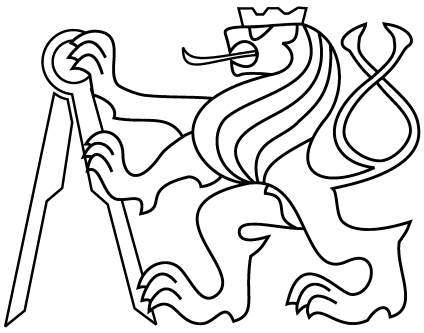
Bakalářská práce

Detekce vzorů v časových řadách

Jiří Bystroň



Květen 2015

Vedoucí práce: Ing. Martin Mudroch, Ph.D.

České vysoké učení technické v Praze
Fakulta elektrotechnická
Katedra řídící techniky

České vysoké učení technické v Praze
Fakulta elektrotechnická

katedra řídicí techniky

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Student: **Jiří Bystroň**

Studijní program: Kybernetika a robotika
Obor: Systémy a řízení

Název tématu: **Detekce vzorů v časových řadách**

Pokyny pro vypracování:

1. Proveďte rešerši používaných metod pro vyhledávání neurčitých vzorů v těžko predikovatelných časových řadách (typicky nad ekonomickými, meteorologickými a obdobnými daty).
2. Vybranými metodami zpracujte konkrétní data.
3. Srovnejte možnosti použití a kvalitu výstupu jednotlivých metod a diskutujte jejich případné odchylky a nedostatky.

Seznam odborné literatury:

[1] Last, Mark, Kandel, Abraham, and Bunke, Horst, eds. Data Mining in Time Series Databases. SGP: World Scientific Publishing Co., 2004. ProQuest ebrary.
[2] Lilly, John H.. Fuzzy Control and Identification. Hoboken, NJ, USA: John Wiley & Sons, 2010. ProQuest ebrary.

Vedoucí: Ing. Martin Mudroch, Ph.D.

Platnost zadání: do konce letního semestru 2015/2016

L.S.

prof. Ing. Michael Šebek, DrSc.
vedoucí katedry

prof. Ing. Pavel Ripka, CSc.
děkan

V Praze dne 3. 3. 2015

Poděkování

Tímto bych rád poděkoval vedoucímu této práce, Ing. Martinu Mudrochovi, Ph.D., za velice konstruktivní poznámky a vstřícný přístup při sestavování jak individuálního projektu, tak samotné bakalářské práce.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze dne ..... Podpis autora .....

Abstrakt

Tato bakalářská práce se zabývá rozpoznáváním vzorů ve finančních časových řadách metodami rule-based, fuzzy a pomocí klasifikátoru založeného na podobnosti s průměrem korektně určených vzorů. Součástí práce je popis, návrh a implementace těchto metod v jazyce Java. Jako podkladová data pro vyhledávání vzorů byly vybrány finanční řady z trhu Forex. Výstupem této práce je jak prokazatelná schopnost vzory těmito metodami detekovat, tak srovnání těchto metod.

Klíčová slova

pattern recognition, fuzzy, rule-based, time series, candlestick

Abstract

This bachelor thesis deals with the pattern recognition in financial time-series using rule-based method, fuzzy method and classification method, which is based on similiarity to an average of correctly specified patterns. This thesis consists of method description, design and implementation in Java language. As underlying data for pattern recognition were chosen time-series from Forex market. The outcome of this thesis is both a demonstrable ability to recognize patterns with these methods and an evaluation of these methods.

Keywords

pattern recognition, fuzzy, rule-based, time series, candlestick

Obsah

Slovník použitých výrazů

časový rámec	···	time frame
částečné učení s učitelem	···	semi-supervised learning
členská funkce	···	membership function
clustering; shlukování	···	clustering
dopředná neuronová síť	···	feedforward neural network
exponenciální klouzavý průměr	···	exponential moving average
holá data	···	raw data
jednoduchý klouzavý průměr	···	simple moving average
k-nejbližších sousedů	···	k-nearest neighbors
klasifikační a regresní strom	···	classification and regression tree
klasifikační krok	···	classification step
klouzavý průměr	···	moving average
křížová validace	···	cross validation
lichoběžníková	···	trapezoidal
nejbližší sousedé	···	nearest neighbors
ostrá data	···	crisp data
ostrá množina	···	crisp set
přeučení	···	overfitting
regresní stromy	···	regression trees
rekurentní neuronová síť	···	recurrent neural network
rozhodovací stromy	···	decision trees
rozpoznávání vzorů	···	pattern recognition
samoorganizující mapy	···	self-organizing maps
schodový graf	···	bar chart
strojové učení	···	machine learning
svícový graf	···	candlestick graph
testovací množina	···	test set
trénovací množina	···	training set
trojúhelníkový	···	triangular
učení bez učitele	···	unsupervised learning
učení s učitelem	···	supervised learning
učící krok	···	learning step
umělá neuronová síť	···	artificial neural network
validační množina	···	validation set
vážený klouzavý průměr	···	weighted moving average

Seznam použitých zkratek

ARIMA	···	AutoRegressive Integrated Moving Average
ARMA	···	AutoRegressive Moving Average
BS	···	velikost těla svíce (Body Size) – pouze pro účely práce
CART	···	Classification And Regression Tree
CBR	···	Case-Based Reasoning
CHF	···	švýcarský frank
CSV	···	Comma-Separated Values či též Character-Separated Values
DCT	···	Discrete Cosine Transform
DFT	···	Discrete Fourier Transform
DS	···	dolní stín svíce (Dolní Stín) – pouze pro účely práce
EUR	···	euro
FOREX	···	FOReign EXchange
GUI	···	Graphical User Interface
HMM	···	Hidden Markov Models
IPCC	···	Intergovernmental Panel on Climate Change
KDD	···	knowledge Discovery from Data
LHC	···	Large Hadron Collider
MA	···	Moving Average
MC	···	Markov chain
NN	···	Nearest Neighbor
OHLC	···	Open High Low Close
PAA	···	Piecewise Aggregate Approximation
RBS	···	velikost reálného těla svíce (Real Body Size) – pouze pro účely práce
S&P 500	···	Standard & Poor's 500
SAX	···	Symbolic Aggregate approXimation
SL	···	délka stínu svíce (Shadow Length) – pouze pro účely práce
STING	···	STatistical INformation Grid-based method
US	···	horní stín svíce (UpShadow) – pouze pro účely práce
USD	···	americký dolar
PPV	···	positive predictive value

1. Úvod

Mezi hlavní cíle této bakalářské práce patří v prvé části vypracování rešerše pro zadané téma vyhledávání neurčitých vzorů ve špatně predikovatelných časových řadách. Tím získám základní přehled o dostupných metodách, technikách a vhodnosti jejich implementace pro dané prostředí.

V části druhé je cílem se již prakticky zaměřit na volbu a vlastní teoretický a programový návrh konkrétních metod pro konkrétní časové řady z oblasti finančních trhů, v nichž budu detekovat konkrétní, obecně uznávané vzory. Výstupem bude demonstrace, že mnou vybrané a implementované metody jsou schopny detekovat tyto vzory. Třetím cílovým bodem je, že tyto metody a výstupy vhodně zvolenými metodami porovnám. Tomu bude odpovídat i struktura práce.

Protože se jedná o specifické téma, ve kterém jsou anglické termíny ustálené, české překlady jsou místy i ke škodě. Budu se tedy držet spíše anglických termínů, jelikož překlad některých termínů de facto ani neexistuje nebo není zcela adekvátní a zavádí akorát nejasnosti. Jsem si vědom toho, že tento způsob prezentace není zcela ideální, nicméně s ohledem na to, že je práce psána česky, se tímto budu snažit vyhnout vytváření nevhodných česko-anglických novotvarů. Uvádím též přehledový slovník, kde lze nalézt překlady vybraných termínů s ohledem na použitý kontext.

Jsem si též vědom faktu, že diskutované teoretické téma nemusí být běžnému čtenáři známo, jako mu nemusí být známa i oblast vybíraných dat – specifických finančních trhů. V některých sekcích této práce proto budu volit bližší popis či názorný příklad než uvedení pouhé definice.

Pojmy z mého pohledu odborné, případně výrazy z cizího prostředí budu při prvním výskytu uvádět v uvozovkách. Při dalších výskytech těchto pojmů a výrazů už uvozovky vynechám. Pokud použiji nadnesené či nepřesné výrazy, vždy je uvedu v uvozovkách.

2. Současný stav

V současné době, která je též přezdívaná „informačním věkem“, se čím dál častěji setkáváme s množstvím oblastí, ve kterých je potřeba zpracovávat enormní množství různorodých dat. Ať už se jedná o data v podobě záznamů z průmyslových senzorů, lékařských přístrojů, klinických databází nebo o data kosmologická, finanční, seismická, meteorologická či data z webových serverů společností jako je například Google, vždy je potřeba tato data vhodnými způsoby uchovávat, třídit a analyzovat.

Zaměříme-li se na aktuální trendy a směr budoucího vývoje, bez újmy na obecnosti se dá hovořit o pojmu „big data“. Ten lze chápat jako „*termín aplikovaný na soubory dat, jejichž velikost je mimo schopnosti zachycovat, spravovat a zpracovávat data běžně používanými softwarovými nástroji v rozumném čase*“ [1]. O vzrůstající popularitě tohoto pojmu svědčí mimojiné fakta, že za poslední rok vzrostla v USA poptávka po datových analyticích se specializací na big data téměř o 100 % [6], dále vzniká velké množství kurzů zaměřených na big data [7] a ostatně minulý rok byl i na Fakultě elektrotechnické ČVUT otevřen volitelný předmět Technologie pro velká data [4].

Pro představu, například v roce 2008 servery společnosti Google zpracovaly požadavky čítající v průměru přes 20 petabajtů dat denně. [2] Jako další příklad můžeme uvést, že každou hodinu je uživateli nahráno na servery společnosti Facebook přes 10 milionů fotografií [5, s. 16]. Nebo také data z měření ve velkém hadronovém urychlovači částic (LHC) čítají přibližně 30 petabajtů za rok [3]. Je zřejmé, že při takových objemech dat je nutné se zabývat metodami, které umožňují s daty efektivně pracovat jak po stránce výpočetní, tak po stránce interpretační.

Svědčí o tom například i fakt, že ačkoliv má být společnost Google schopna mapovat výskyt chřipky díky vyhledávacím požadavkům uživatelů z celého světa stejně dobře, jako jej mapují data z lékařských ordinací [5, s. 19][8], ukazuje se, že to nemusí být úplně pravda, jak rozebírá Steven Salzberg [9]. Problémem je totiž špatné pochopení lidského chování v tomto kontextu a s tím dále spojená interpretace dat jako i jejich vytěžování. V odkazovaném zdroji se k tomuto váže vhodná věta: „*The folks at Google figured that, with all their massive data, they could outsmart anyone.*“ Považuji tedy za vhodné zaměřit se na metody vytěžování klasických dat z časových řad.

3. Teoretická část

3.1 Prostředky pro zpracování dat

Jelikož se budu zabývat oblastmi jako je rozpoznávání vzorů, strojové učení či vytěžování dat a autoři se ne vždy v definici těchto pojmu shodují, je vhodné tyto základní pojmy nejdříve objasnit pro lepší zasazení do našeho kontextu.

3.1.1 Vytěžování dat

Někteří autoři [10, s. 5–6] se pouštějí do polemiky o definici tohoto pojmu a tvrdí, že by se měl jmenovat spíše „knowledge mining from data“. Na což plynule navazují tvrzením, že je tento pojem na jednu stranu chápán jako synonymum pro pojem „knowledge discovery from data“ (KDD), na stranu druhou uvádějí, že může být též chápán jako pouhý jeden krok v komplexním procesu extrakce vědomostí z dat. Později však dochází ke konsensu s jinými autory [11, s. 5] v tom, že vytěžování dat lze popsat jako automatizovaný či částečně automatizovaný proces objevování vzorů ve zpravidla větším množství dat, nalezené vzory musí mít smysluplný význam dle požadovaného zadání a obecně se jedná o řešení problémů analýzou dat, která již existují v databázi. Přičemž databází je zde myšlen v podstatě libovolný, avšak dostatečně objemný informační zdroj.

3.1.2 Strojové učení

Bez újmy na obecnosti lze vyjít z tvrzení, že strojové učení se zabývá metodami, jak se počítačové programy mohou učit automatickému rozpoznávání komplexních vzorů, případně jak se mohou inteligentně rozhodovat na základě vstupních dat. Například klasickou úlohou, která bývá často v tomto kontextu uváděna, je schopnost programu korektně určit ručně psané poštovní směrovací číslo na základě předložených, správně určených vzorů – trénovací množiny, viz dále. [10, s. 24]

3.1.3 Rozpoznávání vzorů

Rozpoznávání vzorů je obecně chápáno jako podmnožina strojového učení, respektive jeho konkrétní aplikace, ačkoliv v některých případech je kladeno do stejné roviny jako samotné strojové učení [12, s. vii]. Rozpoznávaní vzorů je možno uplatnit na rozličná vstupní data, textem nebo zvukem počínaje a symboly na dopravních značkách konče. V případě této práce se jedná o časové řady, respektive data, kterými jsou tyto řady reprezentovány.

3.2 Časové řady

Časovou řadu je možné obecně chápat jako soubor hodnot získaný sekvenčními měřeními za určitý časový úsek. Formální definici je možné zapsat následovně.

Definice 1. Časová řada T délky n je taková posloupnost dvojic

$$T = [(p_1, t_1), (p_2, t_2), ..., (p_i, t_i), ..., (p_n, t_n)], \tag{3.1}$$

kde $t_1 < t_2 < ... < t_i < ... < t_n$ a kde každé p_i představuje datový bod v d-dimenzionálním prostoru a každé t_i představuje čas, kdy byl p_i změřen. [14, s. 11]

Je zřejmé, že se vzrůstajícím počtem dat a dimenzí prostoru se dá očekávat větší náročnost ať už co se týče výpočtů či definování podobnosti časových řad. Vyvstávají poté základní otázky a problémy. [13, s. 12:2]

- Reprezentace dat
  Jak je možné reprezentovat základní tvarovou charakteristiku časové řady, jaké by měla mít vlastnosti? Reprezentace by měla ideálně redukovat dimenzi dat se zachováním podstatných charakteristik datové řady.

- Měření podobnosti
  Jak může být mezi dvěma libovolnými časovými řadami nalezena shoda či jak mohou být odlišeny? Jak je možné formalizovat vzdálenost těchto dvou řad, případně jak je možné rozpoznat intuitivní podobnost řad, ačkoliv nejsou po matematické stránce identické?

- Indexovací metoda
  Jak by měly být organizovány velké objemy dat, které časové řady reprezentují, aby bylo možné v nich rychle vyhledávat? S přihlédnutím k minimální výpočetní a datovému objemu?

Výčet však není konečný, jde jen o jádro problematiky vytěžování dat z časových řad.

3.2.1 Reprezentace časové řady

Jelikož výpočetní operace na holých datech by byly náročné, zavádí se pojem reprezentace. Vedlejším jevem zavedení reprezentace bývá též snížení šumu, jako i snížení datového objemu uložených dat. [13, s. 12:13]

Definice 2. Reprezentací časové řady T délky n nazveme takový model $\bar{T}$ s redukovanými dimenzemi, pro který platí, že $\bar{T}$ aproximuje T. [14, s. 11]

Mezi obecné požadavky na optimální reprezentaci dat, která představují časové řady, patří zejména následující body. [13, s. 12:13]

- významná redukce dimenze dat
- zachování tvarových charakteristik časové řady v lokálním i globálním měřítku
- rekonstrukce původních dat z redukované reprezentace je kvalitní
- necitlivost vůči šumu nebo implicitní potlačení šumu

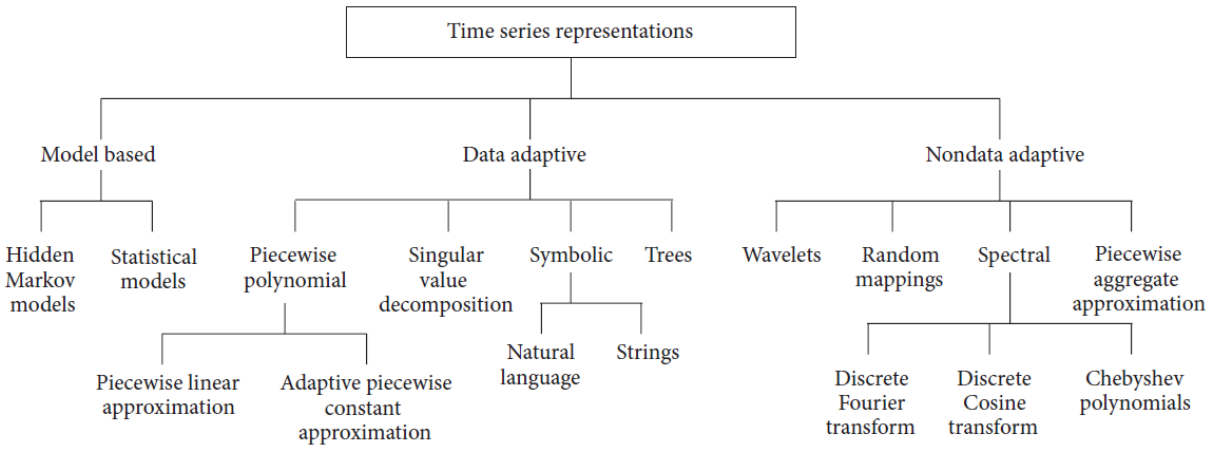
Mezi základní metody a techniky reprezentace dat, respektive časových řad patří zejména následující. [13, s. 12:13]

- Non-data adaptive
  Parametry transformace respektive redukce dimenze jsou stejné pro jakoukoliv časovou řadu nehledě na podstatu dat, která řadu tvoří. Patří zde zejména diskrétní Fourierova transformace (DFT), diskrétní kosinová transformace (DCT) nebo například piecewise aggregate approximation (PAA). Ta je unikátní v tom, že časovou řadu rozdělí na N segmentů stejné délky, pro které spočte střední hodnotu, čímž vzniká nová řada o N bodech. Dle některých studií [15] však poskytuje nepřesné výsledky vlivem velké ztráty informace.

