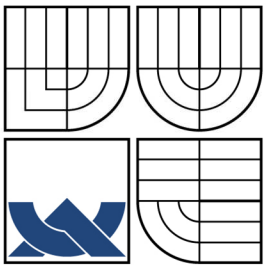
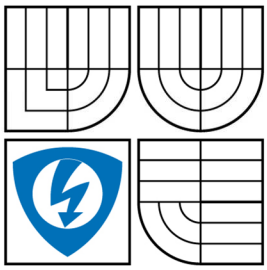


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV AUTOMATIZACE A MĚŘÍCÍ TECHNIKY

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF CONTROL AND INSTRUMENTATION

VYHLEDÁVÁNÍ VZORŮ V DYNAMICKÝCH DATECH

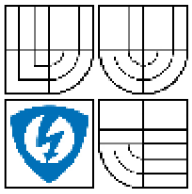
PATTERN FINDING IN DYNAMICAL DATA

DIPLOMOVÁ PRÁCE
MASTER´S THESIS

AUTOR PRÁCE — Bc. JAN BUDÍK
AUTHOR

VEDOUCÍ PRÁCE — Ing. PETR HONZÍK, PhD.
SUPERVISOR

BRNO 2009



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav automatizace a měřicí techniky

Diplomová práce

magisterský navazující studijní obor
Kybernetika, automatizace a měření

Student:	Bc. Jan Budík	***ID:***	83894
Ročník:	2	***Akademický rok:***	2008/2009

NÁZEV TÉMATU:

Vyhledávání vzorů v dynamických datech

POKYNY PRO VYPRACOVÁNÍ:

Cílem diplomové práce je provést rešerši v oblasti popisu a analýzy dynamických dat a dále detekce a vyhledávání vzorů v časových řadách. Dále navrhněte a realizujte systém, který bude v dynamických datech vyhledávat vzory, na jejichž základě budou rozpoznávány požadované události. Využijte jednu z metod rozpoznávání vzorů (pattern recognition). Můžete využít data z oblasti burzy cenných papírů nebo komoditních trhů.

DOPORUČENÁ LITERATURA:

Dle vlastního literárního průzkumu a doporučení vedoucího práce.

Termín zadání:	9.2.2009	***Termín odevzdání:***	25.5.2009

Vedoucí práce: Ing. Petr Honzík, Ph.D.

prof. Ing. Pavel Jura, CSc.
Předseda oborové rady

UPOZORNĚNÍ:

Autor diplomové práce nesmí při vytváření diplomové práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

Vysoké učení technické v Brně

Fakulta elektrotechniky a komunikačních technologií

Ústav automatizace a měřící techniky

Vyhledávání vzorů v dynamických datech

Diplomová práce

Specializace studia:	Kybernetika, automatizace a měření
Student:	Bc. Jan Budík
Vedoucí:	Ing. Petr Honzík, PhD.

Abstrakt :

Ve své diplomové práci se zaměřuji na problematiku rozpoznávání vzorů v dynamických datech. Dynamická data jsou volená burzovní data, která reprezentují cenový vývoj finančních instrumentů. Je vytvořena aplikace, která slouží k nalezení takových vzorů, které disponují určitou predikční schopností a je provedeno několika násobné ověřování na nových datech. Je volena metoda učení založená na instancích. Výstupem programu je série vzorů, které vykazují predikční schopnosti i mimo oblast učících dat.

Klíčová slova: IBL, časové řady, burza, forex, strojové učení, analýza časových řad, rozpoznávání vzorů, statistické předpovědi, klouzavé průměry

Brno University of Technology

The Faculty of Electrical Engineering and Communication

Department of Control, Measurement and Instrumentation

Pattern Finding in Dymanical Data

Master's Thesis

Specialisation of study: Cybernetics, Control and Measurement

Student: Bc. Jan Budík

Supervisor: Ing. Petr Honzík, PhD.

Abstract :

In my Master't thesis I'm focused on pattern finding in dynamical data. I used as a dynamical data data from electronics markets, which are represented by price moving of financial instruments. I did applpication, which can find patterns in these data and founded patterns have predictive quality. After learning proces are used verification.

There are choosed learning method, which is based on instance – Instance based learning. Output of application are patterns, which have predictive quality after verification.

Key words : Instance based learning, time series, markets, forex, datafeed, pattern recognition, machine learning, time series analyses, moving averages

Anotace

V první kapitole je nastíněna problematika rozpoznávání vzorů. Druhá kapitola pojednává o možných řešeních problému za použití umělé inteligence a popisuje základní teorie statistiky a chaosu. Třetí kapitola je zaměřena na problematiku časových řad, jejich typů, problémů a předzpracování. Je zde také popsán typ časových řad ve finančnictví. Čtvrtá kapitola pojednává o problematice rozpoznávání vzorů a predikce. Je zde popsána metoda učení, která je použita. Poslední kapitola popisuje vývoj programu a jeho jednotlivé části a jsou zde zobrazeny dosažené výsledky.

Resumé

Při rešerži a vývoji programu jsem se seznámil s aplikací metod umělé intelligence při vyhledávání vzorů. Seznámil jsem se s použitou metodou učení založené na podobnosti instancí. Pochopil jsem důležitost metod ověřování mimo oblast učících dat, kdy je tento prvek klíčový při rozhodnutí, zda-li vzor má predikční schopnosti. Seznámil jsem se s časovými řadami ve finančnictví a provedl nalezení vzorů, které za určitých podmínek predikovaly úspěšně budoucí vývoj.

Annotation

First chapter is about basic information pattern learning. Second chapter is about solutions of pattern recognition and about using artificial inteligence and there are basic informations about statistics and theory of chaos. Third chapter is focused on time series, types of time series and preprocessing. There are informations about time series in financial sector.
Fourth charter discuss about pattern recognition problems and about prediction.
Last charter is about software, which I did and there are informations about part sof program.

Summrary

In this master's thesis I did an application, which are using artificial inteligence for pattern recognition. I introduced with method of machine learning – Instance Based Learning. I understand important of verification in out of sample data file. This is one of most important process. I understand time series in financial sector now and I found patterns, which have predictive quality.

Bibliografická citace

BUDÍK, J. Vyhledávání vzorů v dynamických datech. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2009. 76 stran. Vedoucí diplomové práce Ing. Petr Honzík, PhD.

P r o h l á š e n í

„Prohlašuji, že svou diplomovou práci na téma Vyhledávání vzorů v dynamických datech jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem úmyslně neporušil autorská práva třetích osob, zejména jsem úmyslně nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.“

V Brně dne 25.května 2009 Podpis:

P o d ě k o v á n í

Děkuji vedoucímu diplomové práce Ing. Petru Honzíkovi, PhD. za odbornou a metodickou pomoc při hledání řešení problémů zpracování mé diplomové práce. Dále děkuji rodině za podporu při studiu na vysoké škole.

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

SEZNAM ROVNIC

1. ÚVOD

Hlavním cílem diplomové práce je udělat rešerši v oblasti rozeznávání vzorů v dynamických datech. Dynamická data bude reprezentovat časová řada, která vyjadřuje změnu ceny finančních instrumentů.

