocení systému

Samotné testování systému bylo provedeno pomocí skriptu, který se pokusil provést strojové učení s nejrůznějšími parametry. Celý proces učení popsaný v kapitole 5.1 Strojové učení byl spuštěn nad všemi komoditami. Dále bylo pro každou komoditu hledáno vhodné nastavení pro velikost zpracovávaného okna, stop lossu, očekávaného profitu, minimálního profitu, míry s jakou se cenové řady mohou propadnout (a tím snížit již získaný zisk), timeframe a nakonec i časem který je spekulant ochoten na zisk čekat než kontrakt uzavře. Jelikož se jedná o hledání optimální kombinace mnoha vlastností systému, je celý proces (jednalo se o mnoho běhů jednoho procesu učení) učení prováděn několikrát, což výrazně prodlužovalo dobu učení.

```
$ bash run_all.sh
```

Ukázalo se, že výsledný program je schopen nalézt nejrůznější paletu úspěšných vzorů, které umožnují praktické sestavení nejednoho obchodního systému. Každý spekulant by si tak měl být schopen najít v systému svůj způsob obchodování.

Jelikož bylo nutné otestovat mnoho kombinací (od různých timeframe, přes nejrůznější profit targety a nesmí být zapomenuto na různé komodity) byl celý výpočet prováděn paralelně na 8 výkonných serverech. I přes paralelní výpočty trval celý proces učení mnoho hodin.

Výsledné hodnoty však ukázaly, že optimální počet svíček, se kterými se pracuje, by neměl překračovat 10. Pokud bylo zpracováváno více svíček, staly se jednotlivé kódové kombinace určující vzor příliš specifickými (kromě svíček byl vzor definován i mnoho indikátory) a jejich četnost výskytu rapidně klesala jen k pár jednotkám za několik let. Vzory s takto nízkým výskytem pak nemohou působit příliš věrohodně a obchodování s nimi by v případě neúspěchu pravděpodobně vedlo k porušování obchodní strategie. Na druhou stranu dané vzory měly obecně mnohem vyšší úspěšnost než vzory ostatní.

Dále se ukázalo, že pro vyšší timeframe hodnoty klesá schopnost systému generovat zisk. Pokud jednotlivé svíčky představovaly agregaci několika dnů, výsledný systém většinou skončil ve ztrátě. To jen potvrzuje, že pro poziční obchodování je vhodné využívat spíše fundamentální než technické analýzy. Zatímco na kratším timeframe při využití intradenního obchodování se jednotlivé fundamenty příliš neprojevují (spíše jen dlouhodobým trendem trhu) a tak se cenový graf řídí aktuální nabídkou a poptávkou obchodujících spekulantů. Pro poziční obchodování již fundamenty silně ovlivňují trh a z technické analýzy je nebylo možné předvídat.

Systém byl sice při pozičním obchodování schopen dosáhnout zisku, ale za cenu velmi vysokých stop lossů. Při několika neúspěšných obchodech v řadě pak byla ztráta i kolem 15 tisíc dolarů, což pro většinu malých spekulantů mohlo představovat celý jejich kapitál. Tyto vlastnosti systému opět jen potvrdily, že poziční obchodování je vhodné pouze pro obchodníky s opravdu velkým kapitálem, který jim dovoluje takovéto propady překonat. Spekulant by totiž neměl na jeden obchod riskovat více jak 5 procent kapitálu. [11]

Programu se v různých kombinacích podařilo nalézt mnoho vzorů s profity přes několik tisíc bodů. Jeden bod může dle komodity představovat několik dolarů. Například pro komoditu YM se podařilo najít obchodní systém o 20 vzorech, které by dohromady představovaly zisk přes 80 tisíc bodů za 6 let, což odpovídá průměrnému profitu 66 tisíc dolarů za rok.

Samotné hledání obchodního systému bylo zaměřeno buď na profit, četnost výskytu nebo úspěšnost predikce. Ukázalo se, že pro jednotlivé typy systém nalézá úplně jiné vzory s jinými vlastnosti.

Výsledky dopadaly nejlépe pro zaměření se na celkový profit, neboť ten vyžadoval, aby vzory příliš nechybovali a zároveň aby se vyskytovaly dostatečně často a profit bylo možné uskutečnit. Nalezené vzory byly tedy nejuniverzálnější:

Akce-vzor	Profit	Četnost	Úspěšnost	Prům.profit
U-4,6,0,6,0,6,0,6,2,6	6548.0	1161	69 %	6
D-4,3,0,3,2,3,0,3,5,3	4171.0	313	79 %	13
D-1,3,0,3,0,3,2,3,5,3	3296.0	249	77 %	13
U-1,6,0,6,2,6,5,6,2,6	3245.0	160	71 %	20
D-4,3,5,3,2,3,2,3,0,3	2728.0	101	75 %	27
U-5,2,4,2,5,2,0,2,2,2	2711.0	247	61 %	11
D-0,3,5,3,4,3,5,3,2,3	699.0	139	71 %	5
D-5,3,2,3,5,3,5,3,1,3	795.0	52	69 %	15

Příklad vzorů, které systém vyhledával se zaměřením na profit.