- Data adaptive
  Na rozdíl od předchozí metody tato metoda již podkladová data zohledňuje a téměř každý non-data adaptive postup se stává data adaptive tím, že přidáme do metody krok, který vybírá konkrétní parametry metod. V případě diskrétní Fourierovy transformace je to například vhodná selekce koeficientů, v případě PAA je to volba dynamické [16] délky segmentů. Unikátní metodou je též symbolic aggregate approximation (SAX), která vychází z PAA, nicméně získané segmenty na stejných či blízkých úrovních označuje písmeny, čímž získáváme posloupnosti písmen. Dle druhu aplikace dosahuje též lepších výsledků než například DFT a další [17].

- Model based
  U této metody se předpokládá, že data reprezentující časovou řadu byla generována nějakým implicitním modelem. Cílem je tedy najít parametry daného modelu, čímž je nalezena i reprezentaci dat. Zde se nejvíce uplatňují například Markovovy řetězce (MC), autoregressive moving average (ARMA) modely, autoregressive integrated moving average (ARIMA) modely či Hidden Markov Models (HMM).

Nakonec uvádím na obr. 3.1 pro přehlednost detailnější rozdělení reprezentace. Kvalitnější obrázek je možné nalézt v příloze B.



Obrázek 3.1: Detailní rozdělení reprezentace časových řad (převzato z [18])

3.3 Klasifikace

S analýzou časových řad souvisí některé základní úlohy, jmenovitě například: „query by content“, „anomaly detection“, „motif discovery“, „prediction“, „clustering“, „classification“, „segmentation“ [13, s. 12:1]. Pro potřeby této práce se zaměřím pouze na klasifikaci. V kontextu časových řad se klasifikace dá popsat jednoduše podle následující definice:

Definice 3. Mějme neklasifikovanou časovou řadu T. Klasifikací časové řady nazveme takový proces přiřazení časové řady do jedné z tříd c_i z množiny $C = \{c_i\}$, kde C reprezentuje množinu předdefinovaných tříd [13, s. 12:7].

Tato definice platí analogicky jak pro podposloupnosti časové řady, tak pro jednotlivé vzory, které se v ní vyskytují. Klasifikace v obecném kontextu je tedy proces, který vybraným vstupům přiřazuje vybrané výstupy.

3.3.1 Dělení dle supervize

První metodou je „učení s učitelem“. Jedná se o dvoukrokový proces [10, s. 328], kdy prvním krokem je krok učící, ve kterém dochází ke konstrukci klasifikačního modelu na základě trénovací množiny. Ta obsahuje vybraná data – vstupy – pro která jsou již známy korektní klasifikační třídy – výstupy. Výstupy jsou známy nejčastěji na základě manuálního označení. Druhým krokem je krok klasifikační, ve kterém již dochází ke klasifikaci konkrétních tříd pro data z množiny testovací, pro která není klasifikační třída známa.

Problémem při tomto postupu bývá přeučení, což představuje stav, kdy je výběr testovací množiny příliš úzce zaměřen, množina není dostatečně obecná [10, s. 330]. Pro účely ověření správnosti klasifikátoru je možno použít validační množinu, přičemž platí obecné pravidlo, že by trénovací, testovací a validační množina měly být navzájem disjunktní [19].

Opačnou metodou je „učení bez učitele“, kdy nejsou známy požadované výstupy. Jsou k dispozici jen data, na která jsou aplikovány metody, které vychází zejména z podobností ve vstupních parametrech dat. Jedná se zejména o metodu shlukování. [10, s. 330]

Existuje minimálně ještě 1 další metoda – „částečné učení s učitelem“, která stojí na pomezí dvou výše uvedených. Dle některých autorů je však obecné zavedení nové metody diskutabilní a dle dílčí konfigurace ji zařazují pod metodu učení s učitelem. [19, s. 15–16].

3.3.2 Dělení dle klasifikačního modelu

Existují v zásadě 2 přístupy. „Lazy learning“ a „eager learning“.

První jmenovaný je založen na pouhém uložení trénovací množiny. Ke klasifikaci dochází až po kontaktu s testovací množinou na základě podobnosti s trénovací množinou, respektive nějakým jejím prvkem. Mezi typické zástupce lazy learning metody patří metoda „nejbližších sousedů“ či metoda „case-based reasoning“ (CBR) [10, s. 422–423]. Metoda CBR se však uplatňuje hlavně ve znalostních databázích, pro časové řady existují daleko vhodnější metody, jak ukáži záhy.

Naproti tomu přístup eager learning spočívá v tom, že na základě trénovací množiny je přímo vytvořen klasifikační model již před kontaktem s testovací množinou. Tento model je poté aplikován na samotnou testovací množinu. Mezi zástupce této metody patří de facto všechny zbylé

metody mimo CBR a nearest neighbors [10, s. 422–423].

V následujících sekcích stručně popíšu některé základní metody klasifikace, z nichž vycházejí další, pokročilejší metody klasifikace [10, s. 393]. Z těchto metod budu také dále vycházet v této práci.

3.3.3 Rule-based klasifikace

Jedná se o triviální IF-THEN přístup, kdy je možné klasifikační pravidla zapisovat ve tvaru

IF *rule antecedent* THEN *rule consequent*,

přičemž „rule antecedent“ má význam podmínky, „rule consequent“ má význam úsudku [10, s. 355]. Je zřejmé, že podmínek může být více. Ty jsou poté dávány do vztahů logickými spojkami AND či OR. Jedná se o metodu, kdy je nutné, aby rozhodovací pravidla byla přesně specifikována, nejčastěji ve spolupráci s doménovým expertem [23, s. 11]. Ten má hlubší náhled do problematiky, případně se též může jednat o komunitní znalosti, k čemuž se dostanu v praktické části. Tyto metody obecně bývají též označovány jako „hard computing“ metody.

3.3.4 Rozhodovací stromy

Pojem „strom“ je zde chápán v kontextu teorie grafů, přičemž klasifikace atributů pomocí rozhodovacích stromů spočívá ve vytvoření hierarchické stromové struktury tak, že každý uzel (větvení) reprezentuje test daného atributu a každá větev směřující z tohoto uzlu reprezentuje rozhodnutí. Je-li větev zakončena listem, pak se jedná přímo o zařazení do klasifikační třídy. [19, s. 52–53] Rozhodovací stromy je možné převést do klasického IF-THEN přístupu, aniž by docházelo ke kolizím; klasifikační pravidla se tedy budou navzájem vylučovat. De facto tedy spadají pod rule-based metody, kam se též někdy zařazují [10, s. 358].

3.3.5 Soft computing

Jedná se o množinu více metod, které však na rozdíl od rule-based metod včetně rozhodovacích stromů nepotřebují detailní rozhodovací pravidla, ale při těchto metodách postačuje základní nutné minimum rozhodovacích pravidel či požadovaný výsledek klasifikace. Daná metoda se již samostatně snaží dosáhnout požadovaných výsledků. Absence doménového experta v těchto případech tedy bývá daleko méně citelnější než v případě rule-based metod. [23, s. 12]

3.4 Metody rozpoznávání vzorů v časových řadách

V této části již stručně shrnu vybrané metody respektive techniky rozpoznávání (detekce) a klasifikace vzorů v časových řadách. Při sestavování seznamu metod jsem vycházel z více zdrojů, které se však v mnoha bodech shodují [20][21][22][24].

3.4.1 Nejbližší soused

Metoda nejbližšího souseda, 1-NN respektive k-NN, kde k představuje počet sousedů, je relativně stará metoda, která byla popsána již v 50. letech 20. století. Jedná se o metodu z množiny učení s učitelem. Klasifikace probíhá ve 2 fázích. Trénovací fáze spočívá v pouhém uložení objektů z trénovací množiny společně s klasifikovanou třídou. V klasifikační fázi je klasifikovanému objektu přiřazena stejná třída, jakou má k objektů z trénovací množiny, které jsou klasifikovanému objektu nejblíže. Pojem „nejblíže“ zahrnuje různé druhy metrik, například euklidovská či manhattanská a další. [10, s. 423]

Metoda nejbližších sousedů obecně je pro klasifikaci časových řad dle dostupných zdrojů [26] úspěšně použitelná. Konkrétně metoda 1-NN je udávána [26][27] ve spolupráci s křížovou validací jak standardní metoda pro měření a vyhodnocování přínosnosti různých reprezentací časových řad, tak jako i standardní metoda pro měření jejich podobností. Její značnou nevýhodou je však špatná odolnost vůči šumu.

3.4.2 Umělé neuronové sítě

Jedná se o metodu, jejíž kořeny sahají až do roku 1942 [30][31], která, dle konkrétního typu neuronové sítě, umožňuje všechny možnosti supervize. Základem umělé neuronové sítě je matematický model biologického neuronu, respektive spojení více těchto neuronů. Neuronová síť je v čase proměnlivá, je možné [23, s. 19] rozlišit 3 stavy této sítě.

- Organizační stav, ve kterém dochází ke změně topologie (architektury) sítě. V základě existují dva typy architektur a to rekurentní síť a dopředná síť.
- Aktivní stav, ve kterém se specifikují inicializační stavy sítě a který definuje způsob změny stavu sítě při pevně dané architektuře a konfiguraci.
- Adaptivní stav, ve kterém dochází ke změnám vah dílčích neuronových spojení. Cílem adaptace je nalézt takovou konfiguraci, aby síť v aktivním režimu realizovala požadovanou funkci.

Problematika a dělení neuronových sítí je značně hluboké téma, nicméně ve zkratce je možné říci, že je tato metoda pro naše účely použitelná [10, s. 398–408][28]. V tomto kontextu mají též úspěchy samoorganizační mapy [38]. Často zmiňovanou komplikací však bývá netransparentnost metody, relativně komplexní implementace různých metod, na druhou stranu jsou však neuronové sítě značně flexibilní, odolné vůči šumu a jsou obecně robustní [26] [10, s. 398] [29, s. 333–353].

3.4.3 Rozhodovací stromy

Jak jsem již uváděl v sekci 3.3.4, jedná se o vytvoření hierarchické rozhodovací struktury. Pro klasifikaci časových řad a vyhledávání vzorů v těchto řadách je však tato metoda značně nevhodná. A to hlavně z důvodů vícedimenzionality časových řad či z neodolnosti vůči šumu. Vzniklé stromy jsou udávány jako příliš hluboké a husté. [32] Což znamená výpočetní náročnost

a v kombinaci s udávanou nepřesností je činí nevhodnou volbou.

Tento problém zčásti řeší zavedení metody regresních stromů [33]. Narozdíl od klasifikačních nejsou přiřazovány objektům konkrétní třídy, ale jsou pro ně odhadovány numerické atributy. Obě tyto metody zastřešuje metoda „classification and regression tree“ (CART). Další zlepšování výsledků poté už záleží jen na konkrétních použitých algoritmech. [34]

Často uváděnou nevýhodou u metody CART je fakt, že tato metoda není založena na pravděpodobnostním modelu při vyvozování predikcí, ale spoléhá se pouze na splnění požadované predikce za určených podmínek. Na druhou stranu mezi její výhody patří mimo jiné schopnost vypořádat se s vyšší dimenzionalitou analyzovaných dat [35].

3.4.4 Clustering

Metody shlukování nevyžadují supervizi a reprezentují techniky, kdy jsou datové objekty shlukovány do shluků neboli clusterů, přičemž objekty v clusteru jsou si podobné a zároveň jsou nepodobné objektům v jiných clusterech [10, s. 108]. Jedním z hlavních problémů při identifikování clusterů v datech je specifikace podobnosti objektů a způsob, jak tuto podobnost měřit [36, s. 3].

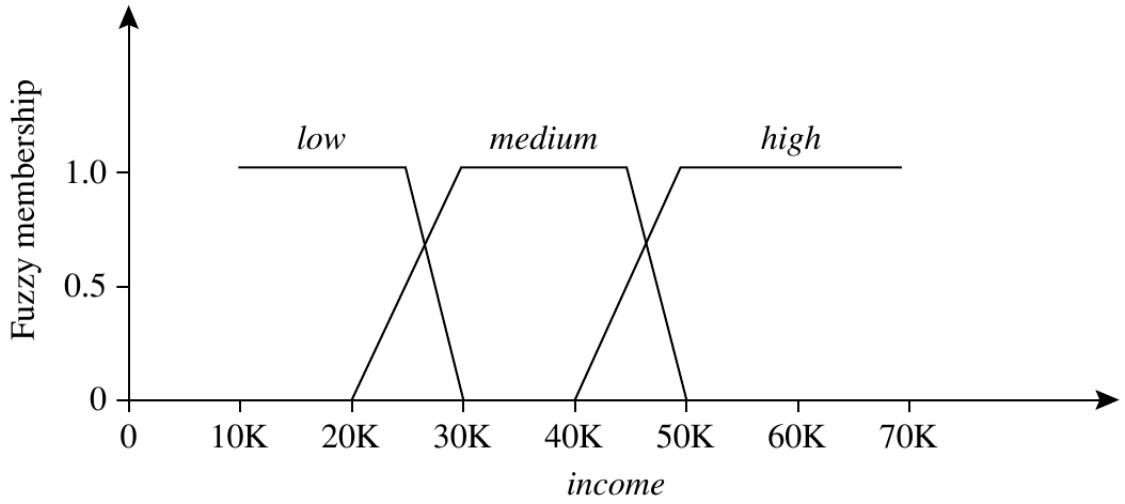
Časové řady je možné shlukovat dle tří základních přístupů. V prvním se uvažuje časová řada jako celek, dále je možné uvažovat dílčí podposloupnosti této časové řady a nakonec samotné dílčí body v časové řadě. [37] Při těchto postupech je možné užívat následujících základních metod clusteringu [37] [10, s. 448–451, 491].

- Partitioning method – nejprve je vytvořena množina k segmentů, kde k představuje počet těchto segmentů. Poté je použita „iterative relocation technique“, která se pokouší o zlepšení segmentace přesouváním dílčích objektů mezi segmenty. Mezi nejznámější metody patří zejména „k-means“.
- Hierarchical method – spočívá ve vytvoření hierarchické struktury. Existují dva přístupy. Prvním je „bottom-up“, kdy dochází ke shlukování menších clusterů do větších, druhým je „top-down“, kdy se jeden velký cluster rozpadá na více menších. Výhodou je přehledná vizualizace, nevýhodou uplatnitelnost spíše na menší datové sady, neboť tato metoda má kvadratickou složitost.
- Density-based method – objekty jsou shlukovány buď na základě hustoty sousedních objektů nebo na základě hustotní funkce.
- Grid-based method – v první fázi jsou objekty uspořádány do mřížkového prostoru a ve druhé fázi je clustering prováděn v rámci tohoto prostoru. Jednou z metod je například STING (z angl. „STatistical INformation Grid“).

Jak se ukazuje, je metoda shlukování úspěšně aplikovatelná například v situaci, kdy je časová řada rozdělena na segmenty, které jsou následně hierarchicky shlukovány metodou bottom-up [39]. Je též uplatnitelná pro hierarchickou top-down metodu [41]. Rozdělením časové řady na podposloupnosti, které jsou poté různými metodami shlukovány, se obecně zabývalo větší množství studií. Výsledky jejich bádání však byly často nejasné a ukazuje se též, že efektivita bývá sporná s ohledem na potřebné paměťové zdroje [37]. Některé zdroje dokonce tvrdí, že shlukování podposloupností je bezvýznamné [40].

3.4.5 Fuzzy logika

Fuzzy logika, respektive fuzzy množiny zjemňují striktní binární klasifikaci, černobílý pohled na věc. Jsou tedy daleko blíže intuitivnímu lidskému uvažování. Pro lepší pochopení uvádím ilustrativní obr. 3.2. Jedná se o klasifikaci platových tříd v závislosti na výši příjmu.



Obrázek 3.2: Příklad fuzzy množin při klasifikaci platových tříd (převzato z [10])

Je zřejmé, že pokud by se klasifikovalo binárně, výsledkem by byly 3 ostré, neprotínající se množiny. Jelikož ale je uvažován fuzzy přístup, je výsledkem věrohodnější popis, kdy každé hodnotě *income* z osy x odpovídá stupeň příslušnosti na ose y z intervalu [0,1].

V kontextu detekce vzorů v časových řadách bývá tato metoda používána nejčastěji ve spolupráci s neuronovými sítěmi [41], případně existují aplikace ve spolupráci se shlukovacími metodami [42]. Zahrnutí fuzzy elementu je též velice populární i ve finanční sféře [43, s. 279]. Dále existují i samotné modely fuzzy časových řad, které najdou uplatnění jak v detekci vzorů v nejistém (ve smyslu nepřesném) prostředí, tak v predikci těchto časových řad [44].

3.5 Popis zvolených dat a jejich významu

Jako data, ve kterých budu vzory detekovat, jsem zvolil finanční časové řady z trhu Forex (z angl. „FOReing EXchange“). Ten minimálně v posledním desetiletí zažívá obrovský rozvoj, co se týče podpory potenciálních investorů. Jedná se zejména o vznik brokerských společností, internetových komunit, vydávání knih a další šíření informační a vzdělávací osvěty. Tím se dostávám k nejpodstatnějšímu důvodu volby těchto časových řad. Lze si relativně snadno opatřit historická data i v minimálním časovém rámci 1M, tedy 1 minuta, a to od dob samotného vzniku trhu Forex. V případě této práce se jedná o data z obchodní platformy společnosti Oanda. Jako další důvod lze uvést fakt, že v těchto časových řadách existují obecně popsané a definované vzory, které je možné hledat.