Cílem rozpoznávání vzorů je klasifikace konkrétní situace. Na základě statistické informace pak zpravidla následuje další rozhodnutí. Ve své podstatě jde o to, že dle statistických výpočtů jsme schopni určit, zda daný jev má momentálně velkou pravděpodobnost výskytu. Rozpoznávání vzorů je podoblast strojového učení.

Rozpoznávání vzorů klasifikuje vzory na základě apriorní znalosti nebo na základě statistických informací získaných z dat. Vzory jsou většinou určeny na základě měření nebo pozorování.

Časová řada je chronologicky uspořádaná posloupnost hodnot určité sledované veličiny. Prakticky to znamená, že časová řada je řada čísel. Tuto řadu tvoří hodnoty určité (např. ekonomické, fyzikální) veličiny, které jsou uspořádány od nejstarších po nejmladší nebo naopak. Z formálního hlediska je časová řada realizací náhodného procesu. Typickým příkladem časové řady je například tabulka vývoje inflace v České republice, záznam průběhu teploty ovzduší během dne, záznam průběhu akciového, komoditního nebo forexového subjektu během obchodní seance.

Cílem práce je tedy vyvinou aplikaci, ve které jsou implementovány algoritmy umělé inteligence pro rozeznávání vzorů v dynamických datech. Dynamická data budou tvořena časovým průběhem vývoje ceny finančních instrumentů. Výstupem aplikace bude série vzorů vykazující predikční schopnosti.

2. MOŽNÁ ŘEŠENÍ PROBLÉMU

2.1 NEURONOVÉ SÍTĚ

S vývojem počítačů vzniká vývoj umělé inteligence, kdy se člověk snaží přenechat řešení složitých a komplexních problémů na počítači. Pro aplikaci těchto metod je třeba mít kvalitní a dostatečně velká vstupní data.

Principem metod neuronových sítí je napodobování funkčnosti lidského mozku, který se učí podle zkušeností. Základ těchto sítí tvoří neurony, které jsou vzájemně spojeny, což znamená, že výstup neuronu závisí na předchozích neuronech. Jednotlivé vstupy jsou ohodnoceny váhami. Při tvorbě sítě jsou váhy zadány náhodně a je třeba síť podrobit procesu učení, čili lze říci, že síť je na počátku „hloupá“ a na konci procesu vykazuje v rámci mezí inteligenci nabytou učením.

Jako vstupní data může být použita právě finanční informace (zadlužení, zisk na akcii, …) a výstupem je výsledný kurz akcie. Tento výsledek se porovnává s realitou a síť se učí do té doby, než bude vykazovat ve většině případů stejné výstupy. Základním rozdílem od standardních jednodušších algoritmů je možnost adaptace. Nevýhodou je výpočetní, finanční a programová náročnost.

Ve finančnictví se neuronových sítí využívá k predikci cen instrumentů, výnosů a dalších ekonomických jevů či odhadů. V ostatních odvětvích slouží tyto sítě například k rozeznání rukopisu.

2.2 DATAMINING

Systém dolování z dat je založen na získání informace, která má určitý statistický význam a lze na jejím základě provádět predikci.

Různorodost vstupních dat je příčinou využití dataminingu v mnoha odvětvích. Jako příklad lze uvést rozhodování v marketingových strategiích nebo aplikace v medicíně, kdy lze predikovat potencionální výskyt infarktu.

Ve finančním sektoru lze aplikovat datamining pro predikci vývoje ceny finančních instrumentů. Vstupní data tvoří databáze historických cen a je zde pomocí metodických postupů dataminingu hledána určitá nenáhodnost.

2.3 EXPERTNÍ SYSTÉMY

Expertní systém je specifickou aplikací umělé inteligence, která se snaží simulovat rozhodování experta při řešení úloh. Bázi znalostí vypracovává expert, který zná dokonale prostředí a situace, kde bude systém aplikován. Na základě těchto znalostí jsou za pomoci uživatelského rozhraní uživateli pokládány otázky a podle odpovědí je postupně dosaženo výsledku.

Výhodou expertních systémů je schopnost rozhodování i při neurčitém stavu. Ve finančním sektoru může být expertního systému využito například při žádostech o půjčky, hypotéky a také pro vyhodnocení tržní situace a následných kroků.

2.4 ZÁKLADNÍ TEORIE

2.4.1 Statistika

Statistika je odvětví analytické matematiky, které se zabývá definicí vlastností a pravidel vyskytujících se v pozorovaných datech. Využití je velmi široké, od výzkumu veřejného mínění až po analýzu vývoje cen finančních instrumentů.

U predikce vývoje cen je využíváno statistických výsledků pozorovaných dat a následného rozhodování dle těchto výsledků. K statistické analýze slouží sofistikované programy, jejichž jádro bývá často tvořeno metodami umělé inteligence.

Cílem je vytvořit program, který dokáže naklonit na naši stranu statistickou výhodu. Tím se v praxi myslí, že dokážeme například s 60% úspěšností odhadnout pohyb ceny při poměru ztráta/profit rovno 1/1.

2.4.2 Teorie chaosu

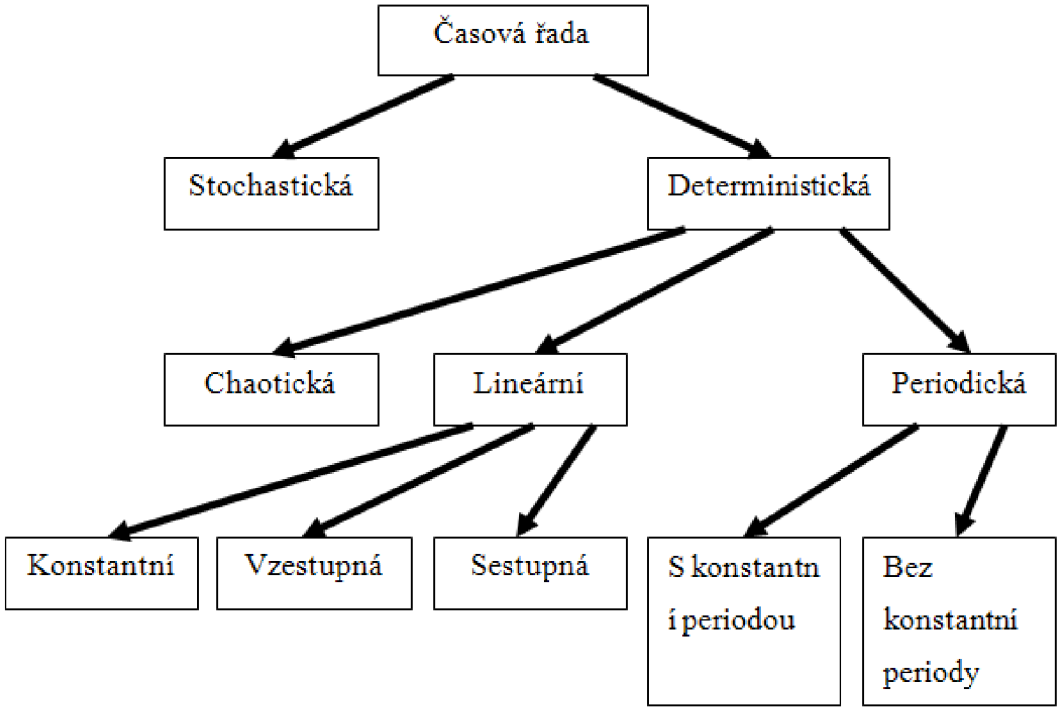
Mnoho lidí, kteří spekulují nad predikcí časových řad, se postupem času alespoň seznámí s teorií chaosu.