Naopak při zaměření systému na vysokou četnost, hledané vzory musely být velmi obecné, což způsobovalo jednak vyšší chybovost, ale hlavně i predikci na kratší dobu. Systém tak vůbec nevyhledával riskantní situace, které ale mohly přinést výrazný zisk:

Akce-vzor	Profit	Četnost	Úspěšnost	Prům.profit
U-1,6,0,6,0,6,0,6,0,6	4268.0	1197	64	4
D-4,6,0,6,0,6,0,6,5,6	2532.0	1103	71	2
U-4,6,0,6,0,6,0,6,0,6	2366.0	1020	61	2

Příklad vzorů se zaměřením na četnost, je znatelná nízká míra průměrného profitu

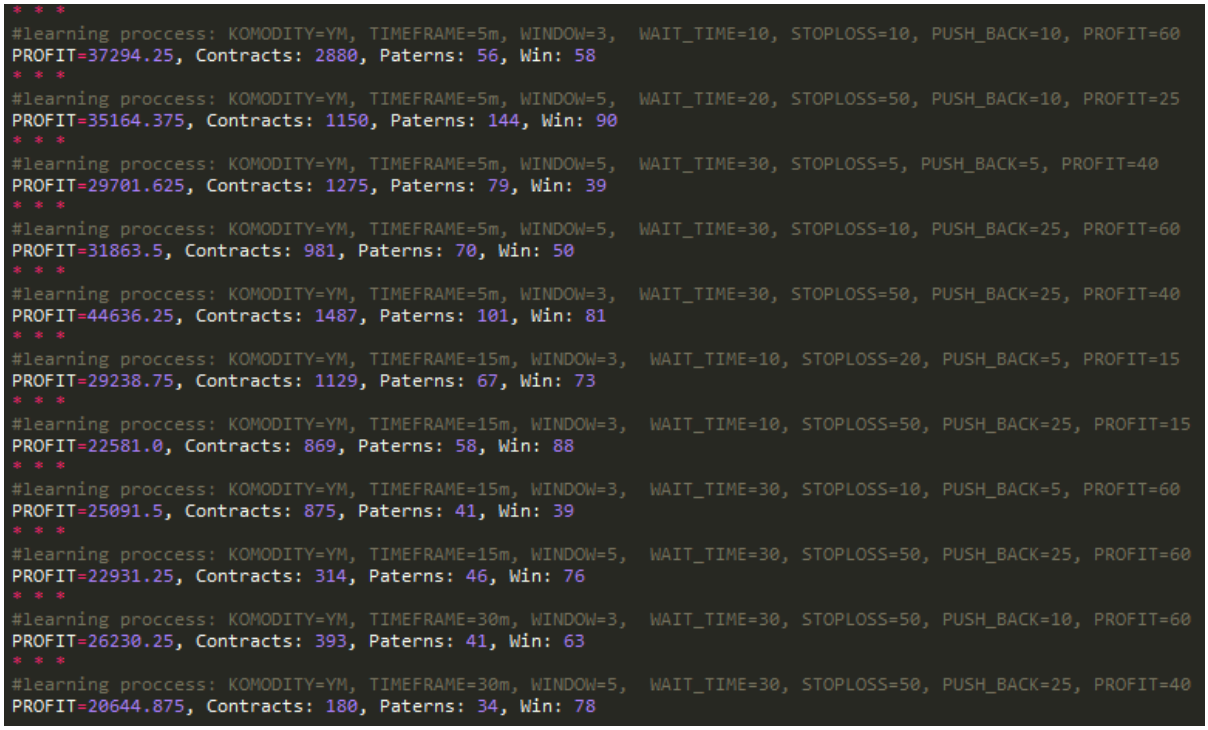
Nejhůře dopadaly systémy, které byly vyhledávány podle vysoké pravděpodobnosti úspěšné predikce. Systém pak pracoval převážně s naprosto specifickými charakteristikami vzorů, které se v cenové řadě téměř nevyskytovaly, a proto tyto vzory nemohou být považovány za odolné, neboť pravděpodobně tvoří jen statistickou odchylku. Přesto jejich průměrný profit na jeden obchod je velmi příznivý:

Akce-vzor	Profit	Četnost	Úspěšnost	Prům.profit
U-2,3,6,3,5,3,2,3,2,3	442.0	10	100 %	44
U-2,3,2,3,2,3,6,3,2,3	273.0;	4	100 %	68
D-5,2,5,2,5,2,5,2,3,2	299.0;	15	93 %	20

Příklady vzorů zaměřených na úspěšnost predikce. Vysoký průměrný profit, ale mizivá četnost.