Na obr. 3.3 je uvedena ukázku grafu, který se skládá z elementárních svící různých druhů, jejichž význam vysvětlím dále. Jedná se o časovou řadu s časovým rámcem 1 hodina (1H). Tato řada charakterizuje vývoj kursu eura (EUR) vůči americkému dolaru (USD). Je zobrazen pouze omezený časový interval, datová sada od roku 1999 pro časový rámec 1H čítá přibližně 100 000 svící. Mým cílem je nalézt v takových řadách všechny výskyty například vzoru zvýrazněného

na obr. 3.3 či jeho „mírně“ odlišné varianty. Takové vzory musí být nejdříve formálně zapsány. Navíc proces vyhledání vzoru je nutno zautomatizovat, jelikož kontrolovat ručně například 8 různých datových sad po 100 000 svících pro několik různých časových rámců není optimální jak z hlediska časového, tak z hlediska chyby lidského faktoru.

Obrázek 3.3: Příklad časové řady reprezentované svícovým grafem

Motivací k vyřešení takového problému může být například snaha získat vstupní informace pro technickou analýzu daného finančního instrumentu, respektive trhu. Na základě mimo jiné těchto znalostí je možné poté například registrovat nově vznikající vzory, které se již objevily v minulosti, a tudíž s určitou pravděpodobností je možné určit další vývoj aktuální situace. Termín „technická analýza“ je možné chápat jako: „*...analýza cenových pohybů, rychlosti jejich změn a objemu z hlediska historie, vychází tedy ze studia minulého tržního chování měny, indexu či komodity...*“ [50]. Podstatná je zejména z toho důvodu, že: „*...je jedním z nejvýznamnějších nástrojů používaných k prognóze chování finančních trhů. Osvědčila se jako efektivní nástroj investorů a stále více účastníků na trhu ji používá...*“[51]. Na tomto místě je vhodné uvést, že svícové grafy se netýkají pouze finanční sféry, ale jejich uplatnění lze nalézt i v jiných oblastech [56][57].

3.5.1 Struktura grafu a dat

Jako trénovací sadu dat pro tuto práci jsem vybral měnový pár EURUSD (euro/americký dolar) s časovým rámcem 1H a jako testovací datovou sadu jsem zvolil USDCHF (americký dolar/švýcarský frank) se stejným časovým rámcem. Důvodem volby těchto párů je fakt, že se jedná o jedny ze 7 hlavních a také nejvíce obchodovaných párů [52]. Navíc se jedná o měnový

pár s negativní korelací limitně se blížící hodnotě *c = -1* [70][71]. To znamená, že pokud kurs EURUSD poroste, kurs USDCHF poklesne o stejnou hodnotu a vice versa. Dá se tedy s určitým zobecněním předpokládat, že klasifikátor vytvořený na základě trénovací množiny reprezentované daty měnového páru EURUSD je úspěšně aplikovatelný na testovací množinu reprezentovanou měnovým párem USDCHF. Pakliže by byla uvažována jako testovací množina dat jiný měnový pár než USDCHF, bylo by vhodné data nejdříve analyzovat, aby se zjistilo, zdali je tento postup korektní.

Časový rámec si lze představit jako časový interval, který je charakterizován nejčastěji 4 případně 5 hodnotami, které shrnují dění na trhu (měnovém páru) v daném časovém intervalu. Nejčastěji se jedná o rámce o hodnotách 1, 5, 15, 30, 60, 240, 1 440 nebo i 10 080, kde číselná hodnota reprezentuje délku časového rámce v minutách. Je zřejmé, že čím větší je délka časového rámce, tím menší je počet svící v datové sadě a tím větší je jejich informační hodnota. Mnou zvolený časový rámec 1H je běžně udáván v kontextu intradenního obchodování [69]. Pokud bych chtěl provádět hlubší analýzu detekce vzorů, bylo by vhodné, abych bral v úvahu různé časové rámce.

Zmíněných 5 hodnot se běžně označuje jako *Open*, *High*, *Low*, *Close* a *Volume*. Představují následující informace:

- *Open* – otvírací cena; hodnota v počátku časového rámce
- *High* – nejvyšší dosažená cena v časovém rámci, za kterou se obchodovalo
- *Low* – nejnižší dosažená cena v časovém rámci, za kterou se obchodovalo
- *Close* – uzavírací cena; hodnota v konci časového rámce
- *Volume* – množství zobchodované měny

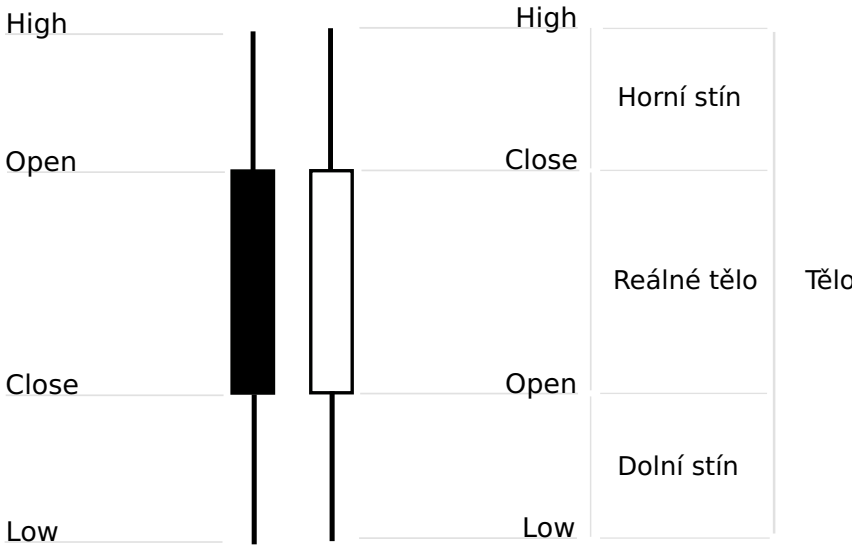
Následuje ukázka, v jakém formátu jsou použitá historická data exportována z obchodní platformy do souboru ve formátu CSV. Každý řádek reprezentuje jednu svící, celkem tedy má soubor přibližně dříve zmíněných 100 000 řádků. Jedná se o měnový pár EURUSD, časový rámec 1H.

```
Datum, Čas, Open, High, Low, Close, Volume
...
2011.11.11,12:00,1.37769,1.37944,1.37466,1.37473,4694
2011.11.11,13:00,1.37472,1.37706,1.37446,1.37523,3522
2011.11.11,14:00,1.37527,1.37716,1.37526,1.37672,1787
...
```

3.5.2 Svícový graf

V praxi se k zobrazování uvedených hodnot používají nejčastěji schodové či svícové grafy. Oba dva typy spadají do množiny „OHLC“ (*Open, High, Low, Close*) grafů. Hlavní rozdíl mezi nimi je v čitelnosti pro člověka. Je pak pochopitelné, že se prosadily svícové grafy na úkor schodových a jsou dnes de facto standardem. Ačkoliv jsou svícové grafy starší než schodové, masově se rozšířily relativně nedávno – jejich expanze po světě začala z Japonska přibližně v roce 1989 díky Stevu Nisonovi. [53]. Dále se budu zabývat již jen svícovými grafy.

Elementárním prvkem svícového grafu je svíce, která je charakterizována čtyřmi, respektive pěti parametry. Základní typy svící [54, s. 3][55, s. 4] se základními i odvozenými parametry jsou znázorněny na obr. 3.4.



Obrázek 3.4: Základní typy svící s vyznačenými parametry svící

Z něj je zřejmé, že černá svíce reprezentuje stav, kdy kurs měny klesá, jelikož hodnota *Open* je větší než hodnota *Close*. U bílé svíce je tomu naopak, tedy hodnota *Open* je menší než hodnota *Cl*ose, takže kurs měny roste.

Svíce mohou nabývat různých dílčích parametrů a podle toho se také bude měnit jejich podoba – například pozice a velikost reálného těla a s tím související délka stínů, apod. Bližší specifikací parametrů se budu zabývat dále.

3.5.3 Trend

Tento pojem si lze zjednodušeně představit jako směr, kterým se vydává trh, respektive časová řada reprezentována svícemi. Buď řada poroste, bude klesat nebo bude stagnovat. Pokud bych chtěl být opravdu korektní a chtěl bych, aby tato práce měla o něco větší praktický přínos, musel bychom pro každou svícovou formaci uvažovat i trend, v jakém se formace nachází. Neboť jak uvádí Morris [58, s. 212–213], existence daného vzoru ve svícovém grafu není dána pouze vztahem mezi daty, které reprezentují vzor, ale je také dána trendem, který předchází výskytu tohoto vzoru. Což, jak dodává, je v současné tématické literatuře často opomíjeno.

3.5.4 Klouzavý průměr

K určování trendu je možné použít nástroj zvaný klouzavý průměr (MA), který časovou řadu „vyhlazuje“, takže nejsou tolik patrná lokální minima či maxima. V této práci jej využiji pro stanovování průměrné velikosti reálného těla svíce. Tato velikost je určena pro další klasifikaci svíce a jejich vzorů, jak ukáži v další kapitole. Je totiž nutné si uvědomit, že je potřeba uvažovat svíci v určitém omezeném časovém kontextu, není možné použít například aritmetický průměr či medián dat za např. posledních 10 let, výsledky by nebyly přesné. Klouzavý průměr zadefinujeme následovně [59].

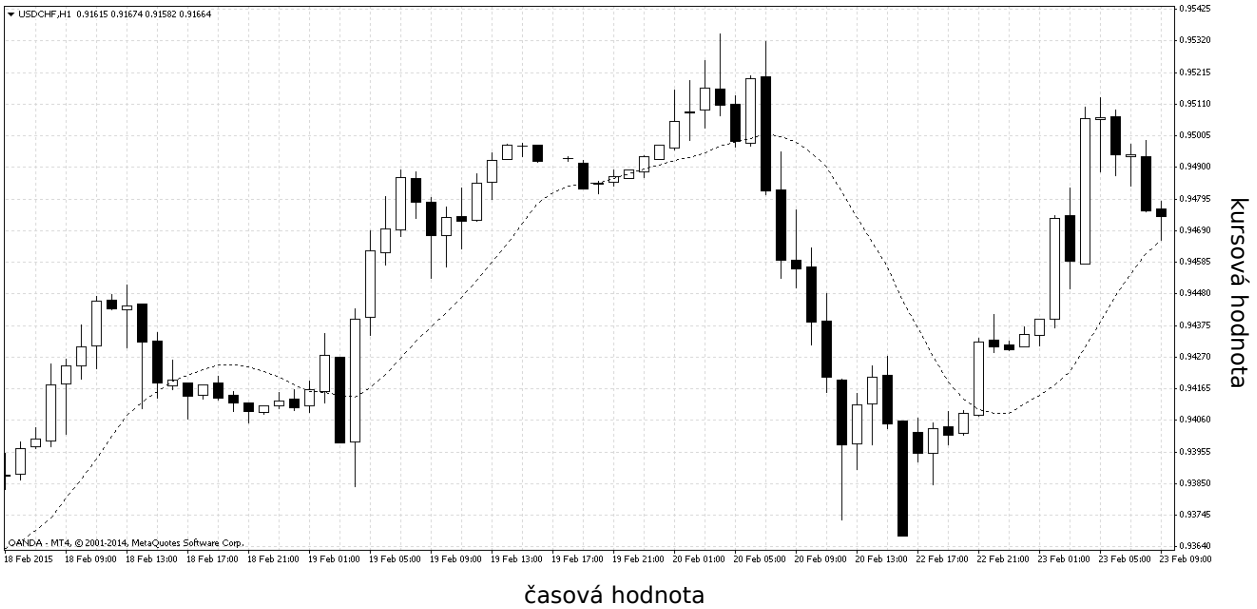
Definice 7. Mějme posloupnost P reálných čísel $r_1, ..., r_n$. Klouzavým průměrem $(MA)_b$ s bází b pro prvek r_n nazveme takovou posloupnost reálných čísel, pro kterou platí, že $(MA)_b = \frac{\sum_{i=1}^{b} D_i}{b}$, kde D_i představuje prvky posloupnosti $r_{n-1}, ..., r_{n-b}$.

Jelikož tato definice nemusí být srozumitelná, uvádím následující praktický příklad. Převzato a upraveno [60].

Příklad 1. Mějme posloupnost $P = (2, 4, 6, 5, 4, 3, 2, 4)$, která reprezentuje hodnoty Close pro hodinový časový rámec. Dále mějme bázi $b = 3$, která reprezentuje, pro jak dlouhá období chceme klouzavý průměr počítat.

1. člen MA se spočte jako: $\frac{2+4+6}{3} = 4$.
2. člen MA se spočte jako: $\frac{4+6+5}{3} = 5$.
3. člen MA se spočte jako: $\frac{6+5+4}{3} = 5$.
...
Nakonec získám klouzavý průměr pro tuto posloupnost jako: $4, 5, 5, 4, 3, 3, 3$. Je tedy zřejmé, že klouzavý průměr je de facto aritmetický průměr za určité období.

Ilustrační zobrazení klouzavého průměru je na obr. 3.5. Jedná se o měnový pár americký dolar/švýcarský frank na časovém rámci 1 hodina. Báze klouzavého průměru odpovídá 12 hodinám. Je patrné, že přerušována křivka reprezentující klouzavý průměr hodnot *Close* má z definice zpoždění a časovou řadu opravdu „vyhlazuje“.



Obrázek 3.5: Demonstrace MA na měnovém páru USDCHF 1H pro b = 12

Mimo tento typ klouzavého průměru, který se též nazývá jednoduchý klouzavý průměr, existují i další typy klouzavých průměrů, například exponenciální či vážený. V této práci uvažuji jen jednoduchý.

4. Aplikace zvolených metod

4.1 Volba vhodných metod pro detekci vzorů

Pro detekci vzorů v časových řadách jsem se rozhodl využít metod rule-based, fuzzy a modifikované klasifikační z oblasti učení s učitelem, což rozeberu dále. Hlavními prioritami pro mě byla transparentnost a intuitivnost použitých metod jako i dostupná možnost vlastního návrhu a programové implementace od základů společně s klasifikační architekturou. V kombinaci se zvolenými daty se jednalo o přijatelnou volbu, jelikož příbuzným tématem a metodami se již některé studie zabývaly [45].

4.1.1 Rule-based metoda

Pro tuto metodu není nutné rozebírat další detaily, vystačí zde popis, který byl již rozebírán v teoretické části.

4.1.2 Fuzzy množiny

S ohledem na výběr fuzzy metody se v této části zaměřím na bližší popis a použití fuzzy množin, což jsem dříve popsal jen zběžně. Tyto poznatky využiji k praktickému návrhu.

Definice 4. Mějme pevně zvolenou univerzální množinu X. Fuzzy množinou A univerza X budeme rozumět objekt popsaný charakteristickou funkcí

$$\mu_A : X \to [0, 1], \tag{4.1}$$

kterou též nazýváme funkce příslušnosti. Pro každý prvek $x \in X$ hodnota $\mu_A(x) \in [0, 1]$ říká, do jaké míry je x prvkem fuzzy množiny A. Každá funkce z X do [0, 1] určuje jednoznačně nějakou fuzzy množinu. [48]

Libovolnou fuzzy množinu je možné popsat její funkcí příslušnosti, která je též známá jako členská funkce. Jako příklad [48] je možné uvést univerzum $X = \mathbf{R}$ a množiny A, B, které je možné zapsat předpisem

$$\mu_A(x) = \begin{cases} 0 & \text{pro } x < 0, \\ x & \text{pro } x \in [0, 1], \\ 2 - x & \text{pro } x \in (1, 2], \\ 0 & \text{pro } x > 2, \end{cases}$$

$$\mu_B(x) = \begin{cases} \frac{1}{2} & \text{pro } x = 3, \\ 1 & \text{pro } x = 4, \\ \frac{1}{4} & \text{pro } x = 5, \\ 0 & \text{jinak.} \end{cases}$$

Pro ilustrační zobrazení fuzzy množin odkáži opět na obr. 3.2. Mimo toto zobrazení funkce příslušnosti, známé též jako lichoběžníkové, existují i další základní typy. Například trojúhelníkové, „Gaussian“, „illogical“, „asymmetrical Gaussian“ a další. [47, s. 12–15]

Pro potřeby této práce ještě popíši některé logické operace nad fuzzy množinami. Pro tyto účely mějme fuzzy množiny M^1, M^2 definované pro x z univerza X a s nimi asociované členské

funkce $\mu^1(x)$, $\mu^2(x)$ [47, s. 16–17].

Definice 5. Fuzzy konjunkcí (AND) $M^1 \cap M^2$ nazveme takovou fuzzy množinu, pro kterou bude platit $\mu^{M^1 \cap M^2}(x) = \min \{\mu^1(x), \mu^2(x) : x \in X\}$.

Definice 6. Fuzzy disjunkcí (OR) $M^1 \cup M^2$ nazveme takovou fuzzy množinu, pro kterou bude platit $\mu^{M^1 \cup M^2}(x) = \max \{\mu^1(x), \mu^2(x) : x \in X\}$.

Z toho je tedy zřejmé, že se budu držet klasického přístupu zakladatele fuzzy logiky a fuzzy množin, kterým je Lotfali Askar Zadeh. V kontextu fuzzy přístupu je nutné popsat další procesy, které souvisí s jejich praktickým použitím. Jedná se zejména o následující kroky [49, s. 18].

- Fuzzification – spočívá v převodu klasických či ostrých dat do fuzzy dat nebo do členských funkcí; jedná se například o definici oblastí *low*, *medium* a *high* na obr. 3.2
- Fuzzy inference process – spočívá v kombinaci členských funkcí společně se zvolenými pravidly, čímž tvoří *fuzzy output*
- Defuzzification – na základě vstupu vybírá konkrétní „fuzzy output“; jedná se například o klasifikaci konkrétní třídy

4.1.3 Modifikovaná klasifikační metoda

Nechť existují 2 datové sady, které slouží jako trénovací a testovací množina. V první sadě naleznu požadované vzory metodami rule-based a fuzzy. Z takto nalezených vzorů následně vytvořím matici průměrného vzoru, tedy jak by měl pravděpodobně vypadat ideální vzor. Tímto postupem si de facto vytvořím 2 trénovací množiny. Poté využiji druhou sadu, testovací, kde se pokusím tyto vzory nalézt na základě procentuální odchylky testovaných dat od ideálního, průměrného vzoru. Bližší popis jako i implementační detaily této metody uvedu záhy.