Jedná se o teorii, která se snaží popsat nenáhodnost chování nelineárních systémů, které na první pohled vykazují prvky chaotického chování. V této práci se nelineárním systémem myslí trh, jehož výstupem je časová řada, která reprezentuje cenový vývoj.

Vývoj cen se na první pohled může zdát chaotický, ale provedeme-li analýzu, dokážeme, že tomu tak ve skutečnosti často není a že i cena se pohybuje s určitou pravděpodobností podle pravidel a dokážeme-li tuto pravděpodobnost naklonit na svou stranu, jsme schopni do dostatečné míry správně predikovat vývoj.

3. ČASOVÉ ŘADY

3.1 ZÁKLADNÍ DĚLENÍ ČASOVÝCH ŘAD



Obrázek 1: Dělení časových řad[2]

Deterministická řada se vyznačuje tím, že neobsahuje prvek náhodnosti a lze ji tedy se 100% úspěšností předpovídat podle definovaného vztahu. U tohoto typu řady neexistuje neznámá proměnná, která by způsobovala náhodné chování. Příkladem deterministické časové řady je posloupnost hodnot funkce sinus, kdy jsme schopni v každém bodě vypočíst následující prvek.

Opak deterministických řad jsou řady stochastické. Stochastické časové řady obsahují prvek náhodnosti, tedy prvek, který způsobuje náhodné generování průběhu. V elektrotechnice náhodnost představuje šum, nepřesnosti či jiné nepříznivé vlivy a v ekonomice lze za tyto jevy brát neočekávané situace, jako například přírodní pohromy, války či hodnoty fundamentálních ukazatelů.

Řady se dělí také podle informace, které jsou nositelem:

- řada absolutních ukazatelů
- řada odvozených charakteristik

Časová řada absolutních ukazatelů je řada, která vznikla pozorováním nebo měřením. Časová řada odvozených charakteristik je v podstatě transformace, která může být provedena například klouzavými průměry či jinými technickými indikátory.

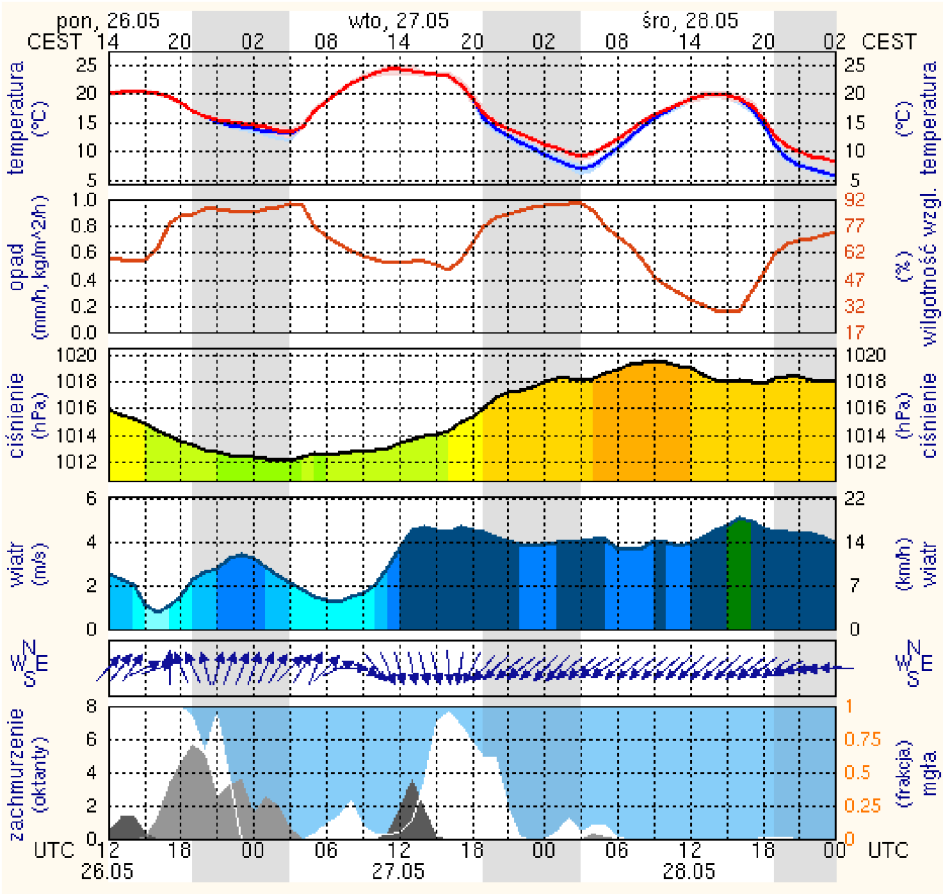
Některé transformace mění charakter časové řady a proto je třeba brát v úvahu, že transformované řady mají prvky závislé na rozdíl od netransformovaných řad.

3.2 PŘÍKLADY ČASOVÝCH ŘAD

Řada na obr. č. 2 reprezentuje záznam průběhu změny cenové hladiny v čase. Jednotlivé úsečky zaznamenávají informaci po dobu zvolené periody. U tohoto příkladu je volena časová perioda 5-ti minut, čili jedna úsečka nese informaci vývoje ceny po dobu 5-ti minut.

Obrázek 2: Příklad časové řady ve finančnictví [15]

Řada na obr. č. 3 je záznamem změny fyzikálních jevů po dobu tří dnů. V této situaci nejde o příliš citlivá data a je tedy volena perioda záznamu 1 hodina, což je pro vyhodnocování těchto veličin dostačující.



Obrázek 3: Příklad časové řady fyzikálních veličin [14]

3.3 VÝZNAM A CÍLE ANALÝZY ČASOVÝCH ŘAD

Cíle analýzy časových dat spočívají v porozumění principům, dle kterých hodnoty časových řad vznikají. Většinou se snažíme sestrojit vhodný model, podle kterého vznikají hodnoty. Tyto analýzy mohou následně sloužit k predikci budoucího vývoje časové řady.