Provést srovnání získaného systému s ostatními systémy není zrovna jednoduchá záležitost. Je mnoho kritérií, které systém může do určité míry naplňovat a každý analytik preferuje jiné. Tak jak může být někomu nepohodlné obchodovat na příliš krátkém timeframe, tak může být různým spekulantům nepohodlné obchodovat se systémem, který často prohrává, přestože jeho profit je kladný. Samotný program je možné zaměřit na vyhledání vzorů podle nejlepšího profitu, úspěšnosti, četnosti atd. Bohužel se však zatím nedá stanovit procentuální důležitost jednotlivých parametrů.

Systém nalézá vzory od zhruba 40% úspěšnosti pro predikci až po zhruba 80% úspěšnost. Vzory s vyšší úspěšností již mají příliš nízký počet výskytů a tak nemusí být úplně věrohodné. Samotný poměr zisku a ztráty může nabývat hodnot od 1:1 až po 5:1.



Obrázek 5.6.1: Příklady nalezených vzorů a jejich vlastností (šedě jsou parametry učení a pod nimi jsou vypsány vlastnosti)

Autorem vytvořený systém tedy poskytuje možnost nalézt celou plejádu možných systému s různými vlastnostmi, které si musí každý spekulant specifikovat sám. Tato vlastnost je velkou výhodou vytvořeného systému, neboť jen spekulant ví, jaké jsou jeho psychické možnosti, a i kdyby pracoval s kvalitním systémem, který by ale neodpovídal jeho osobnímu obchodnímu profilu, jen stěží by s takovým systémem dokázal dlouhodobě generovat zisk.

Právě vysoká variabilita je velkou výhodou daného systému. Mnoho placených systému je již nějak na kombinováno, aby odpovídaly obchodní strategii toho, kdo obchodní systém vytvořil. Daný systém, který je přednastaven od někoho jiného, může mít velmi specifické vlastnosti, a pokud se spekulantovi některá nelíbí a on ji pozmění (vyhozením vzoru s příliš velkým riskem), může tím zničit konzistenci celého obchodního systému. Spekulant je tedy svázán danou podobou zakoupeného systému a jen těžce ji může přizpůsobit svým potřebám.

Podařilo se najít systém, který je schopen dosahovat profitů a zároveň respektovat psychické možnosti spekulanta. Práce jako taková byla velmi rozsáhlá a muselo být spojeno mnoho znalostí o strojovém učení, burze a číselných řadách. Výsledný program však přináší nejen mnoho různých vzorů a obchodních systémů, ale i zajímavých poznatků.

Velmi pozoruhodná byla provázanost tří hlavních vlastností: Profitu, úspěšnosti a četnosti výskytu. Čím vyšší měl být profit, tím riskantnější obchody bylo nutné vykonat. Tyto riskantní obchody pak mnohem vícekrát neuspěly (průměrná úspěšnost těchto vzorů byla od 45 do 55%), ale když se jim podařilo uspět, přinesly až pětinásobek riskované ztráty. Naopak pokud bude uživatel vyžadovat od systému vysokou úspěšnost, je poměr riskované a získané částky již velmi blízký. Kdyby takový systém měl úspěšnost jen 50%, nebyl by schopen generovat zisk, a proto je nutné volit málo riskantní pravidla s úspěšností okolo 80%. Při tlaku na často se vyskytující vzory (které by vyloučili možnost statistické odchylky) nemohly být vzory tak specifické, a tak získaly nejvyšší fitness ohodnocení vzory obecné a tolerantní, které ovšem nemívaly příliš vysokou úspěšnost. Mnoho chybných předpovědí pak snižovalo možný zisk. Celkový profit takových systému sice nebyl tak vysoký (jako když se hledání vzorů zaměřovalo na profit), ale zase velmi vysoký počet výskytu vzoru byl schopný spekulantovi garantovat dobrou obecnost vzoru.

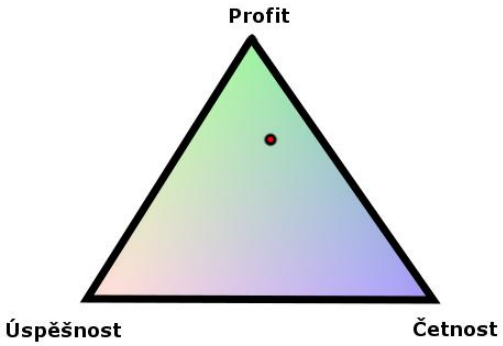
Pro dlouhá časová okna je technická analýza nevhodná, a i program jako takový selhává. Při širokém timeframe by technická analýza měla být skombinována s analýzou fundamentální, protože ta již velmi výrazně ovlivňuje cenový vývoj. Program také není příliš vhodný pro trhy, které jsou obchodovány jen velmi málo (jedná se o okrajové a netypické suroviny), neboť cenové řady pak neposkytují dostatek obchodních příležitostí.

Program by celkově pro začínající obchodníky mohl být velkým přínosem, i když je pravdou, že je na něm ještě mnoho možného vývoje.