4.2 Modelace svící a vzorů

V této části zmíním obecné základní poznatky k modelaci svící a jejich vzorů, které platí napříč metodami. Modelaci v rámci konkrétních metod rozvedu v následující sekci 4.3.

4.2.1 Úvodní slovo k modelaci

Než se pustím do formalizace svící, vzorů a jejich modelace, považuji za vhodné uvést některá fakta. Zdrojem pro tuto sekci je *Encyclopedia of Candlestick Charts*, jejímž autorem je Thomas N. Bulkowski [55]. Ten strávil analýzou grafových vzorů značnou část svého života, na toto téma napsal několik knih a je možné jej považovat za autoritu v této oblasti. V této konkrétní knize analyzoval a třídil svícové grafy s ohledem na frekvenci jejich výskytu, schopnost měnit trend, dále i s ohledem na jejich schopnost generovat zisk atp. Analýzu prováděl na datech, která reprezentují kompletní akciový trh S&P 500 za dobu 10 let [55, s. 4].

Zde je vhodné podotknout, že analýza akciových trhů se příliš neslučuje s mnou analyzovaným trhem Forex. Je však vhodné brát v úvahu, že v knize uvedené svícové formace či elementární svíce jsou až na výjimky k nalezení v de facto libovolném finančním trhu a běžně jsou v komunitě používány. Jak autor sám dále uvádí, tak ostatní výzkumníci mohou docházet k jiným výsledků než k těm, které prezentuje on. Dle něj mohou být na vině zejména metody detekce vzorů, data použitá při testování, použitá časová perioda, ale též i různá měřítka

výkonnosti dílčích vzorů. [55, s. 6]

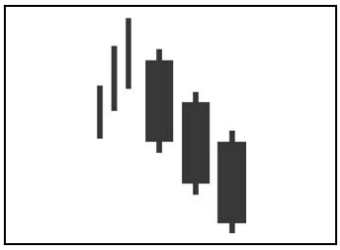
V tomto kontextu považuji dále za vhodné zmínit též výrok mezivládní vědecké organizace IPCC (*Intergovernmental Panel on Climate Change*), která je nechvalně známá v souvislosti s tzv. „hockey stick controversy“ [63][65] či o něco později s aférou obecně známou jako „Climategate“ [64][66]:

„Plně uznáváme, že mnohá z uvedených tvrzení jsou do jisté míry založena na subjektivním vědeckém vnímání a obsahují komunitií a osobní vědomosti. Například pouhý výběr proměnných a procesů, které jsou do modelu zahrnuty, je většinou založen pouze na dojmech a zkušenostech modelovací komunity.“ [62] Převzato z [23, s. 105].

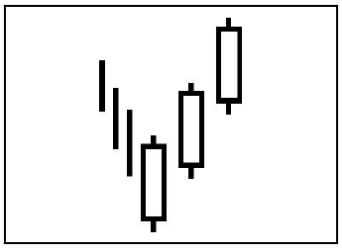
Nemělo by se taktéž zapomínat na to, že tato práce se nezabývá analýzou a tříděním dle různých kritérií jako *Encyclopedia of Candlestick Charts*, nýbrž jen samotným formulováním a vyhledáním vzorů. Považoval jsem však za vhodné výše uvedené skutečnosti zmínit a uvést na pravou míru.

4.2.2 Volba vzorů

Jako demonstrační vzory pro schopnost rozpoznání jsem zvolil 2 poměrně základní a velmi známé vzory, které uvádí Bulkowski [55]. Jedná se o vzor „three black crows“ znázorněn na obr. 4.1 a vzor „three white soldiers“ na obr. 4.2.



Obrázek 4.1: Ilustrační zobrazení vzoru three black crows (převzato z [55])



Obrázek 4.2: Ilustrační zobrazení vzoru three white soldiers (převzato z [55])

V případě three black crows se jedná o vzor tří po sobě jdoucích *dlouhých* černých svící s *krátkými* stíny a s klesající tendencí *Open* a *Close* hodnot svící, což reprezentuje pokles daného kurzu. Analogie je v případě three white soldiers zjevná. Pojem jako „dlouhých“ či „krátkými“ je značně vágní a ani autoři samotní jej exaktně nedefinují, ačkoliv jisté snahy existují, viz [58, s. 215–218][67, s. 16–21]. Z těch také budu dále vycházet. Vzor three black crows dále budu označovat jako vzor „**A**“, vzor three white soldiers jako vzor „**B**“. Dále je vhodné dodat, jak též uvádí Bulkowski, že exaktní podobu těchto vzorů fakticky nelze nalézt, vždy je nutné

uvažovat jistou vizuální odlišnost. Též uvádí, že u těchto vzorů stačí kontrolovat podmínku barvy svící a délky jejich reálných těl a dále je možné si vystačit jen s hodnotami *Open* či jen s hodnotami *Close*. Údajně na základě jeho pozorování, jaká je nízká pravděpodobnost, že se vyskytnou těsně za sebou 3 značně nadprůměrně dlouhé svíce stejné barvy s rostoucí nebo klesající tendencí.

4.2.3 Parametry modelu dílčí svíce

Při vytváření modelů svíce jsem částečně vyšel z [45]. Parametry je možné si pracovně rozdělit na „statické“ a „dynamické“. V případě statických parametrů se jedná o hodnoty, které nezávisí na modelacích, které provádějí různí autoři.

Mezi statické parametry svíce se řadí barva svíce. Ta je bez újmy na obecnosti dvojí: černá a bílá. Svíci typu „doji“ – kdy hodnota *Open* odpovídá hodnotě *Close* a svíce tedy barvu nemá – zde s ohledem na vybrané demonstrační vzory neuvažuji. Dále mezi statické parametry patří numerické hodnoty velikosti těla svíce, velikosti reálného těla svíce, jako i délky stínů svíce.

Za dynamické parametry je možné považovat dělení statických parametrů do velikostních tříd. Délku reálného těla svíce jsem rozdělil do 5 velikostních tříd: `XS`, `S`, `M`, `L`, `XL`. Klasifikaci do konkrétní třídy jsem určoval na základě hodnoty p_1, podílu velikosti reálného těla svíce s klouzavým průměrem reálného těla svíce za posledních N časových rámců, tedy

$$p_1 = \frac{\textit{velikost reálného těla svíce}}{\textit{klouzavý průměr reálného těla svíce za posledních } N \textit{ rámců}}.$$

Autoři se obecně neshodují na konkrétní bázi klouzavého průměru, jelikož záleží na konkrétních obchodních strategiích, nicméně je možné vyjít z báze 21 dnů, která je v obchodní komunitě relativně běžná a uznávaná [68].

Dále jsem uvažoval ve 3 velikostních třídách délku stínu svíce. Jedná se o třídy: `S`, `M`, `L`. Klasifikaci do konkrétní třídy jsem určoval na základě podílu p_2, který stíny tvoří v těle svíce neboli

$$p_2 = \frac{\textit{velikost horního stínu + velikost dolního stínu}}{\textit{velikost těla svíce}}.$$

Jelikož parametr *Volume* nemá žádný vliv na to, jak libovolná svíce vypadá, dovolil jsem si provést odebrání tohoto parametru. V případě, že bych se zabýval rozpoznáváním vzorů v kontextu například ekonomického přínosu či predikce vývoje časové řady, pak bych jej zohledňovat měl.

4.2.4 Parametry modelu svícových vzorů

Při modelaci vzoru **A** vyjdu z následujících předpokladů, které musí platit zároveň.

- vzor se skládá ze tří těsně po sobě následujících černých svící C_n, C_{n+1}, C_{n+2}
- velikostní třída reálného těla každé svíce je `XL`

- velikostní třída stínu každé svíce je `S`
- pro posloupnost hodnot *Open* svící C_n, C_{n+1}, C_{n+2} platí, že je klesající

Při modelaci vzoru **B** vyjdu z následujících předpokladů, které musí platit zároveň.

- vzor se skládá ze tří těsně po sobě následujících bílých svící C_n, C_{n+1}, C_{n+2}
- velikostní třída reálného těla každé svíce je `XL`
- velikostní třída stínu každé svíce je `S`
- pro posloupnost hodnot *Open* svící C_n, C_{n+1}, C_{n+2} platí, že je rostoucí

4.3 Modelace svící a vzorů navrženými metodami

V této části již uvádím modelace svící a vzorů pomocí konkrétních metod společně s numerickými hodnotami a pseudokódy. Popsané metody tedy již umožňují vyhledávání vzorů.

4.3.1 Určení základních parametrů svící

Nejdříve popíši, jak jsem určoval základní parametry dílčích svící. Jmenovitě barvu svíce, dále velikosti horního a dolního stínu a nakonec velikost těla svíce, jako i velikost reálného těla svíce. Všechny tyto parametry jsou nezávislé na použití metod a platí tedy univerzálně. Je důležité zmínit, že pokud chci klasifikovat velikostní třídy jako i vzory, musím nejdříve určit hodnoty uvedených parametrů.

V následujícím pseudokódu popisuji určení barvy `COLOR` svíce `C`, která je charakterizována hodnotami `OPEN` a `CLOSE`.

```
Pseudokód pro přiřazení barvy svíce
1: if (C.OPEN > C.CLOSE) then
2:     C.COLOR = BLACK
3: end if
4: if (C.OPEN < C.CLOSE) then
5:     C.COLOR = WHITE
6: end if
```

Dále popisuji přiřazení velikostí stínů pro svíci `C`, která je charakterizována hodnotami `OPEN`, `HIGH`, `LOW`, `CLOSE`. Velikost horního stínu značím `UPSHADOW`, velikost dolního stínu `LOWSHADOW`.

```
Pseudokód pro přiřazení velikosti stínů svíce
1: if (C.OPEN ≥ C.CLOSE) then
2:     C.UPSHADOW = C.HIGH - C.OPEN
3:     C.LOWSHADOW = C.CLOSE - C.LOW
4: end if
5: if (C.OPEN < C.CLOSE) then
6:     C.UPSHADOW = C.HIGH - C.CLOSE
7:     C.LOWSHADOW = C.OPEN - C.LOW
8: end if
```

Nakonec popisuji přiřazení velikosti těla svíce `BS`, dále velikosti reálného těla svíce `RBS` a též délky stínu `SL` pro svíci `C`. Ta je opět charakterizována hodnotami `OPEN`, `HIGH`, `LOW`, `CLOSE`. Zkratka `abs` představuje absolutní hodnotu.

```
Pseudokód pro přiřazení velikosti těla svíce, reálného těla svíce a délky stínu svíce
1: C.BS = C.HIGH - C.LOW
```

2: C.RBS = abs (C.OPEN - C.CLOSE)
3: C.SL = C.UPSHADOW + C.LOWSHADOW

Tím jsem určil základní parametry svící, dále již rozeberu konkrétní metody, které klasifikují velikostní třídy reálného těla svíce a velikosti stínu svíce.

4.3.2 Rule-based metoda

V následujících pseudokódech popisuji metody klasifikace délky reálného těla `RBS` svíce `C`. Značení `MA(RBS)` reprezentuje klouzavý průměr délky reálného těla svíce `C` za dříve zmíněných 21 dní. Jedná se o 5 metod pro 5 velikostních tříd.

```
Pseudokód metody, která určuje, zdali je velikostní třída reálného těla svíce XS
```
1: **boolean** isRBSTypeXS()
2: **if** (C.RBS ≤ C.MA(RBS) × 0.1) **then**
3: **return** true
4: **end if**
5: **return** false

```
Pseudokód metody, která určuje, zdali je velikostní třída reálného těla svíce S
```
1: **boolean** isRBSTypeS()
2: **if** C.RBS > C.MA(RBS) × 0.1 **and** C.RBS ≤ 0.65 × C.MA(RBS) **then**
3: **return** true
4: **end if**
5: **return** false

```
Pseudokód metody, která určuje, zdali je velikostní třída reálného těla svíce M
```
1: **boolean** isRBSTypeM()
2: **if** (C.RBS > C.MA(RBS) × 0.65 **and** C.RBS ≤ 1.35 × C.MA(RBS) **then**
3: **return** true
4: **end if**
5: **return** false

```
Pseudokód metody, která určuje, zdali je velikostní třída reálného těla svíce L
```
1: **boolean** isRBSTypeL()
2: **if** (C.RBS > C.MA(RBS) × 1.35 **and** C.RBS ≤ 1.55 × C.MA(RBS) **then**
3: **return** true
4: **end if**
5: **return** false

```
Pseudokód metody, která určuje, zdali je velikostní třída reálného těla svíce XL
```
1: **boolean** isRBSTypeXL()
2: **if** (C.RBS > C.MA(RBS) × 1.55) **then**
3: **return** true
4: **end if**
5: **return** false

Nyní popíši metody klasifikace délky stínu `SL` pro svíci `C`. Zřejmě platí, že délka stínu je dána součtem horního a dolního stínu neboli `SL = UPSHADOW + LOWSHADOW`. Zkratka `BS` odpovídá velikosti těla svíce `C`.

```
Pseudokód metody, která určuje, zdali je velikostní třída stínu svíce S
```
1: **boolean** isSLTypeS()
2: **if** (C.SL / C.BS ≤ 0.45) **then**
3: **return** true
4: **end if**
5: **return** false

```
Pseudokód metody, která určuje, zdali je velikostní třída stínu svíce M
```
1: **boolean** isSLTypeM()
2: **if** (C.SL / C.BS > 0.45 **and** C.SL / C.BS ≤ 0.75) **then**

```
3:    return true
4: end if
5: return false
```

`Pseudokód metody, která určuje, zdali je velikostní třída stínu svíce L`

```
1: boolean isSLTypeL()
2: if (C.SL / C.BS > 0.75) then
3:    return true
4: end if
5: return false
```

Nakonec uvádím pseudokód pro metody, které ověřují barvu `COLOR` svíce `C`.

`Pseudokód metody, která určuje, zdali je svíce černá`

```
1: boolean isBlack()
2: if (C.COLOR == BLACK) then
3:    return true
4: end if
5: return false
```

`Pseudokód metody, která určuje, zdali je svíce bílá`

```
1: boolean isWhite()
2: if (C.COLOR == WHITE) then
3:    return true
4: end if
5: return false
```

Tím jsem tedy popsal fundamentální metody, které jsou dále použity mimo jiné k detekci vzorů **A** a **B** v datové sadě. Pseudokódy pro vyhledání těchto vzorů za použití výše zmíněných metod jsou uvedeny následovně. Je zřejmé, že vycházím z předpokladů uvedených v sekci 4.2.4.

`Pseudokód pro detekci vzoru A pomocí rule-based`

```
1: if C1.isBlack() and C1.isRBSTypeXL() and C1.isSLTypeS() then
2:    if  C2.isBlack() and C2.isRBSTypeXL() and C2.OPEN < C1.OPEN and C2.isSLTypeS() then
3:       if  C3.isBlack() and C3.isRBSTypeXL() and C3.OPEN < C2.OPEN and C3.isSLTypeS() then
4:          A found
5:       end if
6:    end if
7: end if
```

`Pseudokód pro detekci vzoru B pomocí rule-based`

```
1: if C1.isWhite() and C1.isRBSTypeXL() and C1.isSLTypeS() then
2:    if  C2.isWhite() and C2.isRBSTypeXL() and C2.OPEN > C1.OPEN and C2.isSLTypeS() then
3:       if  C3.isWhite() and C3.isRBSTypeXL() and C3.OPEN > C2.OPEN and C3.isSLTypeS() then
4:          B found
5:       end if
6:    end if
7: end if
```

4.3.3 Fuzzy metoda

Fuzzy metoda se od rule-based metody liší pouze v definici délky reálných těl svící a délky stínu. Pro délku reálného těla svíce jsem vycházel opět ze stejných tříd a numerických hodnot jako při metodě rule-based, stejně tak pro délku stínu svíce. S tím rozdílem, že jsem pochopitelně zavedl fuzzy množiny a to následovně.