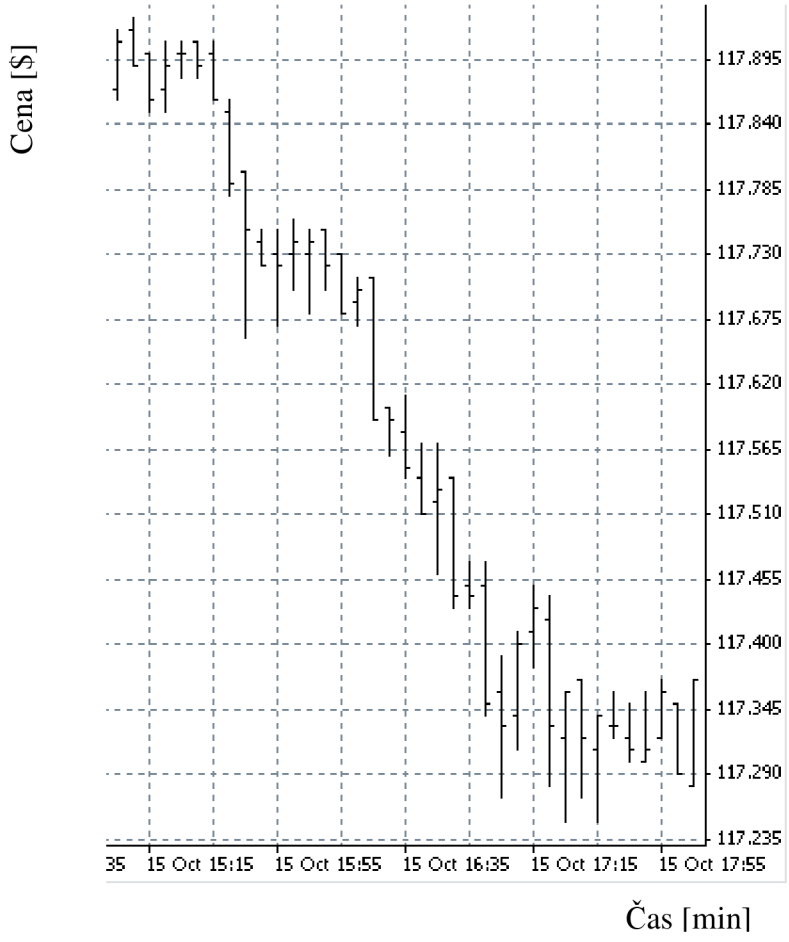
3.4 PROBLÉMY ČASOVÝCH ŘAD

- problémy s volbou časových bodů pozorování
- problémy s kalendářem
  - různá délka měsíců
  - různý počet víkendů v měsíci
  - různý počet pracovních dnů v měsíci
  - pohyblivé svátky
- problémy s délkou časových řad
- problémy nesrovnatelností dat

Obecně nelze stanovit univerzální dobu periody, jelikož každý sledovaný jev vykazuje jiné parametry změny. Při záznamu průběhu cen finančních instrumentů je třeba volit co nejmenší periodu vzorkování. Perioda je volena řádově v desetinách sekund a je dále podle potřeby uživatele transformována do požadované periody. Existují také jevy, které nepotřebují být takto přesně zaznamenávány, jelikož se podle nich neprovádějí závažná rozhodnutí a jejich velká skoková změna je na základě fyzikálních zákonů vyloučena. Jako příklad lze uvést měření teploty ovzduší, kdy pro dostačující měření lze volit periodu jedné hodiny a pro citlivé měření například periodu 15-ti minut.

3.5 GRAFICKÁ A PSYCHOLOGICKÁ ANALÝZA

Grafické metody analýzy časových řad jsou asi nejjednodušším způsobem analýzy. Ve své podstatě jde o rozhodnutí trendu podle subjektivních pocitů z grafu. Na obr. č. 4 můžeme dle pohledu rozhodnout, že jde o klesající trend. Tato tvrzení můžeme dále potvrdit pomocnými indikátory (klouzavé průměry, oscilátory…). Pro tento přístup je však do jisté míry potřebný lidský cit pro čtení grafů.



Obrázek 4: Klesající trend časové řady [15]

3.6 ZÁKLADNÍ ÚPRAVY ČASOVÝCH ŘAD

Při zpracování časových řad může nastat situace, kdy mohou chybět některé prvky vynecháním snímacího procesu například poruchou. Jde-li o proces, kde záznam dat je prováděn pouze jedním prvkem (snímačem, počítačem), dojde k nenávratné ztrátě. Je-li záznam realizován více prvky, lze tato data poskládat dohromady.

Doplnění chybějících hodnot:

Jde-li o proces, který při analýze časové řady není schopen akceptovat chybějící data, můžeme vhodně tato data doplnit sami. Doplněná data téměř nikdy nebudou tak kvalitní, jako by byla data skutečná, ale v určitých případech je toto řešení možné.

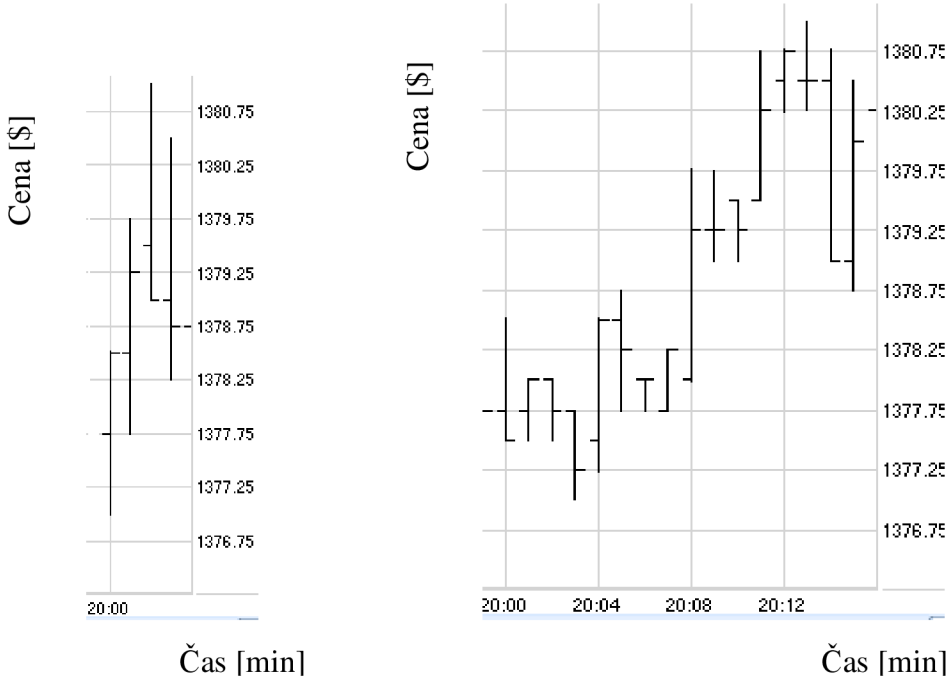
- Náhrada nulami, nevíme-li o řadě nic nebo pouze to, že její průměrný prvek je roven nule.
- Doplnění dat například aritmetickým průměrem, klouzavým průměrem či mediánem.
- Lineární interpolací. Tato metoda se hodí pro řady, které vykazují setrvačnost.
- Nahrazení trendem v celém souboru, který lze získat regresí.
- Odhadem, který je založen na identifikovaném modelu chování.

3.7 TYPY ČASOVÝCH ŘAD VE FINANČNICTVÍ

Řada založena na čase:

Hodnoty časové řady jsou zaznamenávány v přesně definovaných časových intervalech. Jako příklad lze uvést srovnání dvou řad na obrázku č. 5, kdy u první řady (vpravo) je volena časová perioda 1 minuta a u druhé řady (vlevo) perioda 5-ti minutová. Záznam s menší časovou periodou poskytuje citlivější data. Volba periody záznamu je závislá na sledovaném jevu.