5.7 Možný rozvoj

Jak to u většiny programů bývá, i při této práci bylo vymyšleno mnoho vylepšení programu ještě před dopsáním posledního řádku kódu. Přestože výsledky programu jsou dobré, je zde stále možný rozvoj. Úpravy by se jednak týkaly usnadnění použití programu třetí osobou, která nemá technické vzdělání, ale také použitelnost programu pro obchodování v reálném čase. Následuje seznam některých úprav, který by se systémem mohly být provedeny pro vytvoření kvalitnějšího systému:

- **GUI** – k programu by mohlo být přidáno grafické rozhraní, které by ochránilo netechnické uživatele před zadáváním parametrů. Dané GUI by také mohlo vykreslovat průběh zpracovávaného grafu a označovat místa vstupu a výstupu (při simulaci, nebo v reálném běhu programu).
- **Poměrové vlastnosti** – profit, úspěšnost a četnost jsou na sobě velmi závislé a hledání obchodního systému jen podle jedné z těchto vlastností nemusí být úplně optimální. Bylo by vhodné pomoci grafického rozhraní dodat uživateli nástroj, aby sám specifikoval jak je pro něj která váha důležitá. K pohodlnému zadávání takové hodnoty by pak sloužil posuvník v trojúhelníkovém poli.



Obrázek 5.8.1: Příklad jak specifikovat zaměření programu na profit, úspěšnost a četnost

Dále se musí upravit program samotný, aby pro fitness ohodnocení zvažoval všechny tři hodnoty ve zvoleném poměru.

- **Externí indikátory** – bylo by vhodné rozšířit schopnost systému o externí přidání indikátorů, se kterými by pak učení probíhalo. Tím by si každý spekulant mohl vyhledat vhodné vzory ke svému již používanému obchodnímu systému
- **Komunikace s brokerem** – aby se systém dal použít pro automatické obchodování bez zásahu člověka, musí mít schopnost rozšířit své rozhraní takovým způsobem, aby byl schopen zadávat příkazy k nákupu a prodeji (či o zavření pozice) přímo brokerovi. Problém však může být fakt, že různé brokerské společnosti mají různé komunikační rozhraní.

- **Zohlednění počátečního kapitálu** – jedná se o velmi vhodnou úpravu, která by zohledňovala počáteční kapitál spekulanta. Aktuální systém najde nejvýhodnější pravidla pro obchodování, ty však mohou vyžadovat vysoký kapitál, který spekulant nemusí mít. Kdyby se systém rozšířil o možnost zadat výši kapitálu a veškeré učení by pak této výši bylo přizpůsobeno. Celkové výsledky by tak více odpovídaly potřebám uživatelů.
- **Odhalování výstupních vzorů** – jednotlivé vstupní vzory se v programu podařilo predikovat s dostatečným profitem. Přesto by bylo možné zvýšit celkový profit sestaveného obchodního systému při predikci vzorů výstupních. Tyto vzory by predikovaly nejvhodnější setrvání v otevřené pozici, aby se nestávalo, že se ztratí již nabytý profit.

Jednotlivá zmíněná vylepšení pomohou pozvednout program na profesionální úroveň. V příštích verzích programu by tak uživatelé měli být schopni snadno a pohodlně generovat velmi kvalitní obchodní systémy, které jim pomohou zpracovávat cenové řady na burze. Tyto obchodní systémy by mohly být schopné konkurovat kvalitním placeným systémům, a tím by program přinesl úsporu mnoha spekulantům.

6 Závěr

Cílem práce bylo provedení analýzy současného stavu číselných řad, strojového učení a vzájemné aplikace při technické analýze nad burzovními daty. Na základě získaných informací měl být sestaven program pro predikci cenových řad, jehož výsledky by byly využitelné. Tento cíl se podařilo split.

V práci byla představena problematika číselných řad a základní prvky, které analytici provádějí při technické analýze. Tyto znalosti pak zefektivnily způsoby učení a výsledky predikce. Byly představeny problémy získání a čištění dat. Následovalo vysvětlení vhodných metod pro strojové učení, které se pro predikci číselných řad a vyhledávání obchodních vzorů používají.

Implementační část představila problémy, stanovená řešení a jednotlivé výsledky strojového učení. Výsledné vzory byly dle uvažovaných metrik schváleny, jako vhodné pro predikci číselných řad. Práce se zabývala i možností zrychlení celého systému pomocí přenesení některých částí do embedded systémů.

Programem sestavené systémy vzorů jsou schopny dosahovat profitu a mohu být specifikovány a přizpůsobeny dle potřeby spekulanta, který je bude využívat. Autorem vytvořený program byl schopen nalézt vzory s úspěšností predikce až okolo 80% a stále dostatečnou četností výskytu, což vylučuje možnost statistické odchylky. Mnoho navrhovaných systémů bylo schopno dosahovat profitu přes 300 tisíc dolarů za období 7 let.

Práce potvrdila, že technická analýza komodit je velmi efektivní pro kratší časové intervaly a pro intradenní obchodování, naopak pro intervaly přesahující jeden den, již selhává. Dalším zajímavým zjištěním byla vzájemná provázanost mezi profitem, úspěšností predikce a četností výskytu daného vzoru. Apel na zvýšení jedné vlastnosti vzoru, totiž snižuje hodnotu ostatních dvou.