Pro délku reálného těla svíce jsem navrhnul fuzzy množiny:

$$\mu_{XS}(x) = \begin{cases} 1 & \text{pro } x \leq a, \\ \frac{b-x}{b-a} & \text{pro } a < x < b \\ 0 & \text{pro } x \geq b, \end{cases}$$

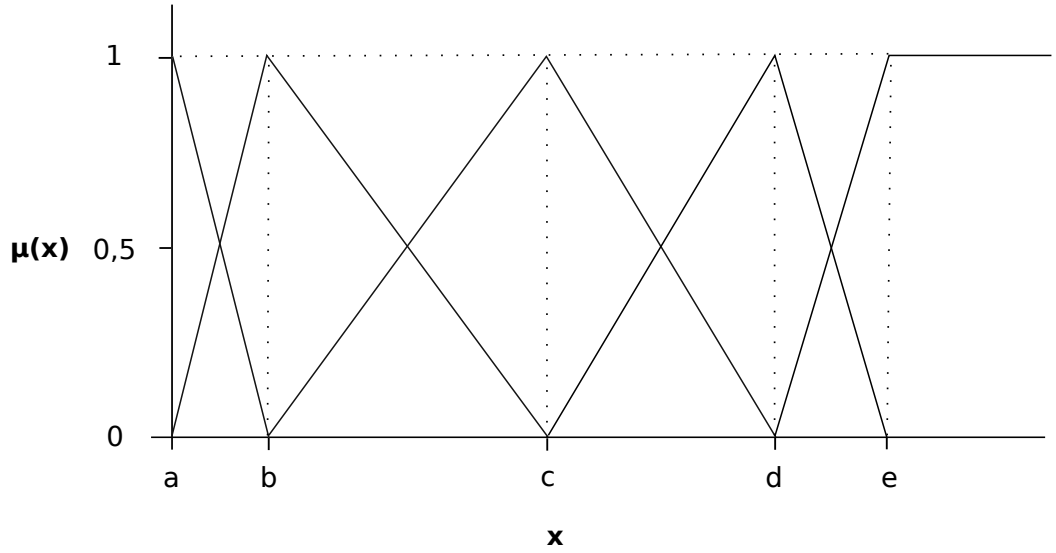
$$\mu_S(x) = \begin{cases} 1 & \text{pro } x = b, \\ \frac{x-a}{b-a} & \text{pro } b > x > a, \\ \frac{c-x}{c-b} & \text{pro } c > x > b, \\ 0 & \text{pro } a \geq x \geq c, \end{cases}$$

$$\mu_M(x) = \begin{cases} 1 & \text{pro } x = c, \\ \frac{x-b}{c-b} & \text{pro } c > x > b, \\ \frac{d-x}{d-c} & \text{pro } d > x > c, \\ 0 & \text{pro } b \geq x \geq d, \end{cases}$$

$$\mu_L(x) = \begin{cases} 1 & \text{pro } x = d, \\ \frac{x-c}{d-c} & \text{pro } d > x > c, \\ \frac{e-x}{e-d} & \text{pro } e > x > d, \\ 0 & \text{pro } c \geq x \geq e, \end{cases}$$

$$\mu_{XL}(x) = \begin{cases} 1 & \text{pro } x \geq e, \\ \frac{x-d}{e-d} & \text{pro } e > x > d \\ 0 & \text{pro } x \leq d, \end{cases}$$

kde $(a; b; c; d; e) = (0; 0,1; 0,65; 1,35; 1,55)$ a $x = \frac{C.RBS}{C.MA(RBS)}$. `RBS` opět reprezentuje délku reálného těla svíce `C` a `MA(RBS)` klouzavý průměr délky reálného těla svíce `C`. Zakreslení těchto členských funkcí je zobrazeno na obr. 4.3.



Obrázek 4.3: Členské funkce pro určení délky reálného těla svíce

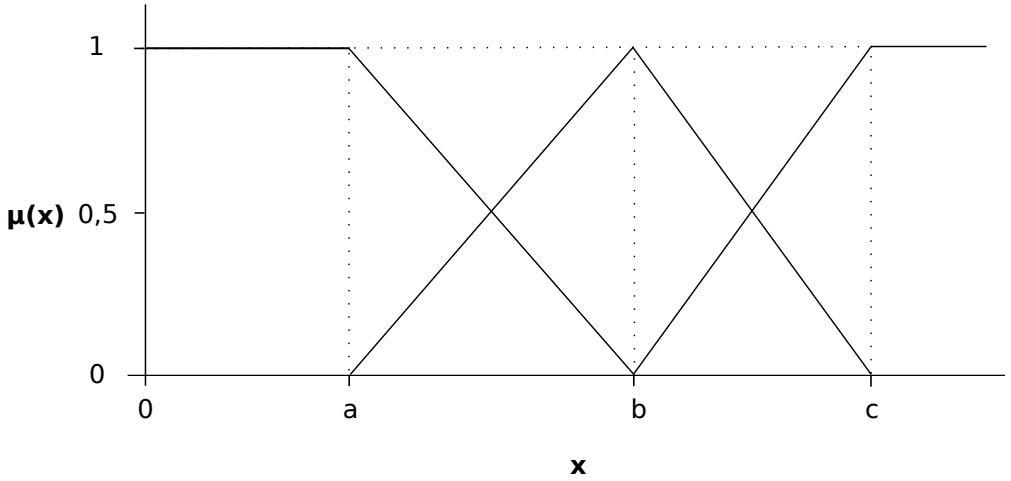
Pro délku stínu svíce jsem navrhnul tyto fuzzy množiny:

$$\mu_S(x) = \begin{cases} 1 & \text{pro } x \leq a, \\ \frac{b-x}{b-a} & \text{pro } a < x < b \\ 0 & \text{pro } x \geq b, \end{cases}$$

$$\mu_M(x) = \begin{cases} 1 & \text{pro } x = b, \\ \frac{x-a}{b-a} & \text{pro } b > x > a, \\ \frac{c-x}{c-b} & \text{pro } c > x > b, \\ 0 & \text{pro } a \geq x \geq c, \end{cases}$$

$$\mu_L(x) = \begin{cases} 1 & \text{pro } x \geq c, \\ \frac{x-b}{c-b} & \text{pro } c > x > b \\ 0 & \text{pro } x \leq b, \end{cases}$$

kde $(a; b; c) = (0,3; 0,6; 0,9)$ a $x = \frac{SL}{BS}$. Z těchto numerických hodnot jsem též vycházel při definici numerických hodnot velikosti stínu svíce pro rule-based přístup, kdy jsem uvažoval poloviny intervalů, tedy $\frac{0,3+0,6}{2} = 0,45$ a $\frac{0,6+0,9}{2} = 0,75$. Připomenu, že parametr `BS` reprezentuje délku těla svíce a `SL` délku stínu svíce. Zakreslení členských funkcí pro tyto množiny je zobrazeno na obr. 4.4.



Obrázek 4.4: Členské funkce pro určení délky stínu svíce

Pro zařazení do velikostních tříd využívám metodu `assignSLFuzzyType` pro velikost stínu a dále `assignRBSFuzzyType` pro velikost reálného těla svíce. Pro určování konkrétních tříd jsem volil fuzzy konjunkci definovanou v úvodu praktické části. S ohledem na rozsáhlost kódu, viz též uvedený zápis fuzzy množin, zde neuvádím kompletní kód metod. Jedná se de facto jen o přepis matematického zápisu těchto množin do programovacího jazyka. Ke zhlédnutí je v přiložených zdrojových kódech, avšak část kódu vysvětlím v sekci, kde se věnuji programové implementaci.

V další fázi již procházím datovou sadu a vyhledávám vzory analogicky jako v případě metody rule-based. Opět uvádím pseudokódy s dodatkem, že metoda `isSLFuzzyS` vrací true, pokud je délka stínu svíce `C` ve třídě `S`. Metoda `isRBSFuzzyXL` vrací `true`, pokud je délka reálného těla svíce `C` ve třídě `XL`. Detaily implementace opět zmíním dále.

```
Pseudokód pro detekci vzoru A pomocí fuzzy
 1: if C1.isBlack() and C1.isRBSFuzzyXL() and C1.isSLFuzzyS() then
 2:     if  C2.isBlack() and C2.isRBSFuzzyXL() and C2.OPEN < C1.OPEN and C2.isSLFuzzyS() then
 3:         if  C3.isBlack() and C3.isRBSFuzzyXL() and C3.OPEN < C2.OPEN and C3.isSLFuzzyS() then
 4:             A found
 5:         end if
 6:     end if
 7: end if
```

```
Pseudokód pro detekci vzoru B pomocí fuzzy
 1: if C1.isWhite() and C1.isRBSFuzzyXL() and C1.isSLFuzzyS then
 2:     if  C2.isWhite() and C2.isRBSFuzzyXL() and C2.OPEN > C1.OPEN and C2.isSLFuzzyS() then
 3:         if  C3.isWhite() and C3.isRBSFuzzyXL() and C3.OPEN > C2.OPEN and C3.isSLFuzzyS() then
 4:             B found
 5:         end if
 6:     end if
 7: end if
```

4.3.4 Modifikovaná klasifikační metoda

Nechť jsem v trénovací datové sadě EURUSD nalezl N vzorů **A** (analogicky též pro vzor **B**) pomocí metody rule-based. Vyberu-li libovolný z nich, vztahy mezi dílčími svícemi tohoto vzoru lze zapsat dle tab. 4.1 o 4 řádcích pro parametry *Open*, *High*, *Low*, *Close* a 3 sloupcích pro počet svící ve vzoru. Dílčí prvky matice reprezentují rozdíly parametrů svící.

$$\begin{array}{c} \\ Open \\ High \\ Low \\ Close \end{array} \begin{array}{c} \begin{array}{ccc} C_1 - C_2 & C_2 - C_3 & C_1 - C_3 \end{array} \\ \begin{pmatrix} \text{C1.O} - \text{C2.O} & \text{C2.O} - \text{C3.O} & \text{C1.O} - \text{C3.O} \\ \text{C1.H} - \text{C2.H} & \text{C2.H} - \text{C3.H} & \text{C1.H} - \text{C3.H} \\ \text{C1.L} - \text{C2.L} & \text{C2.L} - \text{C3.L} & \text{C1.L} - \text{C3.L} \\ \text{C1.C} - \text{C2.C} & \text{C2.C} - \text{C3.C} & \text{C1.C} - \text{C3.C} \end{pmatrix} \end{array}$$

Tabulka 4.1: Rozdílová matice pro nalezený vzor

Tento postup jsem provedl s každým nalezeným vzorem **A**, čímž jsem získal N matic s těmito rozdíly. Následně jsem si z těchto N matic vytvořil 1 matici aritmetických průměrů, která představuje ideální vzor. Znázorněno v tab. 4.2. Zkratka „avg“ značí aritmetický průměr.

$$\begin{array}{c} \\ Open \\ High \\ Low \\ Close \end{array} \begin{array}{c} \begin{array}{ccc} C_1 - C_2 & C_2 - C_3 & C_1 - C_3 \end{array} \\ \begin{pmatrix} \text{avg(C1.O} - \text{C2.O)} & \text{avg(C2.O} - \text{C3.O)} & \text{avg(C1.O} - \text{C3.O)} \\ \text{avg(C1.H} - \text{C2.H)} & \text{avg(C2.H} - \text{C3.H)} & \text{avg(C1.H} - \text{C3.H)} \\ \text{avg(C1.L} - \text{C2.L)} & \text{avg(C2.L} - \text{C3.L)} & \text{avg(C1.L} - \text{C3.L)} \\ \text{avg(C1.C} - \text{C2.C)} & \text{avg(C2.C} - \text{C3.C)} & \text{avg(C1.C} - \text{C3.C)} \end{pmatrix} \end{array}$$

Tabulka 4.2: Rozdílová matice aritmetických průměrů pro nalezený vzor

Analogicky jsem provedl pro metodu fuzzy, čili jsem na trénovací datové sadě EURUSD získal 2 matice aritmetických průměrů. V dalším kroku jsem vždy po třech procházel svíce z testovací sady USDCHF a počítal pro ně matice z tab. 4.1. Tyto matice jsem porovnával vždy s 1 maticí aritmetických průměrů z trénovací sady a to následovně:

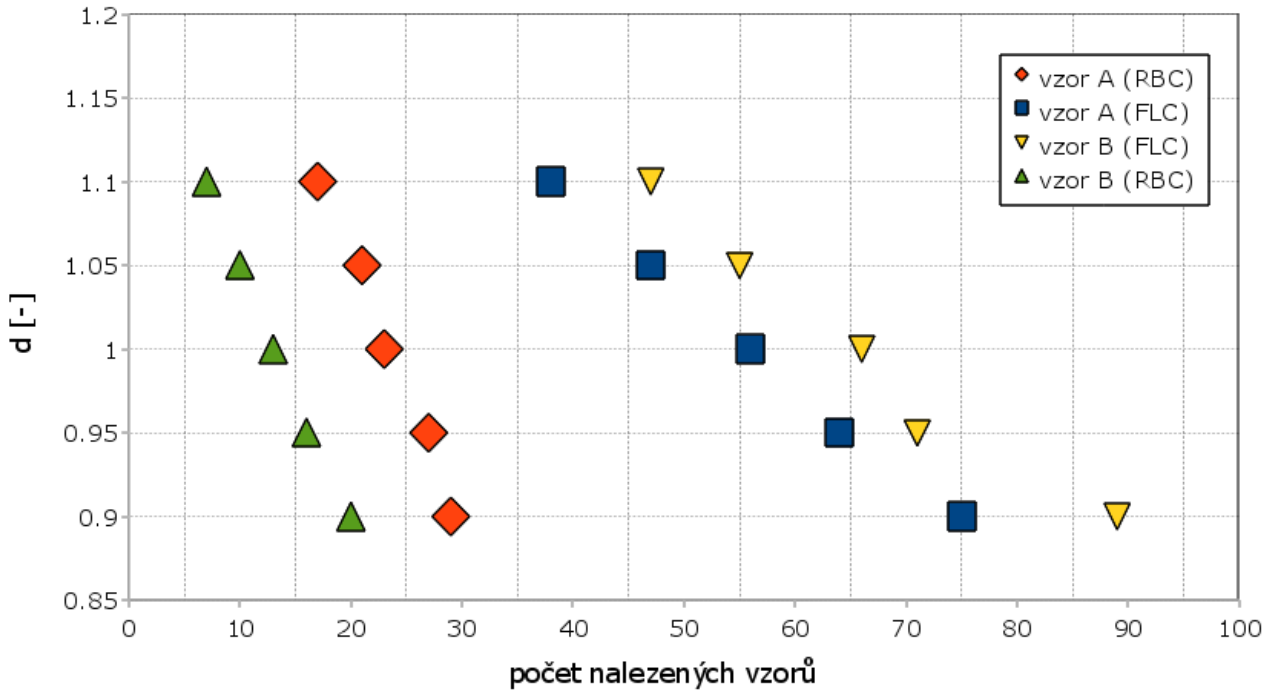
Pokud hodnota $p_3 = \frac{x}{y}$ byla větší nebo rovna zvolené hodnotě d pro všech 12 prvků matice, vyhodnotil jsem vzor jako korektní. Parametr x představuje dílčí prvek matice z tab. 4.1, parametr y představuje odpovídající aritmetický průměr z tab. 4.2. Jako hodnota d byly zvoleny postupně hodnoty d = 0,9; 0,95; 1; 1,05; 1,1. Jedná se de facto o povolenou odchylku od ideálního vzoru v rozmezí 0, ± 5, ± 10 %.

Metodu, kdy byly použity matice aritmetických průměrů, které byly získány z dat nalezených metodou rule-based, označuji jako „RBC“. Analogicky zvolím „FLC“ pro metodu fuzzy. Zde již pomalu přecházím k výsledkům detekce vzorů. Počet nalezených vzorů v závislosti na hodnotě d je možné vidět na obr. 4.5. Tyto údaje vycházejí z tab. 4.3.

d	0,90	0,90	0,95	0,95	1,00	1,00	1,05	1,05	1,10	1,10
metoda	FLC	RBC	FLC	RBC	FLC	RBC	FLC	RBC	FLC	RBC
vzor A	75	29	64	27	56	23	47	21	38	17
vzor B	89	20	71	16	66	13	55	10	47	7

Tabulka 4.3: Četnosti vzorů nalezených v testovací datově sadě metodami RBC a FLC

Je zřejmé, že pro rostoucí hodnotu d obecně klesá počet nalezených vzorů. Což je správně a odpovídá to realitě. Zvyšováním hodnoty d totiž dochází k navyšování minimální nutné hodnoty



Obrázek 4.5: Počet vzorů nalezených v závislosti na parametru d

rozdílů z tab. 4.1 pro to, aby byl vzor uznán jako korektní. V praxi to znamená, že pro to, aby byl vzor uznán korektním za takovýchto vysokých hodnot parametru d, musely by být vertikální vzdálenosti mezi parametry svící větší, což by bylo možné jen v případě značných výkyvů finanční řady. Analogicky realitě odpovídá i rostoucí počet nalezených vzorů pro snižující se hodnotu d. Jedná se totiž o snižování minimální nutné hodnoty rozdílů z tab. 4.1 pro to, aby byl vzor uznán jako korektní. Jsou tedy tolerovány i menší vertikální vzdálenosti mezi dílčími parametry svící.

4.4 Aplikace na datech

4.4.1 Implementace

Pro implementaci jsem si vybral jazyk Java, pro snadnou vizuální kontrolu a prohlížení dat jsem využil knihovnu JFreeChart[1], pro import dat a export nalezených vzorů jsem využil knihovnu jCSV[2]. Níže uvádím seznam tříd s jejich popisem. Kompletní zdrojový kód s komentáři včetně použitých dat je k dispozici na přiloženém CD jako i v elektronické verzi příloh.

- CandleStick.java – implementuje dílčí svíci včetně obou metod detekce (fuzzy i rule-based)
- CandleStickEntryParser.java – formátuje data pro čtení ze souboru CSV
- CandleStickEntryConverter.java – formátuje data pro zápis do souboru CSV
- CandleSticksSelectionDemo.java – pro zvolený CSV soubor otevírá GUI, viz též obr. 3.3
- Main.java – obstarává detekční procesy, klasifikační metodu, načítání a ukládání dat

Než rozvedu další sekci, je vhodné zmínit též relativně vysokou výpočetní náročnost při specifických užitích knihovny JFreeChart. Pro potřeby této práce se jednalo spíše o minoritní

[1] http://www.jfree.org/jfreechart/api.html
[2] https://code.google.com/p/jcsv/wiki/Welcome

problém, nicméně při užití větších datových sad je vhodné to brát v úvahu. Je možné, že by problém byl vyřešen experimentováním s nastavením konkrétních parametrů, viz též klíčová slova: JFreeChart performance.

4.4.2 Popis a ukázky kódu

Zde uvedu vybrané ukázky kódu, jako i princip vybraných metod.

Třída `Candlestick.java` implementuje mimo ostatní kód též dříve odkazované metody `assignRBSFuzzyType`, `assignSLFuzzyType`, `isRBSFuzzyXL` a `isSLFuzzyS`.