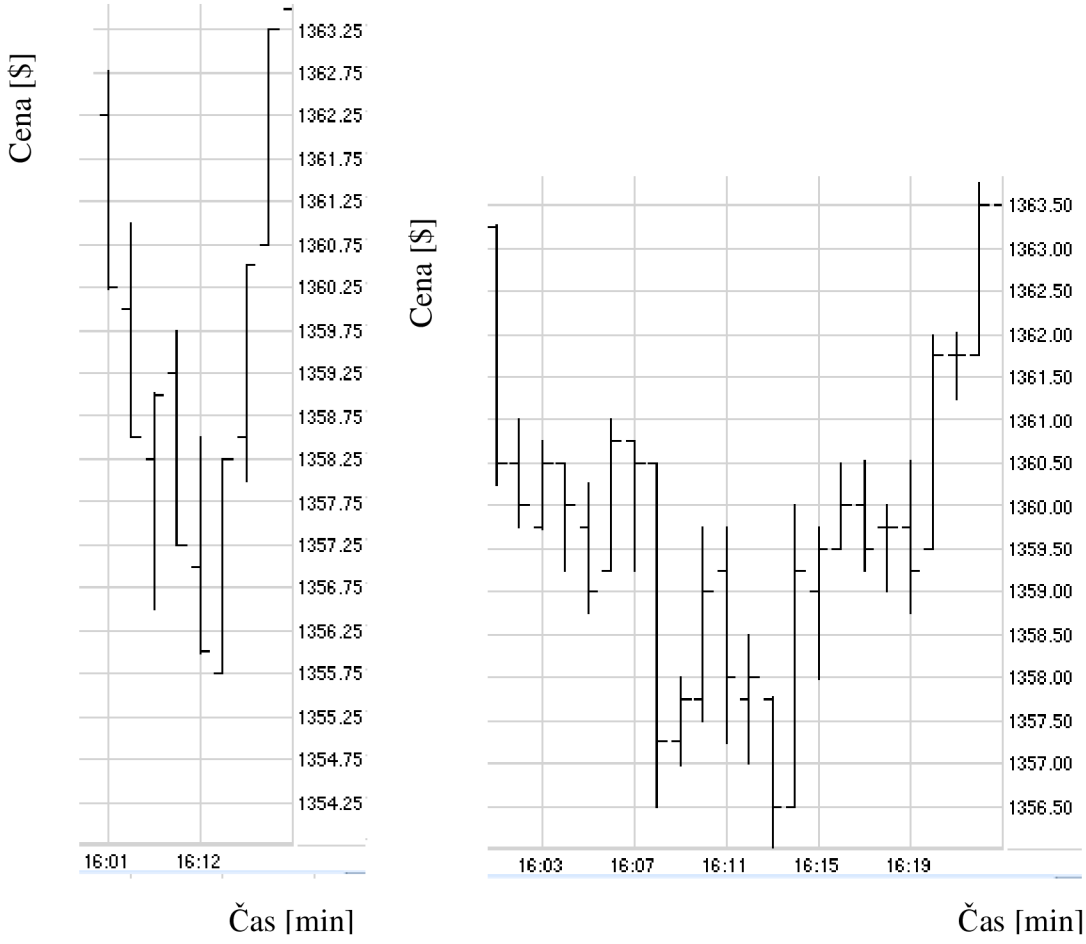
U měření teploty v průběhu dne je zbytečné volit minutovou periodu vzorkování, kdežto například u záznamu průběhu cen finančních instrumentů je nutné použít periodu vzorkování rovno vteřině i méně.



Obrázek 5: Řada založena na čase [15]

Řada založena na velikosti pohybu:

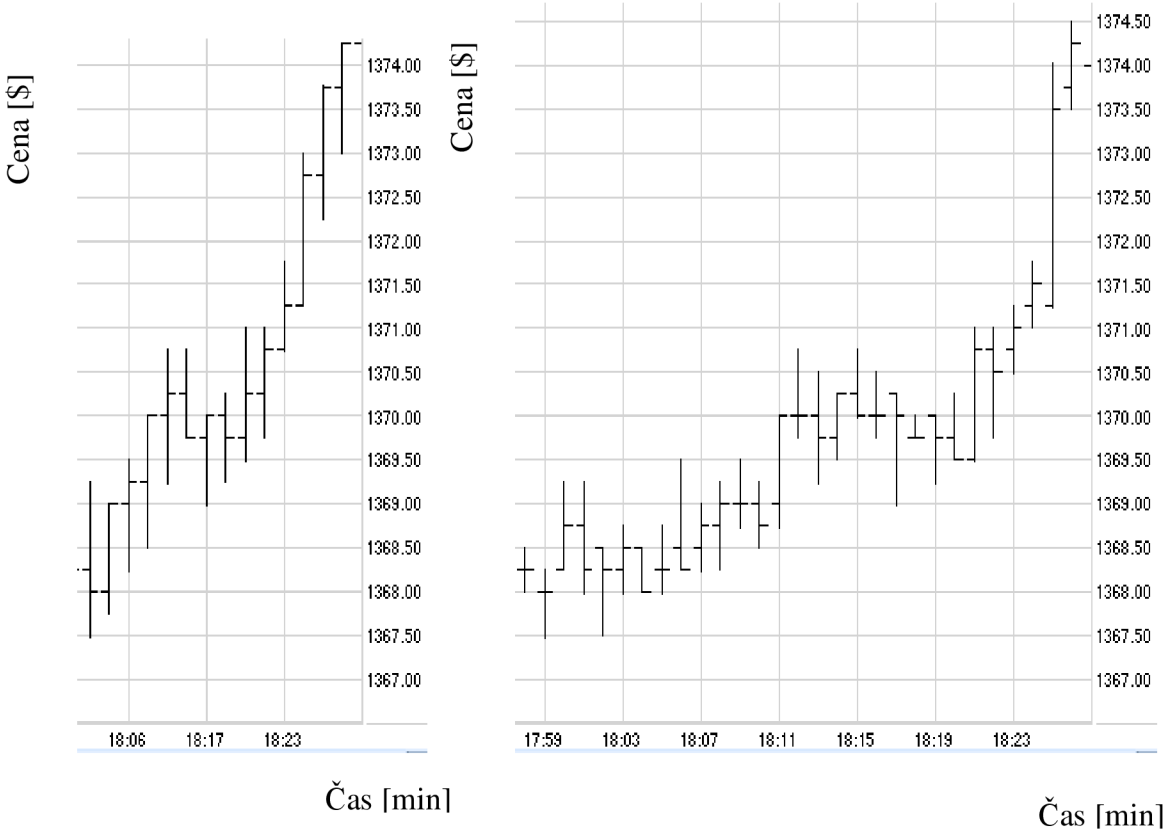
Tento speciální typ je využíván zkušenými obchodníky, jelikož dokáže eliminovat z určité části nechtěný šum. Podstata generování tohoto typu řady spočívá v definované hodnotě pohybu a po těchto částech je průběh rozdělován. Příklad tohoto typu řad je ukázán na obrázku č. 6 (vlevo) a je porovnán se stejným úsekem zaznamenaném pevně stanovenou periodou času 1 minuty (vpravo).



Obrázek 6: Řada založena na velikosti pohybu [15]

Řada založena na počtu provedených obchodů:

Poslední typ speciálních řad je založen na množství provedených obchodů. Uživatel si tedy definuje, po kolika provedených obchodech v trhu s vybraným finančním instrumentem dojde k tvorbě nové úsečky. Tento typ dat dává uživateli tu výhodu, že lze subjektivním pohledem zpozorovat, kdy na trhu probíhá právě zajímavá seance a kdy naopak téměř žádné obchody neprobíhají. V případě se zajímavou seancí probíhá tvorba nových úseček často, kdežto u nezajímavé seance nedochází téměř k žádné tvorbě úseček. Na obrázku č. 7 je porovnání časové řady generované na počtu obchodů (vlevo) a řady generované dle stanovené periody 1 minuta (vpravo).