Přestože výsledky práce jsou velmi povzbudivé, i zde je možný další budoucí rozvoj. Práci by bylo vhodné rozšířit o grafické rozhraní, které by jí poskytovalo pohodlnější ovladatelnost, a dále o možnost zpracovávat více indikátorů, z jejichž seznamu by si budoucí uživatel zvolil ty, které sám chce využívat. Dále by se hodila možnost personifikace přímo pro uživatele. Bylo by vhodné, aby každý spekulant mohl rozložit míru důležitosti mezi profit, četnost výskytu a pravděpodobnost úspěšné predikce, čímž by nalezené vzory více odpovídali psychologickému profilu toho, kdo je bude používat. Programu se daří odhalit vzory indikující otevření pozice. Celý obchodní systém, by však mohl být mnohem efektivnější, kdyby predikoval i vzory pro uzavření pozice. Pro zvýšení pohodlí ovládání celého systému, je možné rozšířit celý program o možnost analyzovat historii obchodů spekulanta a na bázi této historie by se systém sám přizpůsobil danému spekulantovi. Nebylo by tak nutné v programu nastavovat jednotlivé parametry obchodního systému, neboť by se hodnoty nastavily automaticky z analýzy historických obchodů. Těmito úpravami by byl program srovnatelný s placenými systémy.

Literatura

[1] Clenow A. F.: *Following the trend: diversified managed futures trading*. TJ International Ltd, Cornwall, Great Britain, 2013. ISBN 978-1-118-41085-1

[2] Bollinger J.: *Bollinger on Bollinger bands*. McGraw-Hill, New York, USA, 2002. ISBN 0071373683.

[3] Nesnídal T., Podhajský P.: *Burza srozumitelně: VÝBĚR TRHU komodity a ETF's trhu*. Centrum finančního vzdělání, CZ, 2011

[4] Nison S.: *The Candlestick Course*. Wiley, USA, 2003. ISBN-13: 978-0471227281

[5] Williams R. L.: *How I Made One Million Dollars ... Last Year ... Trading Commodities*. Windsor Books, 3rd edition, Great Britain, 1998. ISBN-13: 978-0930233105

[6] Williams R. L: *Long-Term Secrets to Short-Term Trading*. Wiley, USA, 2011. ISBN-13: 978-0470915738

[7] Williams T.: *Master of Market*. TradeGuider Systems, USA, 2005. ASIN: B001GF0LAM

[8] Schwager D. J.: *Schwager on Futures: Technical Analysis*. Wiley, USA, 1995, ISBN-13: 978-0471020516

[9] Schwager D. J., Turner C. S.: *Futures: Fundamental Analysis*. Wiley, USA, 1995, ISBN: 978-0-471-02056-1

[10] Chande S. T.: *Beyond Technical Analysis: How to Develop and Implement a Winning Trading System.* Wiley, 2nd edition, USA, 2001, ISBN: 978-0-471-41567-1

[11] Nowak J.: *Kompletní průvodce psychologie obchodování*. Finančník, CZ, 2005

[12] Dreman N. D.: *Psychology and the Stock Market: Investment Strategy Beyond Random Walk*. American Management Association, USA, 1977, ISBN: 978-0814454299

[13] Zendulka J., Bartík V., Lukáš R., Rudolfová I.: Studijní opora k předmětu Získávání znalostí z databází. Vysoké učení technické - *Fakulta informačních technologií*, Brno, Česká republika, 2006.
Dostupné na URL: <http://www.fit.vutbr.cz/~bartik/ZZN.pdf> (přístup 3. 10. 2015)

[14] Zbořil F.: přednášky k předmětu Soft Computing. Vysoké učení technické - *Fakulta informačních technologií*, Brno, Česká republika, 2013.

[15] Ryšavý M.: *Morfologický analyzátor pomocí konečných automatů* [online]. Vysoké učení technické - *Fakulta informačních technologií*, Brno, Česká republika, 2013.
Dostupné na URL:
<https://www.vutbr.cz/www_base/zav_prace_soubor_verejne.php?file_id=118692>
(přístup 6. 1. 2016)

[16] Kolektiv autorů: *Adaptive Builder User's Guide*, Adaptrade Software, USA, 2015.
Dostupné na URL:
<http://www.adaptrade.com/Builder/AdaptradeBuilderUG.pdf> (přístup 6. 1. 2016)

[17] Donalek C.: *Supervised and Unsupervised Learning*, Astronomy Colloquia. USA, 2011.
Dostupné na URL:
<http://www.astro.caltech.edu/~george/aybi199/Donalek_Classif.pdf>
(přístup 6. 12. 2015)

[18] Dasu T., Johnson T.: *Exploratory Data Mining and Data Cleaning*. Wiley, USA, 2003. ISBN: 978-0-471-26851-2