Každý objekt třídy `Candlestick.java` má dále deklarována 2 jednorozměrná pole typu `double[]`. Jedná se o pole `realBodySizeType` a `shadowType`. Ta uchovávají data užívaná při klasifikaci metodou fuzzy, konkrétně uchovávají stupeň příslušnosti μ_x k dané fuzzy množině.

index	0	1	2	3	4
velikostní třída	XS	S	M	L	XL

Tabulka 4.4: Pole realBodySizeType reprezentující velikostní třídy reálného těla svíce

index	0	1	2
velikostní třída	S	M	L

Tabulka 4.5: Pole shadowType reprezentující velikostní třídy stínu

Níže uvádím část kódu metody `assignRBSFuzzyType`, ve které je vidět, jak je definována velikostní třída reálného těla svíce `XL`. Numerické hodnoty jako i zápis vycházejí z fuzzy množin popsaných v sekci 4.3.3.

```
public void assignRBSFuzzyType()
{
        double x = this.realBodySize / this.realBodySizeMA;
        double a = 0;
        double b = 0.1;
        double c = 0.65;
        double d = 1.35;
        double e = 1.55;
...
// XL
if (x >= e)
{
        this.realBodySizeType[4] = 1;
}
if (e > x && x > d)
{
        this.realBodySizeType[4] = (x-d)/(e-d);
}
if (x <= d)
{
        this.realBodySizeType[4] = 0;
}
...
```

Parametr `realBodySize` představuje velikost reálného těla svíce, parametr `realBodySizeMA` představuje klouzavý průměr velikosti reálného těla svíce.

Analogicky funguje také metoda `assignSLFuzzyType`. Níže uvádím část jejího kódu včetně definice velikostní třídy `S`.

```
public void assignSLFuzzyType()
{
        double x = (this.upShadow + this.lowShadow) / this.bodySize;
        double a = 0.3;
        double b = 0.6;
        double c = 0.9;

// S
if (x <= a)
{
        this.shadowType[0] = 1;
}
if (a < x && x < b)
{
        this.shadowType[0] = (b-x)/(b-a);
}
if (x >= b)
{
        this.shadowType[0] = 0;
}
...
```

Každé svíci poté přiřadím v hlavní třídě `Main.java` odpovídající velikostní třídy následovně.

```
Reader csvFile = new InputStreamReader...
CSVReader<CandleStick> reader = new CSVReaderBuilder...
List <CandleStick> trainSetList = new ArrayList<>();
trainSetList = reader.readAll();
...
for (int i = 0; i < trainSetList.size(); i++)
{
        trainSetList.get(i).assignRBSFuzzyType();
        trainSetList.get(i).assignSLFuzzyType();
}
...
```

Je zřejmé, že si načtu do `trainSetList` data z CSV souboru. Jak jsem uváděl dříve v sekci 3.5.1, tato data reprezentují dílčí svíce.

Nakonec se dostávám k metodě `isRBSFuzzyXL`, kde jednoduše porovnám dříve přiřazené hodnoty μ_x v dílčích prvcích pole `realBodySizeType` a vyberu z nich tu největší. Je zřejmé, že též povoluji rovnost, čili v případě rovnosti μ_x mezi množinami na stejném intervalu zjevně dostávám 2 různé typy velikosti reálného těla svíce.

```
public boolean isRBSFuzzyXL ()
{
  if (this.realBodySizeType[4] >= this.realBodySizeType[1] &&
      this.realBodySizeType[4] >= this.realBodySizeType[2] &&
      this.realBodySizeType[4] >= this.realBodySizeType[3] &&
      this.realBodySizeType[4] >= this.realBodySizeType[0])
  {
     return true;
  }
  else
  {
     return false;
  }
}
```

Analogicky též pro metodu `isSLFuzzyS`, jak uvádím následovně.

```
public boolean isSLFuzzyS ()
{
  if (this.shadowType[0] >= this.shadowType[1] &&
      this.shadowType[0] >= this.shadowType[2])
  {
     return true;
  }
  else
  {
     return false;
  }
}
```

4.5 Výsledky detekce a srovnání metod

4.5.1 Statistický aparát

Ke srovnání metod využiji aparát „positive predictive value“ neboli PPV. Jedná se o pravděpodobnost, s jakou je určený vzor opravdu korektním vzorem. Hodnota PPV je definována jako

$$PPV = \frac{TP}{TP + FP}, \tag{4.2}$$

kde TP je počet prvků množiny „true positive“ neboli vzorů, které byly označeny jako korektní a jsou korektní, FP je počet prvků množiny „false positive“ neboli vzorů, které byly označeny jako korektní, avšak korektní nejsou. Konkrétní hodnoty TP a FP získám z testovacích dat. A to na základě srovnání korektně určených dat a těch, která jsem nalezl metodami RBC a FLC.

K tomu však budu potřebovat korektní vzory z testovací datové množiny páru USDCHF. Jak jsem již zmiňoval, korektní hodnoty de facto neexistují. Zvolím tedy jako testovací množinu taková data, která budou nalezena klasickými metodami rule-based a fuzzy. Poté se pokusím porovnat, zdali je signifikantní rozdíl mezi hodnotami PPV, pokud použiji RBC klasifikátor na data, která byla jako korektní určena fuzzy metodou a mezi hodnotami PPV, pokud použiji klasifikátor FLC na data, která byla jako korektní určena rule-based metodou.

Domnívám se, že párový test použít nelze s ohledem na to, že dílčí výběry (respektive identifikované vzory) nejsou vždy totožné. Předpokládám tedy nezávislé výběry. Využiji oboustranného dvouvýběrového Studentova t-testu, kdy předpokládám stejné rozptyly, což nejdříve ověřím Fisherovým F-testem. Pro konstrukci statistického aparátu vycházím z [72].

Pro oboustranný F-test testuji nulovou hypotézu o rovnosti rozptylů:

$$H_0 : \sigma_1^2 = \sigma_2^2 \tag{4.3}$$

naproti ní existuje alternativní hypotéza

$$H_1 : \sigma_1^2 \neq \sigma_2^2 \tag{4.4}$$

a dále nechť

$$F = \frac{s_1^2}{s_2^2}, \tag{4.5}$$

kde s_1^2, s_2^2 představují výběrové rozptyly souborů. Pokud bude platit, že

$$F < F_{\alpha/2\ (\nu_1,\nu_2)} \ \vee \ F > F_{1-\alpha/2\ (\nu_1,\nu_2)}, \tag{4.6}$$

kde hladina spolehlivosti $\alpha = 0,05$ a dále $\nu_1 = n_1 - 1, \nu_2 = n_2 - 1$ představují stupně volnosti pro počty prvků souborů, pak H_0 zamítám. Uvedené výběrové rozptyly souborů spočtu jako

$$s^2 = \frac{\sum_{n=1}^{n}(x_i - \bar{x})^2}{n-1}, \tag{4.7}$$

kde n je počet prvků x_i souboru a dále $\bar{x}$ je aritmetický průměr tohoto souboru.

Za předpokladu, že H_0 nezamítám, přecházím k oboustrannému t-testu a formuluji hypotézu o ekvivalentní efektivitě PPV pro zmíněné dva přístupy.

$$H_0 : \mu_1 = \mu_2, \tag{4.8}$$

naproti tomu

$$H_1 : \mu_1 \neq \mu_2. \tag{4.9}$$

Testovací t-statistiku volím následovně:

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}, \tag{4.10}$$

kde dílčí parametry odpovídají dříve uvedeným hodnotám. Pokud následně platí

$$t \notin (-t_{1-\frac{\alpha}{2}(n_1+n_2-2)}; t_{1-\frac{\alpha}{2}(n_1+n_2-2)}), \tag{4.11}$$

kde opět $\alpha = 0,05$, pak H_0 zamítám. Na konec statistického aparátu uvedu, že dílčí detailní výpočty a hodnoty, které budou následovat, uvádím v příloze D pro možnost kontroly.

4.5.2 Data nalezená v trénovací množině

V trénovací datové sadě EURUSD jsem navrženými metodami fuzzy a rule-based nalezl vzory, jejichž četnosti jsou zaznamenány v následující tabulce 4.6.

vzor	metoda fuzzy	metoda rule-based
A	216	89
B	247	78

Tabulka 4.6: Četnosti vzorů nalezených v trénovací datově sadě

Na základě těchto nalezených vzorů jsem vytvořil matice reprezentující průměrný vzor dle zápisu v tabulce 4.2. Tyto matice též uvádím v příloze. Dále jsem vyhledával vzory již na množině testovací.

4.5.3 Určování korektních vzorů

Implementaci automatického srovnávání korektních vzorů pro dílčí porovnávání se mi bohužel nepodařilo sestavit, nicméně byl jsem schopen manuálně projít nalezená data a určit průnik množiny vzorů, které byly zvolené jako korektní, s těmi, které byly nalezeny.

Jedná se o použití programů `awk` a `grep` na cílové CSV soubory, do kterých ukládám nalezená data z programu, který je psán v Javě. Tato výstupní data mají následující tvar, který je lehce odlišný od dříve uváděného.

```
2011.11.11 14:00;1.37527;1.37716;1.37526;1.37672;1787
```

Je tedy možné si vypsat jednoznačně identifikující hodnoty do separátních souborů, jak uvádím níže. Zde se jedná o jednoznačně identifikující časovou hodnotu '`2011.11.11 14:00`'.

```
awk -F";" '{print $1}' nalezene_vzory.csv > file1
awk -F";" '{print $1}' soubor_korektni-ch_vzoru.csv > file2
```

A tyto soubory je poté již možné porovnávat.
`grep -Fx -f file1 file2 | wc -l`

Tímto způsobem lze získat minimálně počet prvků v průniku množin.

4.5.4 Korektní data nalezená metodou rule-based

V testovací sadě dat USDCHF jsem pomocí metody rule-based nalezl vzory, jejichž četnosti jsou zaznamenány v tab. 4.7. Tyto vzory jsou uvažovány pro další srovnávání jako korektní.

vzor	**metoda rule-based**
A	98
B	87

Tabulka 4.7: Četnosti korektních vzorů v testovací datově sadě

Dále jsem v testovací sadě pro tyto korektní vzory hledal odpovídající vzory na základě klasifikační metody RBC a FLC. Četnosti nalezených korektních vzorů, značeno „vzor X OK“, uvádím v tab. 4.8 společně se zvolenou hodnotou d.

d	0,90	0,90	0,95	0,95	1,00	1,00	1,05	1,05	1,10	1,10
metoda	FLC	RBC	FLC	RBC	FLC	RBC	FLC	RBC	FLC	RBC
vzor A	75	29	64	27	56	23	47	21	38	17
vzor A OK	36	17	32	16	30	14	29	13	25	11
vzor B	89	20	71	16	66	13	55	10	47	7
vzor B OK	31	10	25	9	23	7	19	7	17	6

Tabulka 4.8: Četnosti vzorů a korektních vzorů v testovací datově sadě

Nyní využiji dříve uvedeného statistického aparátu a stanovím hypotézu

$$H_0 : \mu_{PPV_{FLC}} = \mu_{PPV_{RBC}},$$

jinými slovy, že na zvolené testovací datové sadě nalezené metodou rule-based dosahuje stejné efektivity PPV jak metoda RBC, tak metoda FLC. Jako demonstraci pro pochopení výpočtu uvádím výpočet sumy dílčích hodnot PPV z tab. 4.8 pro metodu FLC a to následovně dle dříve uvedeného vzorce:

$$PPV_{FLC} = \frac{36}{75} + \frac{31}{89} + ... + \frac{25}{38} + \frac{17}{47}, \tag{4.12}$$

analogicky pro metodu RBC.

Prvním krokem je test shodnosti výběrových rozptylů pomocí F-testu.

$$F = \frac{s^2_{PPV_{FLC}}}{s^2_{PPV_{RBC}}} = \frac{0,01451}{0,0997} = 1,455. \tag{4.13}$$

Dále tedy

$$F < F_{\alpha/2\ (\nu_1,\nu_2)} \ \vee \ F > F_{1-\alpha/2\ (\nu_1,\nu_2)}, \tag{4.14}$$

$$1,455 < F_{0,025\ (9,9)} \ \vee \ 1,455 > F_{0,975\ (9,9)}, \tag{4.15}$$

$$1,455 < 0,25 \ \vee \ 1,455 > 4,03, \tag{4.16}$$

což zjevně neplatí. Hypotézu H_0 o rovnosti rozptylů tedy není možné zamítnout. Dá se předpokládat, že uvedené rozptyly jsou shodné.

Dále vyjdu z uváděného testovací kritéria t,

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{n_1 s^2_{PPV_{FLC}} + n_2 s^2_{PPV_{FLC}}}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}, \tag{4.17}$$

kde $\bar{x_1}$ a $\bar{x_2}$ udávají průměrné hodnoty PPV_{RBC}, respektive PPV_{FLC}. Hodnoty n_1, n_2 udávají velikost souboru, v našem případě $n_1 = n_2 = 10$.

Po dosazení konkrétních hodnot vychází

$$t = \frac{0,455 - 0,611}{\sqrt{10 \cdot 0,0145 + 10 \cdot 0,009}} \cdot \sqrt{\frac{10 \cdot 10(10 + 10 - 2)}{10 + 10}}, \tag{4.18}$$

$$t = -3,05. \tag{4.19}$$

S ohledem na symetrii t-rozdělení je možné uvažovat $t = 3,05$. Dále pokračuji v dosazování do intervalu Studentova t-rozdělení

$$(-t_{1-\frac{\alpha}{2}(n_1+n_2-2)}; t_{1-\frac{\alpha}{2}(n_1+n_2-2)}), \tag{4.20}$$

$$(-t_{0,975(18)}; t_{0,975(18)}), \tag{4.21}$$

$$(-2,101; 2,101) \tag{4.22}$$

a zjevně tedy

$$t \notin (-2,101; 2,101). \tag{4.23}$$

Na hladině $\alpha = 0,05$ mohu zamítnout hypotézu H_0 o rovnosti $\mu_{PPV_{FLC}} = \mu_{PPV_{RBC}}$.

4.5.5 Korektní data nalezená metodou fuzzy

V testovací sadě dat USDCHF jsem pomocí metody fuzzy nalezl vzory, jejichž četnosti jsou zaznamenány v tab. 4.9. Tyto vzory jsou uvažovány pro další srovnávání jako korektní.

vzor	**metoda fuzzy**
A	265
B	252

Tabulka 4.9: Četnosti korektních vzorů v testovací datově sadě

Dále jsem v testovací sadě pro tyto korektní vzory hledal odpovídající vzory na základě klasifikační metody RBC a FLC jako v předchozí sekci. Četnosti nalezených vzorů uvádím v tab. 4.10 společně se zvolenou hodnotou d a korektně určenými vzory, opět značeno „vzor X OK“.

t	0,90	0,90	0,95	0,95	1,00	1,00	1,05	1,05	1,10	1,10
metoda	FLC	RBC	FLC	RBC	FLC	RBC	FLC	RBC	FLC	RBC
vzor A	75	29	64	27	56	23	47	21	38	17
vzor A OK	36	17	31	16	30	14	28	13	24	11
vzor B	89	20	71	16	66	13	55	10	47	7
vzor B OK	34	10	28	9	26	7	22	7	20	6

Tabulka 4.10: Četnosti vzorů a korektních vzorů v testovací datově sadě

Při pohledu na tabulku 4.10jsem se domníval, že jsem při výpočtech nebo v kódu někde udělal chybu. Tabulka z předchozí sekce, kde jsem uvažoval jako korektní vzory ty, které byly nalezeny rule-based metodou, se totiž v podstatě vůbec až na malé odchylky nezměnila. Je zřejmé, že s uvedenými numerickými hodnotami nemá smysl provádět stejnou statistickou analýzu, neboť by evidentně dosahovala totožných výsledků.

4.5.6 Shrnutí

Po opakované kontrole, že se nestala chyba, jsem přišel na důvod, proč se tak stalo. Spatřuji jej ve zobecnění ze sekce 3.5.1, kdy jsem uvažoval běžně udávaný korelační koeficient $c \to -1$ mezi páry. Objasní to též pouhé rozdíly počtů nalezených vzorů, které uvádím znovu pro přehlednost v následujících tabulkách 4.11 a 4.12.

vzor	**fuzzy**	**rule-based**
A	216	89
B	247	78

Tabulka 4.11: Četnosti nalezených vzorů v datově sadě EURUSD

vzor	**fuzzy**	**rule-based**
A	265	98
B	252	87

Tabulka 4.12: Četnosti nalezených vzorů v datově sadě USDCHF

Je zřejmé, že pro mnou navržené modely je na měnovém páru USDCHF nacházeno obecně větší množství vzorů než v případě páru EURUSD. Pár je tedy vertikálně „posunutý“, čímž

znemožňuje dosahování významně odlišných výsledků v případě, kdy uvažuji modifikovanou klasifikační metodu aritmetických průměrů ve srovnání s metodami klasickými (rule-based a fuzzy) jako korektními vzory. Modifikovaná klasifikační metoda je totiž „naučená“ na pár EU-RUSD. Je tedy svým způsobem možné mluvit o přeučení.

Je nicméně zřejmé, že metoda RBC dosahuje obecně lepších výsledků, neboť jak je patrné z výpočtů v příloze, medián hodnoty PPV pro metodu FLC je 42 %, medián hodnoty PPV pro metodu RBC je 60 %, avšak při téměř $3,5\times$ menším počtu obecně nalezených vzorů metodou RBC – tedy jak korektních tak nekorektních.

Na základě výše uvedených skutečností je tedy možné vynést minimálně ten závěr, že mnou navržená modifikovaná klasifikační metoda RBC je v daném kontextu obecně lepší než metoda FLC a je s relativně dobrou přesností použitelná i přes nedostatek zmíněného korelačního koeficientu.

5. Závěr

5.1 Zhodnocení cílů

V této bakalářské práci byl proveden teoeretický popis, návrh a implementace metod za účelem vyhledávání vzorů v neurčitých, špatně predikovatelných časových řadách. Tyto řady byly reprezentovány finančními časovými řadami z trhu Forex.

Metody byly navrženy postupy rule-based, fuzzy a modifikovanou klasifikační metodou, která vychází z podobností s ohledem na ideální vzor. Dle očekávání se návrh a implementace metod setkaly s úspěchem a definované vzory je uvedenými metodami možné jak úspěšně nacházet, tak poté zobrazovat a procházet, což se dá považovat za hlavní přínos této práce.