Obrázek 7: Řada založena na množství provedených obchodů [15]

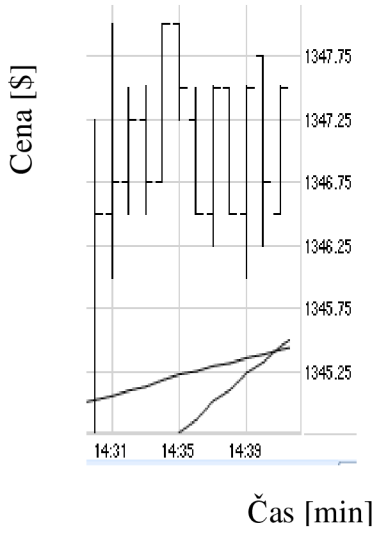
Z výše uvedených ukázek speciálních typů časových řad je patrné, že volbou typu časové řady můžeme také docílit jiných výsledků v oblasti predikce. Tyto typy řad dokážou částečně filtrovat vyskytující se šumy a tím zvýšit výkon predikce. Vzhledem k dostupnosti těchto dat bude pro predikci použito řad, založených na časové periodě.

4. PROBLEMATIKA ROZPOZNÁVÁNÍ VZORŮ A PREDIKCE

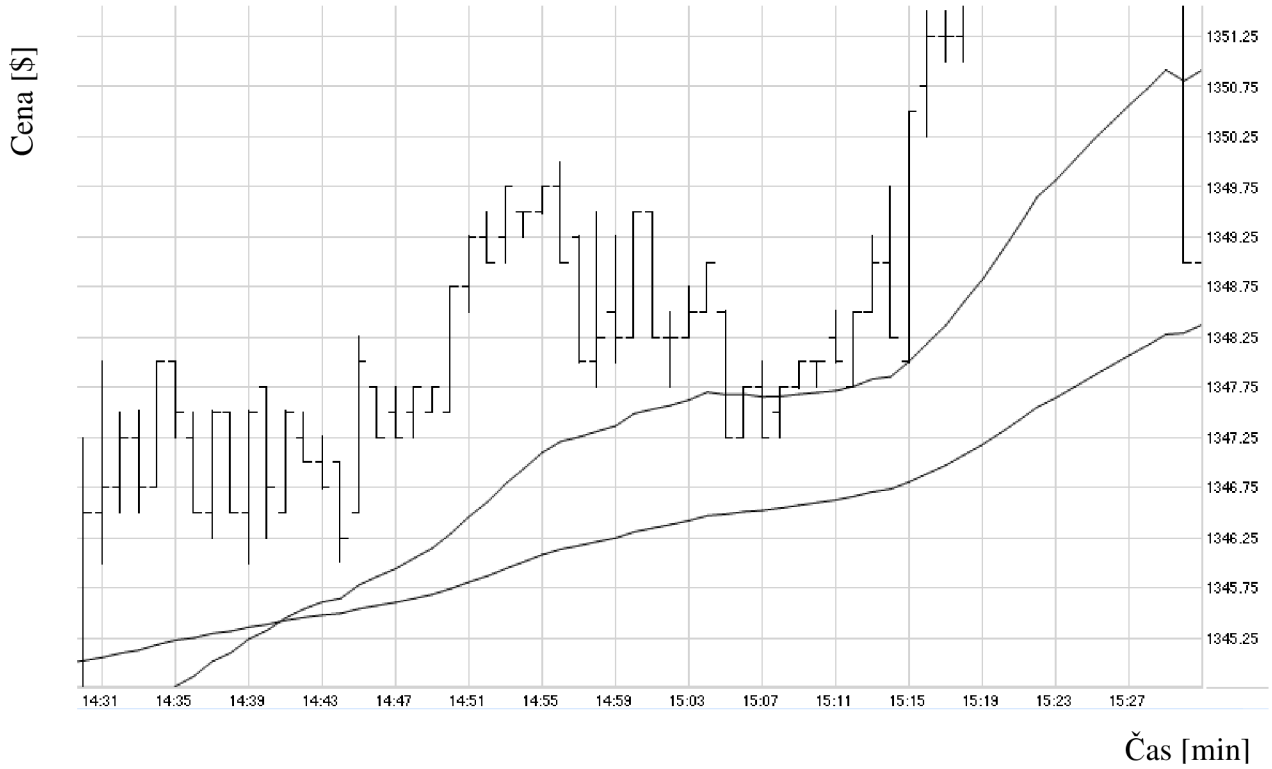
4.1 ÚVOD DO PROBLEMATIKY

Rozpoznávání vzorů je vědní disciplína, která se zabývá predikcí vývoje na základě již známých dat a modelů. Toho lze využít jak v technickém odvětví při predikci stavů různých fyzikálních systémů či jevů, tak i ve finanční analýze při předpovídání pohybu finančního instrumentu (akcie, komodity, měnové páry…).

Získání vzoru pro následnou predikci je možno pomocí apriorní znalosti nebo statistické informace, která je získána a vypočtena z požadované dynamické časové řady. Na obrázku č. 8 je znázorněn jeden z mnoha tisíců možných vzorů pro časovou řadu finančního páru EUR/USD. Tento vzor je založen na principu klouzavých průměrů s různou periodou a jejich křížení. Časová řada má periodu 5 minut. V tomto případě byl detekován vzor, který předpovídá pohyb ceny směrem nahoru. Jak je vidět na obrázku č. 9, došlo u tohoto vzoru k úspěšné predikci vývoje ceny.



Obrázek 8: Vzor na základě křížení klouzavých průměrů [15]



Obrázek 9: Následující pohyb po výskytu vzoru [15]

Pro rozpoznávání vzorů se využívá metod jednoduchých až po metody složité. Mezi metody složitější patří rozpoznávání vzorů pomocí aplikace neuronových sítí a jejich adaptaci na nové situace.

4.2 PREDIKCE

Predikce budoucího chování je potřebná v mnoha odvětvích k predikci různorodých veličin. K tomu účelu bylo vyvinuto mnoho různých metod. Tyto metody se dají dělit na tři základní typy. Metody využívající algoritmů, schopností se učit a metody založeny na heuristice.

Prvním krokem při provádění predikce je specifikace problému a je nutné vědět o povaze a typu zkoumaného prostředí.

Dále je nutné volit vhodnou metodu predikce, časový horizont odhadu a zda-li bude užito předzpracování dat.

Tvoření modelu predikce u ekonomických systémů, jako jsou například burzy, patří mezi nejobtížnější a nejméně spolehlivé. Trh tvoří statisíce lidí, kteří utvářejí cenu a každý se rozhoduje podle toho, jaké prvky na něho působí. Tímto vzniká prostředí, které nejde jednoduše definovat jednoduchým modelem či rovnicí a jeho stoprocentní predikce je téměř nemožná.

Důležitým prvkem při predikci je vyhodnocení úspěšnosti modelu. Úspěšnost je založena na odchylce hodnot predikovaných a hodnot, které skutečně nastaly.