[19] Kolektiv autorů: *Data quality and data cleaning*. Rutgers, USA, 2004.
Dostupné na URL:
<www.cs.rutgers.edu/~muthu/dqted.ppt> (přístup 5. 12. 2015)

[20] Pyle D.: *Data Preparation for Data Mining* [online]. Morgan Kaufmann Publishers, San Francisco, USA, 1999.
Dostupné na URL:
<http://www.temida.si/~bojan/MPS/materials/Data_preparation_for_data_mining.pdf>
(přístup 10. 12. 2015)

[21] García S., Luengo J., Herrera F.: *Data Preprocessing in Data Mining*. Springer, Switzerland, 2015. ISBN: 9783319102467

[22] Witlen I. H, Frank E. Hall M. A.: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3. Vydání, USA, 2011. ISBN-13: 978-0123748560

[23] Kolektiv autorů: *Regresní a korelační analýza*. České vysoké učení technické, Praha, Česká republika, 2013.
Dostupné na URL:
<https://www.fd.cvut.cz/department/k611/pedagog/K611THO_soubory/0_regrese.pdf>
(přístup 10. 10. 2015)

[24] Miklánek T.: *Predikce – regresní modely*. Vysoké učení technické - *Fakulta informačních technologií*, Brno, Česká republika, 2011.
Dostupné na URL:
<http://www.fit.vutbr.cz/study/courses/ZZD/public/seminar0405/miklanek.pdf>
(přístup 10. 10. 2015)

[25] Kolektiv autorů: *Regresní analýzy*. Quonia, Brno, Česká republika, 2011.
Dostupné na URL:
<http://geoinovace.data.quonia.cz/materialy/ZX510_Pokrocile_statisticke_metody_geografickeho_vyzkumu_MU/Regresni_analyza.pdf> (přístup 10. 10. 2015)

[26] Raschka S.: *Python Machine Learning*. Packt Publishing, Great Britain, 2015. ISBN-13: 978-1783555130

[27] Kačer P.: *Vícevrstvá neuronová síť*. Vysoké učení technické - *Fakulta informačních technologií*, Brno, Česká republika, 2013.
Dostupné na URL:
<https://www.vutbr.cz/www_base/zav_prace_soubor_verejne.php?file_id=64938>
(přístup 5. 12. 2015)

[28] Poli R., Langdon B. W., McPhee F. N.: *A Field Guide to Genetic Programming*. Lulu Enterprises, USA, 2008. ISBN-13: 978-1409200734

[29] Goldberg E. D: *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, USA, 1989. ISBN-13: 978-0201157673

[30] Pošík P.: *Paralelní Genetické Algoritmy*. České vysoké učení technické, Praha, 2001.
Dostupné na URL:
<http://labe.felk.cvut.cz/~posik/dipl/Diplomka.htm> (přístup 5. 12. 2015)

[31] Kolektiv autorů: *Analýza časových řad*. České vysoké učení technické: *Fakulta elektrotechnická*, Praha, 2004.
Dostupné na URL:
<http://bio.felk.cvut.cz/~huptycm/Vyuka/IKTZ_prednasky/CasoveRady0910.pdf>
(přístup 10. 10. 2015)

[32] Hančlová J., Tvrdý L.: *Úvod do analýzy časových řad*. Vysoká škola báňská: Institut geoinformatiky, Ostrava, Česká republika, 2003.
Dostupné na URL:
<http://gis.vsb.cz/pan-old/Skoleni_Texty/TextySkoleni/AnalyzaCasRad.pdf >
(přístup 10. 10. 2015)

[33] Kolektiv autorů: *Časové řady*. Vysoká škola ekonomická, Praha, Česká republika, 2008. Dostupné na URL: <http://iastat.vse.cz/casovky/casovky2.htm> (přístup 10. 10. 2015)

[34] Dorba M.: *Časové řady*. Vysoká škola báňská: *Technická Univerzita Ostrava*, Ostrava, Česká republika. 2003. Dostupné na URL: < http://homel.vsb.cz/~dor028/Casove_rady.pdf> (přístup 10. 11. 2015)

[35] Sebera M.: *Časové řady v kinantropologickém výzkumu*. Masarykova Univerzita, Brno, Česká republika, 2012. Dostupné na URL: <http://is.muni.cz/do/rect/habilitace/1451/33088294/33088307/Habilitacni_prace_Sebera.pdf> (přístup 5. 12. 2015)

[36] Kleeman L.: *Understanding and Applying Kalman Filtering*. Clayton: *Monash University Department of Electrical and Computer Systems Engineering*, Dostupné na URL: <http://biorobotics.ri.cmu.edu/papers/sbp_papers/integrated3/kleeman_kalman_basics.pdf> (přístup 5. 12. 2015)

[37] Kolektiv autorů: *Kalmanův filtr.* Vysoké učení technické - *Fakulta informačních technologií*, Brno, Česká republika, 2011. Dostupné na URL: <http://www.uamt.feec.vutbr.cz/~richter/vyuka/0910_mpov/tmp/kalman_filter.html> (přístup 16. 11. 2015)

[38] Kolektiv autorů: *Stavový model a Kalmanův filtr*. České vysoké učení technické: *Fakulta elektrotechnická*, Praha, 2013. Dostupné na URL: <https://www.fd.cvut.cz/personal/provipav/Stochastika/Materialy-test4/Stav.pdf> (přístup 16. 11. 2015)

Příloha A

Slovníček pojmů[1]

Analytik – jedná se o účastníka na burze, který ji neovlivňuje (ne přímo). Analytik zkoumá chod burzy a snaží se ji pochopit, aby se později stal spekulantem, nebo aby jim za finanční odměnu nabízel své rady.