Ačkoliv programová implementace je považována spíše za nástroj než samotný cíl práce, je možné z ní dále vycházet. Jak již bylo uvedeno, problémem při zobrazování většího množství dat může být výpočetní náročnost zapřičiněná užitím knihovny JFreeChart.

V závěrečném srovnání metod je zřejmé, že se ukázal jako chybný výchozí přepoklad korelace měnových párů. Uvedený postup na druhou stranu přinesl zajímavý poznatek, že navržený klasifikátor RBC je i přesto možné použít s akceptovatelnými výsledky.

Také se ukazuje, že pro striktně definované vzory, které mají nejblíže teorii, poskytuje nejpřesnější výsledky klasifikátor založený na metodě rule-based. Pokud by byl uvažován přístup s ohledem na praktičtější použití, zřejmě by se ukázal vhodnější volbou fuzzy přístup, neboť obchodníky v praxi nezajímá ani tak exaktní podoba s teoretickým vzorem, jako spíše obecný vzestup či pokles kursu.

I přes překvapivý výsledek při srovnávání metod je s ohledem na dosažené výsledky možné říct, že tato bakalářská práce byla dokončena dle očekávání a její cíle byly splněny.

5.2 Návrh rozšíření a zlepšení

Jedním z kroků, jak se dá práce zlepšit za současného stavu, je uvažovat pouze jeden měnový pár. Problémem však bude menší množství nalezených vzorů, dle toho bude nutné upravit modelace vzorů a svící.

Dále je vhodné užít více metrik než je uvedená klasifikační metrika založená na odchylce od aritmetického průměru vzorů, která se neukazuje zcela ideální. Je možné zavést například vážené průměry, euklidovské vzdáleností a jiné. Je také možné zavést více vzorů než jen demonstrační. Pak ovšem je nutný jejich návrh.

Poněkud odlišné výsledky by též poskytovalo užití jiného druhu klouzavého průměru, například exponenciálního. Ten totiž zohledňuje daleko více členy těsně předcházející členu, pro který je klouzavý průměr určován. Oproti němu klasický klouzavý průměr uvažuje celistvý interval.

V souvislosti s tím by bylo vhodné zpracovat implementaci automatického srovnávání korektních vzorů, jelikož s rostoucím počtem jak klasifikátorů, tak záznamů není metoda zpracování dat externími skripty a příkazy optimální.

Bylo by též vhodné otestovat i jiná podkladová data než ta z trhu Forex či finančního prostředí obecně. Pokud by existovala oblast s exaktně definovanými vzory jako i takovými časovými řadami, kde by vzory měly přesnější podobu, srovnání metod by mělo daleko větší praktický přínos.

Seznam použité literatury

[1] Dolák, Ondřej. Big data, Nové způsoby zpracování a analýzy velkých objemů dat. In: *Systemonline.cz* [online]. [cit. 2015-04-15]. Dostupné z: `http://www.systemonline.cz/clanky/big-data.htm`.

[2] Dean, Jeffrey – Ghemawat, Sanjay. MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM - 50th anniversary issue: 1958 - 2008* [online]. Volume 51 Issue 1, s. 107-113. [cit. 2015-04-19]. Dostupné z doi: 10.1145/1327452.1327492.

[3] *Computing* [online]. Conseil Européen pour la recherche nucléaire. [cit. 2015-04-21]. Dostupné z: `http://home.web.cern.ch/about/computing`.

[4] *Nový volitelný magisterský předmět Big Data Technologies* [online]. České vysoké učení technické v Praze. [cit. 2015-04-24]. Dostupné z: `http://informatika.fel.cvut.cz/node/721`.

[5] Mayer-Schönberger, Viktor – Cukier, Kenneth. *Big Data - Revoluce, která změní způsob, jak žijeme, pracujeme a myslíme*. Computer Press, a.s., 2014. ISBN: 9788025141199.

[6] Columbus, Louis. Where Big Data Jobs Will Be In 2015. In: *Forbes.com* [online]. 2014-12-29 [cit. 2015-04-15]. Dostupné z: `http://www.forbes.com/sites/louiscolumbus/2014/12/29/where-big-data-jobs-will-be-in-2015/`.

[7] McNulty, Eileen. 10 Online Big Data Courses. In: *Dataconomy.com* [online]. 2014-09-25 [cit.2015-04-21]. Dostupné z: `http://dataconomy.com/10-online-big-data-courses/`.

[8] *About Google Flu Trends* [online]. Google Inc. [cit. 2015-04-17]. Dostupné z: `https://www.google.org/flutrends/about/faq.html`.

[9] Salzberg, Steven. Why Google Flu Is A Failure. In: *Forbes.com* [online]. 2014-03-24 [cit. 2015-04-17]. Dostupné z: `http://www.forbes.com/sites/stevensalzberg/2014/03/23/why-google-flu-is-a-failure/`.

[10] Han, Jiawei – Kamber, Micheline – Pei, Jian. *Data Mining - Concepts and Techniques*. 3rd ed. Morgan Kaufmann Publishers, July 2011. ISBN: 978-0123814791.

[11] Witten, Ian – Frank, Eibe. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. Morgan Kaufmann Publishers, 2005. ISBN: 0120884070.

[12] Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006. ISBN: 978-0-387-31073-2.

[13] Esling, Philippe – Agon, Carlos. Time-series data mining. *ACM Computing Surveys (CSUR)* [online]. Volume 45, Issue 1, November 2012, Article No. 12. [cit. 2015-04-20]. Dostupné z doi: 10.1145/2379776.2379788.

[14] Alles, Irina. *Time Series Clustering in the Field of Agronomy*. Darmstadt: Technische Universität, 2013. Master-Thesis, Technische Universität Darmstadt, Department of Computer Science.

[15] Donghua, Pan – Chonghui, Guo – Hailin, Li. An improved piecewise aggregate approximation based on statistical features for time series mining. In: *Proceedings of the 4th international conference on Knowledge science, engineering and management*. Springer-Verlag Berlin, Heidelberg 2010. Pages 234-244. ISBN:978-3-642-15280-1 (Online).

[16] Chakrabarti, Kaushik – Keogh, Eamonn – Mehrotra, Sharad – Pazzani, Michael. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM Transactions on Database Systems (TODS)* [online]. Volume 27, Issue 2, June 2002. Pages 188-228 [cit. 2015-04-22]. Dostupné z doi: 10.1145/568518.568520.

[17] *Welcome to the SAX (Symbolic Aggregate approXimation) Homepage.* [online]. University of California, San Diego. [cit. 2015-04-22]. Dostupné z: `http://www.cs.ucr.edu/~eamonn/SAX.htm`.

[18] Liu, Wenjia – Chen, Bo – Swartz, R. Andrew. Investigation of Time Series Representations and Similarity Measures for Structural Damage Pattern Recognition. In: *The Scientific World Journal*. Vol. 2013, Article No. 248349, 13 pages. Dostupné z doi: 10.1155/2013/248349.

[19] Chapelle, Olivier – Schölkopf, Bernhard – Zien, Alexander. *Semi-supervised learning*. MIT press Cambridge, 2006. ISBN: 9780262033589.

[20] Kidiyo, Kpalma – Ronsin, Joseph. An Overview of Advances of Pattern Recognition Systems in Computer Vision. In: Goro Obinata and Ashish Dutta. *Vision Systems: Segmentation and Pattern Recognition* [online]. I-Tech Education and Publishing, 2007. ISBN: 978-3-902613-05-9. [cit. 2015-04-20]. Dostupné z doi: 10.1145/2379776.2379788.

[21] Liu, Jie – Sun, Jigui – Wang, Shengsheng. Pattern recognition: An Overview. In: *International Journal of Computer Science and Network Security*. Vol. 6, pp. 57-61, 2006. [cit. 2015-04-22]. Dostupné z: `http://paper.ijcsns.org/07_book/200606/200606A10.pdf`.

[22] Rao, Subba M. – Reddy, Eswara B. Comparative analysis of pattern recognition methods: An overview. In: *Indian Journal of Computer Science and Engineering (IJCSE)*. Vol. 2, Number 3, Pages 385-390, 2011. [cit. 2015-04-21]. Dostupné z: `www.ijcse.com/docs/IJCSE11-02-03-103.pdf`.

[23] Janošek, Michal – Kocian, Václav – Kotyrba, Martin – Volná, Eva. *Umělá inteligence - Rozpoznávání vzorů v dynamických datech*. BEN - technická literatura, 2014. ISBN: 978-80-7300-497-2.

[24] Santosh Kumar, Das – Abhishek, Kumar – Bappaditya, Das – Burnwal, A.P. On Soft Computing Techniques In Various Areas. In: *ACER 2013, CS & IT-CSCP*. Vol. 3, pp. 59–68, 2013. [cit. 2015-04-28]. Dostupné z doi: 10.5121/csit.2013.3206.

[25] Xi, Xiaopeng – Keogh, Eamonns – Shelton, Christian – Wei, Li – Ratanamahatana, Chotirat Ann. Fast Time Series Classification using Numerosity Reduction. In: *Proceedings of the Twenty-Third International Conference on Machine Learning*. 2006. pp. 1033-1040. Dostupné z: `http://www.cs.ucr.edu/~cshelton/papers/index.cgi?XiKeoSheWeiCho06`.

[26] Lin, Jessica – Williamson, Sheri – Borne, Kirk – DeBarr, David. Pattern Recognition in Time Series. In: Michael Way, Jeff Scargle, Kamal Ali, Ashok Srivastava. *Advances in Machine Learning and Data Mining for Astronomy* . Chapman and Hall an imprint of CRC Press (a division of Taylor and Francis), 2012. ISBN: 978-1439841730. [cit. 2015-04-20]. Dostupné z `http://cs.gmu.edu/~jessica/publications/astronomy11.pdf`.

[27] Ding, Hui – Trajcevski, Goce – Scheuermann, Peter – Wang, Xiaoyue – Keogh, Eamonn. Querying and mining of time series data: experimental comparison of representations and distance measures. In: *Proceedings of the VLDB Endowment.*Volume 1 Issue 2, August 2008. Pages 1542-1552. [cit. 2015-04-25]. Dostupné z: `http://www.cs.ucr.edu/~eamonn/vldb_08_Experimental_comparison_time_series.pdf`.

[28] Gülin, Dede – Hüsnü Sazli, Murat. Speech recognition with artificial neural networks. In: *Digital Signal Processing.* Volume 20, Issue 3, May, 2010, Pages 763-768. [cit. 2015-04-21]. Dostupné z doi: 10.1016/j.dsp.2009.10.004.

[29] Bishop, Christopher M. *Neural Networks for Pattern Recognition.* Oxford University Press, Inc. New York, NY, USA ©1995. ISBN: 0198538642.

[30] *G5AIAI - Introduction to Artificial Intelligence* [online]. The University of Nottingham. [cit. 2015-04-28]. Dostupné z: `http://www.cs.nott.ac.uk/~gxk/courses/g5aiai/006neuralnetworks/neural-networks.htm`.

[31] *Neural Networks - History* [online]. Stanford University. [cit. 2015-04-28]. Dostupné z: `http://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/History/history1.html`.

[32] Lin, Jessica – Keogh, Eamonn – Wei, Li – Lonardi, Stefano. Experiencing SAX: a novel symbolic representation of time series. In: *Data Mining and Knowledge Discovery.* Volume 15, Issue 2, October 2007, Pages 107 - 144. [cit. 2015-04-26]. Dostupné z doi: 10.1007/s10618-007-0064-z.

[33] Donghua, Pan. Pattern Extraction for Time Series Classification. In: *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery.* Springer-Verlag London, UK ©2001. Pages 115-127. ISBN:3-540-42534-9.

[34] Wei-Yin, Loh. Classification and regression trees. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.* Vol.1, 2011. Pages 14-23. Dostupné z doi: 10.1002/widm.8.

[35] Hoddinott, John – Yisehac, Yohannes. Classification and regression trees. In: *Classification and regression trees: An introduction*[online]. International Food Policy Research Institute, 1999. Dostupné z `http://www.ifpri.org/sites/default/files/publications/tg03.pdf`.

[36] Jain, Anil K. – Dubes, Richard C. *Algorithms for clustering data* . Prentice-Hall, Inc. Upper Saddle River, NJ, USA ©1988. ISBN:0-13-022278-X.

[37] Zolhavarieh, Seyedjamal – Aghabozorgi, Saeed – Wah Teh, Ying – Lonardi, Stefano. A Review of Subsequence Time Series Clustering. In: *The Scientific World Journal.* Volume 2014 (2014), Article ID 312521. [cit. 2015-04-29]. Dostupné z doi: 10.1155/2014/312521.

[38] Fu, Tak-chung – Chungozorgi, Fu-lai – Ng, Vincent – Luk, Robert. Pattern discovery from stock time series using self-organizing maps. In: *Workshop Notes of KDD2001 Workshop on Temporal Data Mining.* 2001. [cit. 2015-04-29]. Dostupné z: `http://www.researchgate.net/profile/Vincent_Ng10/publication/228771755_Pattern_discovery_from_stock_time_series_using_self-organizing_maps/links/0f31753218a5bd2457000000.pdf`.

[39] Spiegel, Stephan – Gaebler, Julia – Lommatzsch, Andreas – De Luca, Ernesto – Albayrak, Sahin. Fast Time Series Classification using Numerosity Reduction. In: *Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data*. ACM New York, NY, USA ©2011. Pages 34-42. ISBN: 978-1-4503-0832-8. Dostupné z doi: 10.1145/2003653.2003657.

[40] Keogh, Eamonn – Lin, Jessica – Truppel, Wagner. Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Research. In: *Proceedings of the Third IEEE International Conference on Data Mining*. IEEE Computer Society Washington, DC, USA ©2003. Page 115. ISBN:0-7695-1978-4.

[41] Nayak, Maya – Behera, Lalit Kumar. Fuzzy Neural Inference System for Pattern Recognition of Power Quality Events Using Rule Generation. In: *International Journal of Computer Science and Information Technologies*. VOLUME 4, ISSUE 1, January- February 2013. Pages : 121-125. [cit. 2015-05-02]. Dostupné z: `http://www.ijcsit.com/docs/Volume%204/Vol4Issue1/ijcsit2013040128.pdf`.

[42] Min, Ji – Fuding, Xie – Yu, Ping. A Dynamic Fuzzy Cluster Algorithm for Time Series. In: *Abstract and Applied Analysis*. vol. 2013, Article ID 183410, 7 pages, 2013. [cit. 2015-04-27]. Dostupné z doi: 10.1155/2013/183410.

[43] Kovalerchuk, Boris – Vityaev, Evgenii. *Data mining in finance: advances in relational and hybrid methods*. Kluwer Academic Publishers Norwell, MA, USA ©2000. ISBN:0-7923-7804-0.

[44] Song, Qiang – Chissom, Brad S. – Yu, Ping. Fuzzy time series and its models. In: *Fuzzy Sets and Systems*. Volume 54, Issue 3, March 25, 1993. Pages 269 - 277. [cit. 2015-04-18]. Dostupné z doi: 10.1016/0165-0114(93)90372-O.

[45] Lee, Chiung-hon Leon – Liu, Alan – Chen, Wen-sung. Pattern Discovery of Fuzzy Time Series for Financial Prediction. In: *IEEE Transactions on Knowledge and Data Engineering*. Volume:18, Issue: 5. 2006. Page(s): 613 - 625. [cit. 2015-04-27]. Dostupné z doi: 10.1109/TKDE.2006.80.

[46] Lee, Chiung-hon Leon – Liu, Alan – Chen, Wen-sung. Pattern Discovery of Fuzzy Time Series for Financial Prediction. In: *IEEE Transactions on Knowledge and Data Engineering*. Volume:18, Issue: 5. 2006. Page(s): 613 - 625. [cit. 2015-04-27]. Dostupné z doi: 10.1109/TKDE.2006.80.

[47] Lilly, John H. *Fuzzy Control and Identification*. John Wiley & Sons, Hoboken, NJ, USA, 2010. ISBN: 978-0-470-54277-4.

[48] Navara, Mirko – Olšák, Petr. Základy fuzzy množin. In: *Katedra matematiky*. [online] České vysoké učení technické v Praze. 2001. [cit. 2015-04-27]. Dostupné z: `ftp://math.feld.cvut.cz/olsak/fuzzy/fuzzy.pdf`

[49] Kidiyo, Kpalma – Ronsin, Joseph. Advanced Fuzzy Logic Technologies in Industrial Applications. In: Michael J. Grimble and Michael A. Johnson. *Advances in Industrial Control*. Springer ©2006. ISSN: 1430-9491.

[50] Markets.com. *Vše, co byste měli vědět o technické analýze* [online]. [cit. 2015-02-07]. Dostupné z: `http://www.markets.com/cz/education/technical-analysis/`.

[51] Markets.com. *Co je technická analýza?* [online]. [cit. 2015-02-07]. Dostupné z: http://www.markets.com/cz/education/technical-analysis/what-is-technical-analysis.html.

[52] Hartman, Ondřej – Turek, Ludvík. *První kroky na FOREXu*. Computer Press, a.s., 2009. ISBN: 978-80-251-2006-4.

[53] Hartle, Tom. Steve Nison On Candlestick Charting. *Technical Analysis of Stocks & Commodities*. Roč. 9, č. 3, s. 105–108. Technical Analysis, Inc.© 1991.

[54] Lambert, Clive. *Candlestick Charts*. Harriman House, 2009. ISBN: 9781905641741.

[55] Bulkowski, Thomas N. *Encyclopedia of Candlestick Charts*. John Wiley & Sons, Inc., 2008. ISBN: 978-0-470-18201-7.

[56] Blaise, Jean-Yves – Dudek, Iwona. Can simplicity help?. In: *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business* . Article No. 17. New York, NY, USA ©2014. ISBN: 978-1-4503-2769-5. Dostupné z doi: 10.1145/2637748.2638414.