4.3 HLAVNÍ CÍLE ALGORITMŮ ROZPOZNÁVÁNÍ VZORŮ

Cílem správně sestaveného algoritmu pro rozpoznávání vzorů je přesné určení třídy odpovídající danému znakovému vektoru na základě znalosti získané procesem trénování.

4.4 PRVKY PROCESU ROZPOZNÁVÁNÍ VZORŮ

Reálný problém rozpoznávání vzorů – hledáme vzor z reálného světa. Je ho třeba matematicky definovat.

Získání dat – získání dostatečného množství kvalitních dat.

Předzpracování – proces odstranění šumu, filtrování, normalizace.

Extrakce znaků – extrakce důležitých rysů z dostupných dat.

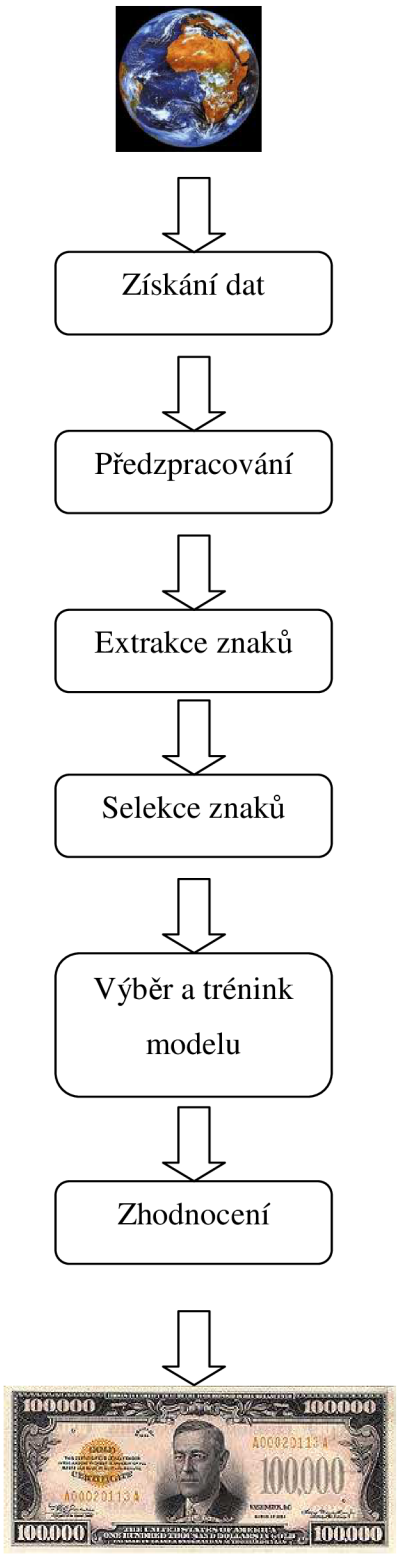
Selekce znaků – výběr množiny nejvíce relevantních znaků.

Výběr modelu – výběr správného typu modelu.

Trénink modelu – trénink vybraného typu modelu pomocí vhodného učícího algoritmu.

Zhodnocení – odhadnutí pravé výkonnosti klasifikátoru v reálném prostředí. Zjištění míry jistoty tohoto odhadu.

Řešení – vyřešení klasifikačního problému z reálného prostředí a následné automatické rozhodnutí.



Obrázek 10: Proces rozpoznávání vzorů [8]

4.5 METODA UČENÍ ZALOŽENÁ NA INSTANCÍCH (IBL)

Použitá metoda predikce v této diplomové práci je jedna z těch jednodušších, založených na algoritmech. Je to metoda využívající podobnosti situace známé se situací novou.

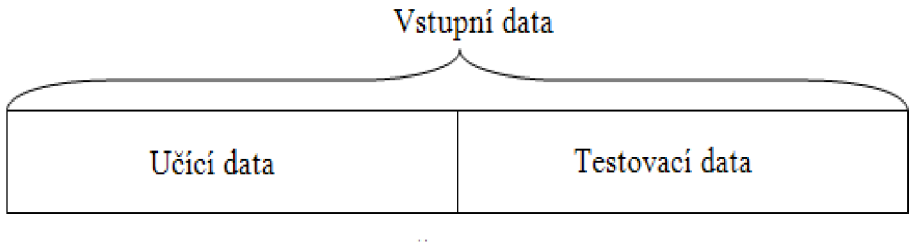
Algoritmus predikce je založen na uložení množiny učících dat do paměti a při vyskytnutí nové situace dojde pomocí algoritmů k prohledání uložených učících dat a zjištění informací, které poslouží pro klasifikaci nové situace.

Nevýhodou je časová náročnost prováděných výpočtů, jelikož tato metoda není založena na sestavení modelu, ale je vytvářen jedinečný model, který bývá často velmi jednoduchý.

Ukládání všech učících dat je mnohdy neefektivní a zbytečné, proto lze v určitých případech uložit pouze ta data, která mají statisticky významnou hodnotu. Tímto krokem zkrátíme čas potřebný pro výpočty a také ušetříme paměťové místo.

Největším úskalím metody založené na instancích je výborná predikce v oblasti učících dat, ale nepříliš úspěšná predikce v nových datech. Tento problém lze řešit sofistikovanými metodami ověřování úspěšnosti metody.

Jedna z těchto metod je založena na principu rozdělení dat na dvě poloviny. První polovina dat je použita jako učící databáze a druhá polovina slouží k otestování funkčnosti a naopak, viz obr. č. 11.



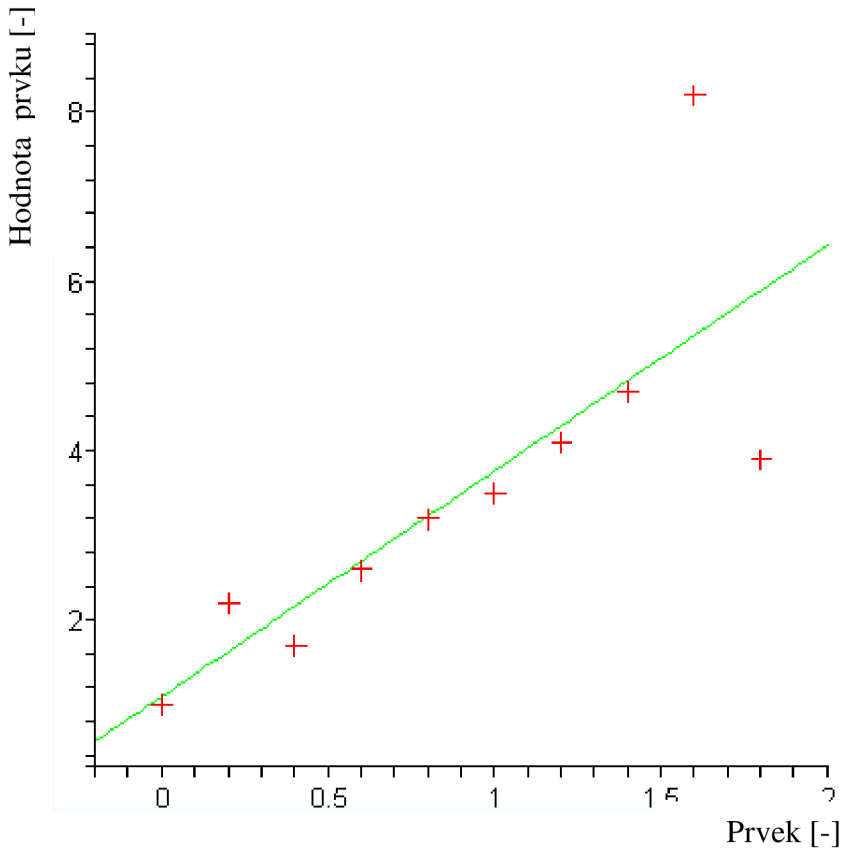
Obrázek 11: Vstupní data pro proces nalezení vzorů

4.6 CHYBOVÁ FUNKCE

Výpočet chybové funkce je základním prvkem metod učení. Cílem je číselně vyjádřit míru podobnosti dvou situací, z nichž jedna situace bude referenční a druhá bude situace nová, kterou je třeba klasifikovat.