Broker – registrovaný zprostředkovatel pro nákup a prodej finančních produktů od burzovního domu jako jsou například komodity a akcie.

Burza – burza je místo, na kterém se za striktního dohledu kontrolních orgánů provádějí jednotlivé burzovní obchody. Burzy jsou různě specializované – existují burzy komoditní, akciové nebo např. opční. Na burze se obchoduje prostřednictvím prostředníků – tzv. brokerů. Párování příkazů dnes stále může probíhat ručně (na tzv. pitu), nebo častěji elektronicky. Obchodníci komodity nakupují od kohokoliv, kdo je ochoten v daný okamžik kontrakt za stanovenou cenu prodat.

Býčí trh – rostoucí trh. Každý trh, který roste (cena komodity, akcií jde nahoru), se nazývá býčí trh (bull). A naopak, každý trh, který klesá (cena komodity jde dolů), se nazývá Medvědí trh (bear). Stejně tak se trhy nazývají stoupající či klesající.

FinWin – kompletní obchodní systém autorů webu financnik.cz – Petra Podhajského a Tomáše Nesnídala. Autoři si obchodní systém FinWin sami postavili k vydělávání peněz na trzích, nyní systém Finančník Winner (FinWin) vyučují v rámci svých seminářů. Systém je postavený na indikátorech CCI14 a CCI50 a základy systému je možné nastudovat například v knize Jak se stát intradenním finančníkem. Systém je vhodný pro futures i pro forex, zejména pak pro intradenní obchodování. Aplikace je možná i na denní grafy, jak na futures, forex, tak na akcie. Systém je univerzální a dá se aplikovat prakticky na libovolný trh.

Futures kontrakt – je dohoda dvou stran o nákupu či prodeji standardizovaného množství komodity v předem specifikované kvalitě za danou cenu a k určitému budoucímu datu. Pěstitel (prodávající) a obchodník (kupující) se dnes dohodnou při uzavírání futures kontraktu na uzavření obchodu v budoucnosti za určitou cenu. Prodávající již dnes ví, kolik v budoucnu dostane peněz, a kupující ví, kolik zaplatí.

Gap – zcela jednoduše řečeno představuje gap takovou oblast grafu, kde nedošlo za danou cenu k žádným obchodům. Nejčastěji se gapy objevují na denních grafech – tedy takových, kde jeden bar představuje jeden obchodní den. Gapy se zde objevují proto, že otevírací cena (open) se liší od uzavírací ceny (close) předchozího dne. Tím vznikne na grafu mezera neboli gap. Gap může vzniknout proto, že se přes noc (kdy se daný finanční instrument neobchoduje) objevily určité zásadní informace (report, katastrofická zpráva, náhlá změna počasí, politické prohlášení atd.) ovlivňující zájem obchodníků, kteří hromadně dávají své obchodní příkazy na open následujícího obchodního dne. Množství příkazů pak způsobí, že se cena na open začne obchodovat výše (chtějí-li obchodníci především nakupovat) nebo níže (chtějí-li obchodníci především prodávat) než byla v době uzavírání trhu předchozí den.

Chop – jedná se o anglický výraz pro trh, který se nikam nehýbe, resp. jde do strany. Chop je pro obchodníky nezajímavý, neboť na trzích, které netrendují (jdou do strany), není v drtivé většině možné vydělávat a spíše se jenom ztrácí.

Komise (ve smyslu poplatek) – poplatek účtovaný za službu brokera (tzv. brokerská komise, commission nebo také fee). Tato komise se pak účtuje většinou za kompletně provedený obchod (tzn. vstup do pozice i výstup z pozice). Komise se obvykle skládá z několika částí: 1. poplatek brokerovi –

[1] Dle Finančník.cz <http://www.financnik.cz/wiki/glosar>

každý broker ho může mít jinak vysoký a pro jednotlivého obchodníka může záviset např. na množství obchodů provedených za měsíc. 2. poplatek burze – může být různý pro různé instrumenty a burzy. 3. regulační poplatek – ze zákona, například v USA je vybírá National Futures Association

Komoditní spread – současný nákup a prodej futures kontraktu. Spread pak představuje rozdíl, který pohybem ceny vzniká.