[57] Patro, Ashish – Govindan, Srinivas – Banerjee, Suman. Observing home wireless experience through WiFi APs. In: *Proceedings of the 19th annual international conference on Mobile computing & networking* . Pages 339-350 . New York, NY, USA ©2013. ISBN: 978-1-4503-1999-7. Dostupné z doi: 10.1145/2500423.2500448.

[58] Morris, Gregory L. *Candlestick Charting Explained*. McGraw-Hill, 1995. ISBN: 978-0071461542.

[59] *Time Series Methods* [online]. ©2014 Pearson Education. [cit. 2015-05-01]. Dostupné z: http://www.prenhall.com/divisions/bp/app/russellcd/PROTECT/CHAPTERS/CHAP10/HEAD03.HTM.

[60] Fio.cz. *Klouzavý průměr (Moving average)* [online]. [Poslední úprava: 9.4.2010 13:13]. Dostupné z: http://www.fio.cz/spolecnost-fio/slovnik/klouzavy-prumer-moving-average

[61] Akcie.cz. *Vývoj indexu Standard and Poor's 500* [online]. [cit. 2015-02-02]. Dostupné z: http://www.akcie.cz/kurzy-svet/indexy-svet/sp500

[62] IPCC.CH. *CLIMATE CHANGE 2001: THE SCIENTIFIC BASIS* [online]. [cit. 2015-05-01]. Dostupné z: http://www.ipcc.ch/ipccreports/tar/wg1/311.htm

[63] Nature.com. *Academy affirms hockey-stick graph* [online]. [cit. 2015-05-01]. Dostupné z: http://www.nature.com/nature/journal/v441/n7097/full/4411032a.html

[64] Nature.com. *Climatologists under pressure* [online]. [cit. 2015-05-01]. Dostupné z: http://www.nature.com/nature/journal/v462/n7273/full/462545a.html

[65] Wikipedia.org. *Hockey stick controversy* [online]. [cit. 2015-05-01]. Dostupné z: http://en.wikipedia.org/wiki/Hockey_stick_controversy

[66] Wikipedia.org. *Climatic Research Unit email controversy* [online]. [cit. 2015-05-01]. Dostupné z: http://en.wikipedia.org/wiki/Climatic_Research_Unit_email_controversy

[67] Logan, Tina. *Getting Started in Candlestick Charting*. John Wiley & Sons, Inc, 2008. ISBN-13: 978-0470182000.

[68] *Moving Average* [online]. CMSForex.com. [cit. 2015-05-01]. Dostupné z: http://www.cmsfx.com/en/forex-education/technical-analysis-articles/moving-average-indicators/moving-average-indicator/calculation/

[69] *What is the 'Best' Time Frame to Trade?* [online]. DailyFX.com. [cit. 2015-05-01]. Dostupné z: http://www.dailyfx.com/forex/education/trading_tips/daily_trading_lesson/2014/03/26/The-Best-Time-Frame.html

[70] Lien, Kathy. Making Sense Of The EUR/CHF Relationship . In: *Investopedia.com* [online]. © 2015, Investopedia. [cit. 2015-05-04]. Dostupné z: http://www.investopedia.com/articles/forex/06/eurchfrelationship.asp.

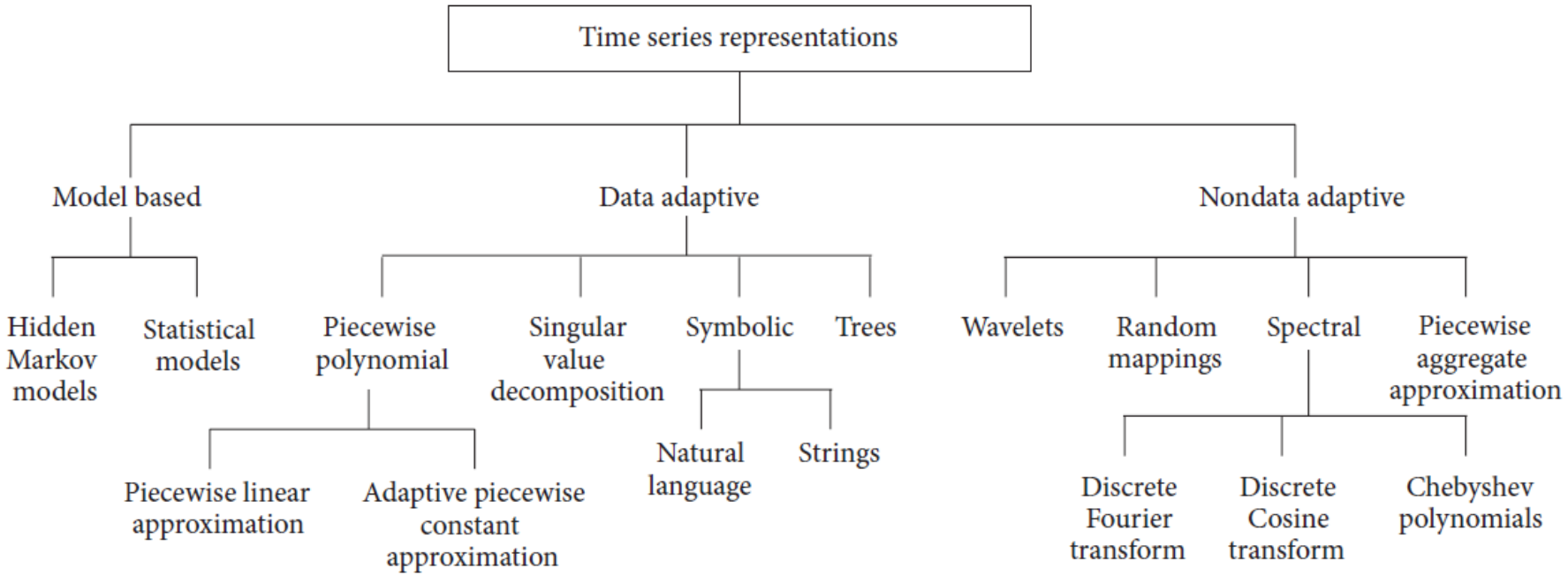
[71] *Currency Pair Correlations* [online]. Cashbackforex.com. [cit. 2015-05-04]. Dostupné z: https://www.cashbackforex.com/en-us/school/tabid/426/ID/437424/currency-pair-correlations

[72] *Špalek, J.* Studijní materiály předmětu ESF:PVSTAP. Masarykova univerzita. [online]. 2006 [cit. 2015-05-17]. Dostupné z: http://is.muni.cz/el/1456/jaro2006/PVSTAP/um/1266212/

A. Seznam obrázků

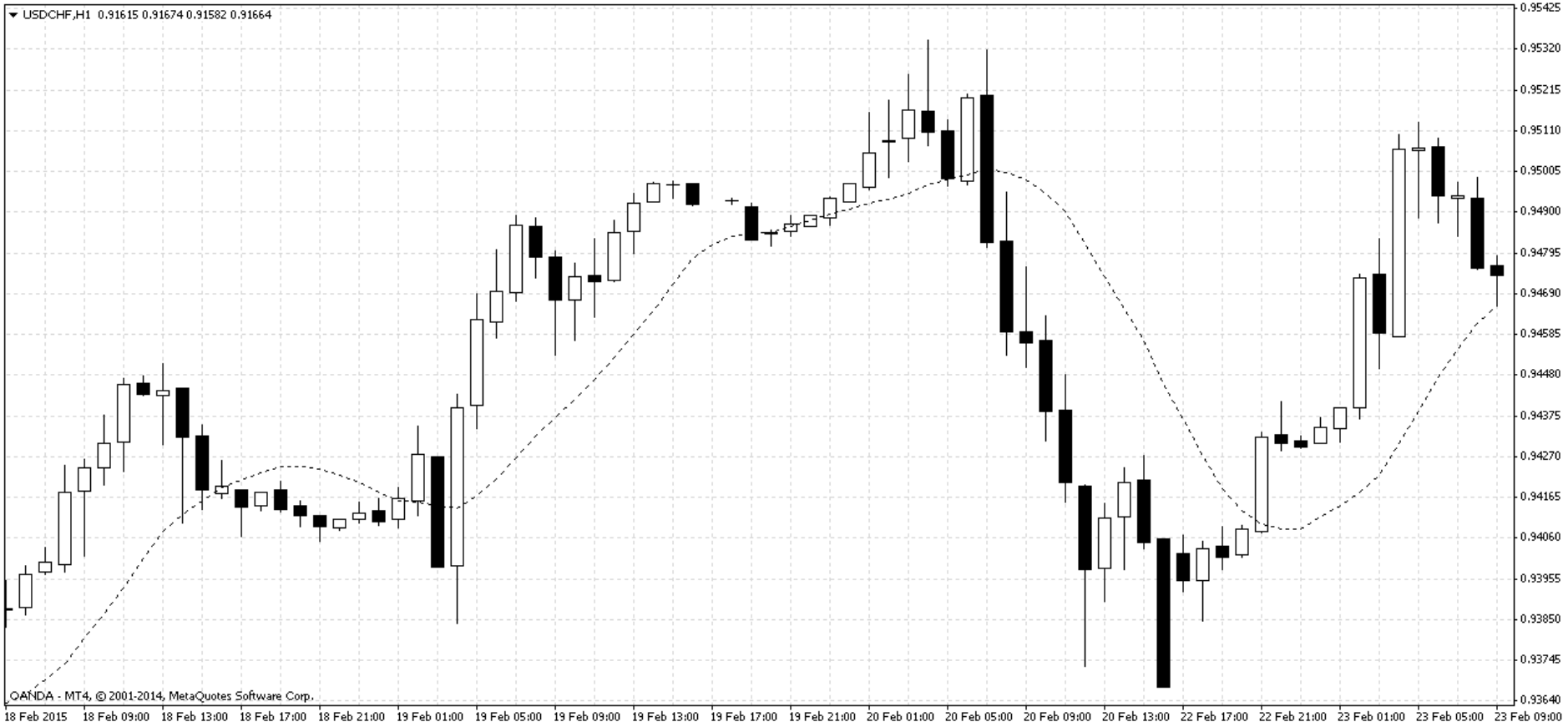
B. Obrázky ve větším rozlišení

Zde jsou uvedeny vybrané obrázky ve větším rozlišení.

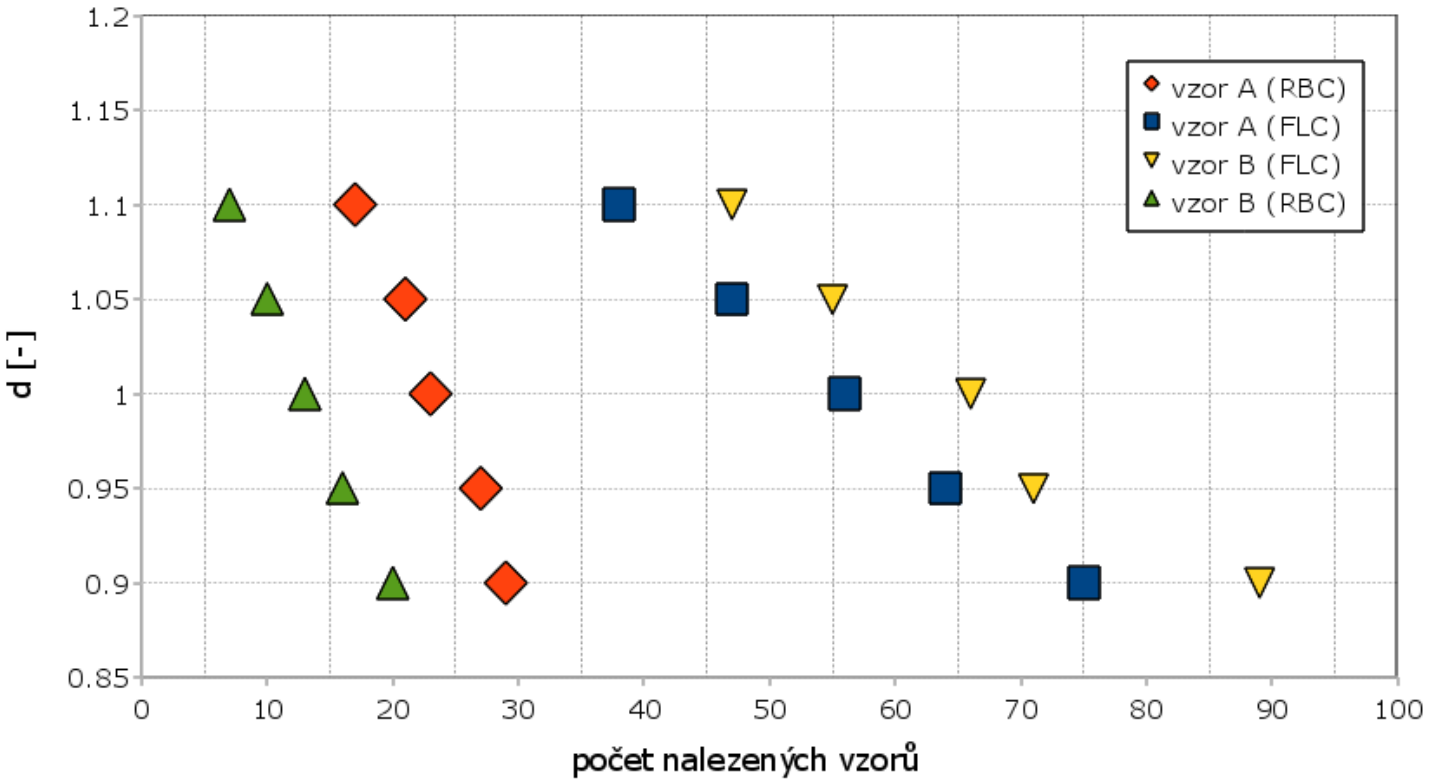


Obrázek B.1: Detailní rozdělení reprezentace časových řad (převzato z [18])

Obrázek B.2: Příklad časové řady reprezentované svícovým grafem



Obrázek B.3: Demonstrace MA na měnovém páru USDCHF 1H pro b = 12



Obrázek B.4: Počet nalezených vzorů v závislosti na parametru d

C. Matice průměrných vzorů

Matice prumeru prvku pro vzor A metodou rule-based:
0.00314 0.00338 0.00652
0.00315 0.00328 0.00643
0.00340 0.00354 0.00694
0.00340 0.00336 0.00676

Matice prumeru prvku pro vzor A metodou fuzzy:
0.00219 0.00250 0.00469
0.00218 0.00249 0.00468
0.00253 0.00284 0.00537
0.00249 0.00268 0.00518

Matice prumeru prvku pro vzor B metodou rule-based:
-0.00299 -0.00352 -0.00651
-0.00367 -0.00371 -0.00738
-0.00285 -0.00358 -0.00643
-0.00352 -0.00371 -0.00722

Matice prumeru prvku pro vzor B metodou fuzzy:
-0.00200 -0.00235 -0.00435
-0.00244 -0.00276 -0.00520
-0.00192 -0.00235 -0.00427
-0.00233 -0.00275 -0.00508

D. Vybrané dílčí výpočty a hodnoty

Následující část výpočtů se vztahuje zvláště ke kapitole 4.5.4.
Výpočty jsou též uvedeny v elektronické příloze.

V tabulce D.1 uvádím opět hodnoty pro metodu FLC s konkrétními výpočty. Hodnota X zde představuje PPV_{FLC}, OK představuje počet vzorů, které byly vyhodnoceny jako korektní.

i	FLC	OK	X_i	$(X_i - \bar{x})^2$
1	75	36	36/75	0,0006416134
2	64	32	32/64	0,002054817
3	56	30	30/56	0,0065681908
4	47	29	29/47	0,0263579661
5	38	25	25/38	0,0413003301
6	89	31	31/89	0,0113114506
7	71	25	25/71	0,0105179863
8	66	23	23/66	0,0112752674
9	55	19	19/55	0,0119279959
10	47	17	17/47	0,0086430086
	$\sum_{FLC} = 608$	$\sum_{OK} = 267$	$\sum X_i = 4{,}5467$	$\sum (X - \bar{x})^2 = 0{,}1306$
			$\bar{x} = 0{,}4547$	Med(X) = 42,085 %

Tabulka D.1: Přepsáno z kapitoly 4.5.4 Korektní data nalezená metodou rule-based

Výběrový rozptyl

$$s^2_{PPV_{FLC}} = \frac{\sum (X - \bar{x})^2}{10 - 1} = \frac{0,1306}{9} = 0,01451.$$

V tabulce D.2 uvádím znovu hodnoty pro metodu RBC s konkrétními výpočty. Hodnota Y zde představuje PPV_{RBC}, OK představuje počet vzorů, které byly vyhodnoceny jako korektní.

i	RBC	OK	Y_i	$(Y_i - \bar{y})^2$
1	29	17	17/29	0,5862068966
2	27	16	16/27	0,5925925926
3	23	14	14/23	0,6086956522
4	21	13	13/21	0,619047619
5	17	11	11/17	0,6470588235
6	20	10	10/20	0,5
7	16	9	9/16	0,5625
8	13	7	7/13	0,5384615385
9	10	7	7/10	0,7
10	7	6	6/7	0,8571428571
	$\sum_{RBC} = 183$	$\sum_{OK} = 110$	$\sum Y_i = 6{,}2117$ $\bar{y} = 0{,}6212$	$\sum (Y - \bar{y})^2 = 0{,}0897$ Med(Y) = 60,064 %

Tabulka D.2: Přepsáno z kapitoly 4.5.4 Korektní data nalezená metodou rule-based

Výběrový rozptyl

$$s^2_{PPV_{RBC}} = \frac{\sum (Y - \bar{y})^2}{10 - 1} = \frac{0,0897}{9} = 0,00997.$$

E. Obsah přiloženého CD

- `fx_bakala` – NetBeans projekt včetně knihoven, zdrojových a nalezených dat
- `vypocty.ods` – statistické výpočty
- `prace.pdf` – text této práce