Kontraktní měsíc – měsíc vypořádání konkrétní komodity. Vypořádáním (fyzickým dodáním) se komodita s dodáním v příslušném měsíci přestává obchodovat.

Margin – vratná záloha umožňující ovládat komoditní kontrakt.

Medvědí trh – klesající trh. Každý trh, který klesá (cena komodity jde dolů), se nazývá medvědí trh (bear).

Obchodník – jedná se o účastníka na burze, který ji ovlivňuje přes svého brokera. Jeho cílem je získat kontrakty (nákupem) a fyzicky dostat zboží, pokud je konzumentem zboží (například podnik zpracovávající ropu a vyrábějící plastové výrobky). Nebo naopak prodat své zboží, pokud se jedná o producenta (například těžařská společnost).

Obchodní systém – systém pravidel definující "jak konkrétně obchodovat". Tato strategie obchodování by měla obsahovat: vstupní strategie (popsaná vzory, pracující s indikátory), metoda umisťování a posouvání stop-lossu, strategie výstupu z obchodů (profit-taking), money-management a position-sizing.

Otevření pozice – jedná se o vstup do obchodování, ať již nákupem nebo prodejem kontraktu. Uzavření pozice je pak provedeno opačnou akcí, která byla na vstupu.

Pit – jedná se o fyzické místo burzy, kde se scházejí obchodníci a zaměstnanci burzy, aby zde uskutečnili své obchody. V dnešních dobách pit ustupuje rychlejšímu a pohodlnějšímu párování obchodů přes internet.

Rollover (Forex) – převrácení (překlopení) je vlastně provedení výměny pozice, která je držena s pozicí následujícího dne vyrovnání. Přeneseně můžeme říct, že rollover provede otočení otevřené pozice za kontaktní den a to uzavření pozice v daném období a otevřením pozice v období následujícím.

Slippage – anglický výraz Slippage, česky přeložitelný jako skluz, označuje rozdíl mezi požadovaným a skutečně získaným plněním. Zejména v rychlejších trzích se běžně stane, že příkaz dostane "skluz" několik ticků, což může podle trhu znamenat rozdíl i několika desítek dolarů. Může se tak stát, že kontrakt bude nakoupen mnohem dráž.

Spekulant – jedná se o účastníka na burze, který ji ovlivňuje přes svého brokera. Jeho cílem je získat kontrakty (nákupem) a pak se jich zase zbavit (prodejem), nebo v opačném pořadí, za účelem dosažení zisku. Spekulant si nepřeje zboží fyzicky získat.

Spread – rozpětí je rozdíl mezi nákupní a prodejní cenou.

Stop-loss – jedná se o předem definovaná krajní hranice ztráty, po které se pozice uzavře. Zadávání stop-lossu není povinné, ale jeho nepoužití může vést k ohromným ztrátám

Svíčkový graf – Jedná se o techniku zvanou „candle stick charting", neboli zobrazování a čtení grafů v podobě tak zvaných „svíček". Svíčkové neboli candle stick grafy jsou speciálními krabicovými grafy.

Time frame – časové měřítko grafu. Poziční obchodníci budou používat např. denní grafy, kde jedna úsečka představuje změnu ceny za jeden den. Jde o tzv. denní timeframe. Intradenní obchodníci používají intradenní timeframe – například tříminutový, kde jedna úsečka představuje změnu ceny za tři minuty.

Volatilita – faktor udávající jak živý a rychlý daný trh je. V živých trzích lze vydělat více peněz za menší časové období, ale současně lze hodně peněz ztratit, nejedná-li obchodník rychle a rozvážně. Volatilita jde vyčíst z grafu a to rozpětím maximální a minimální ceny v rámci dne. Například kukuřice je běžná denní úsečka rozpětí cca 200 dolarů, v době maximální volatility má úsečka rozpětí cca 600

dolarů. Oproti tomu například ropa má rozpětí běžného dne cca 1200 dolarů, při extrémních výkyvech pak i před 3500 dolarů.

Volume – objem obchodů zrealizovaných v rámci dané časové periody (záleží na použitém time frame). Vysoké volume značí vysokou likviditu trhu, ale tyto trhy mohou přinášet nepříjemné slippage, z důvodu frontového zpracovávání obchodů.

Příloha B

Přiložené CD

Na přiloženém CD, se nachází elektronický obsah této práce, zdrojové kódy skriptů i vstupní data pro strojové učení.

CD

/doc – složka s elektronickou verzí textu práce. Text je jednak ve formátu pdf a také v editovatelné podobě programu Microsoft World.

/data – obsahuje vstupní datové soubory pro komoditu YM, tak jak byly získány od společnosti Sierra Chart. Vzhledem k autorským právům, nad tady (a ohromné velikosti) nemohou být šířena všechna

/paterns – složka pro výstupní seznamy nalezených vzorů

Skripty – jednotlivé skripty pro zpracování dat a strojové učení. Skripty již byly vysvětleny v textu této práce a také jsou vysvětleny v souboru README

RADME – obsahuje popis skriptů a jejich příklady spuštění