4.6.1 Metoda nejmenších čtverců

K výpočtu chyby se dá využít metody nejmenších čtverců. Jde o metodu aproximační, kdy hledáme takové parametry funkce, kterou budeme prokládat, kdy bude součet čtverců odchylek vypočtených hodnot od hodnot naměřených nejmenší.



Obrázek 12: Lineární aproximace prvků [16]

4.6.2 Typy metod nejmenších čtverců

Metoda nejmenších čtverců se dá rozdělit pomocí typů funkcí, kterými se aproximuje.

- Aproximace přímkou
- Aproximace parabolou
- Aproximace polynomem

Aproximace přímkou:

Proložení daných hodnot je provedeno přímkou s rovnicí 1. Její koeficienty „a“ a „b“ lze vypočíst pomocí rovnic 2 a 3.

$$y = f(x, a, b) = ax + b$$

Rovnice 1: Rovnice přímky [16]

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

Rovnice 2: Výpočet koeficientu „a“ rovnice přímky [16]

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

Rovnice 3: Výpočet koeficientu „b“ rovnice přímky [16]

Aproximace parabolou:

Proložení daných hodnot je provedeno parabolou s rovnicí 4. Její koeficienty „a“, „b“ a „c“ lze vypočíst pomocí rovnic 5.

$$y = f(x, a, b, c) = ax^2 + bx + c$$

Rovnice 4: Rovnice paraboly [16]

$$\begin{aligned} a\sum x_i^4 &+ b\sum x_i^3 &+ c\sum x_i^2 &= \sum y_i x_i^2 \\ a\sum x_i^3 &+ b\sum x_i^2 &+ c\sum x_i &= \sum y_i x_i \\ a\sum x_i^2 &+ b\sum x_i &+ c\cdot n &= \sum y_i \end{aligned}$$

Rovnice 5: Soustava rovnic pro výpočet koeficientů „a“ , „b“, „c“ [16]

Aproximace polynomem:

Proložení daných hodnot je provedeno funkcí s rovnicí 6, jejích koeficienty lze vypočíst pomocí rovnic 7.

$$P_k = p_0 + p_1 x + \ldots + p_k x^k$$

Rovnice 6: Rovnice polynomu [16]

$$\begin{bmatrix} \sum x_i^{2k} & \cdots & \sum x_i^{k+1} & \sum x_i^k \\ \vdots & \ddots & \vdots & \vdots \\ \sum x_i^{k+1} & \cdots & \sum x_i^2 & \sum x_i \\ \sum x_i^k & \cdots & \sum x_i & n \end{bmatrix} \cdot \begin{bmatrix} p_k \\ \vdots \\ p_1 \\ p_0 \end{bmatrix} = \begin{bmatrix} \sum y_i x_i^k \\ \vdots \\ \sum y_i x_i \\ \sum y_i \end{bmatrix}$$

Rovnice 7: Soustava rovnic pro výpočet koeficientů polynomu [16]

5. VÝVOJ PROGRAMU

Základní princip vyhledávání vzorů v dynamických datech je založen na principu využití statistické výhody. Statistickou výhodou se myslí schopnost např. z 55% odhadnout směr pohybu ceny (obecně predikovat vývoj), při poměru zisk/risk 1/1.

Vzorem se myslí obecně situace, která se stala v minulosti. Na základě výskytu dané situace jsme schopni s určitou statistickou pravděpodobností určit následující vývoj. V případě této práce budou sloužit jako objekt sledování burzovní data.

5.1 REPREZENTACE BURZOVNÍCH DAT

Pro analýzu historických událostí a vyhledávání vzorů je použito burzovních dat. Tato data reprezentují pohyb ceny v závislosti na sledované veličině. Nejčastěji je sledovaná veličina čas, ale v praxi zkušení obchodníci s oblibou využívají generování průběhu ceny v závislosti na počtu provedených obchodů (tzv. Volume grafy) nebo v závislosti na velikosti pohybu (tzv. Range grafy). Tvorba a dostupnost těchto dat je však náročnější, kdežto data závislá na čase jsou levně a často zdarma přístupná.

Úsečka vývoje ceny je popsána čtyřmi základním parametry:

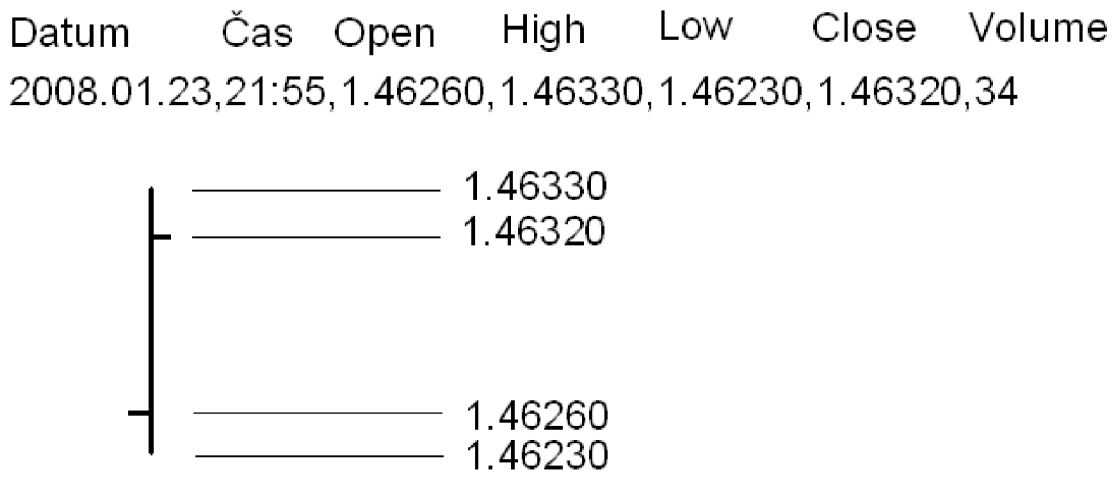
Open - počáteční cena dané časové periody

High - maximální cena dané časové periody

Low - minimální cena dané časové periody

Close - koncová cena dané časové periody

Jako dodatečnou informaci si každá úsečka nese s sebou údaj o datu, čase a množství provedených obchodů v dané periodě (viz. obrázek č. 13).



Obrázek 13: Části záznamu ceny

5.2 NORMALIZACE

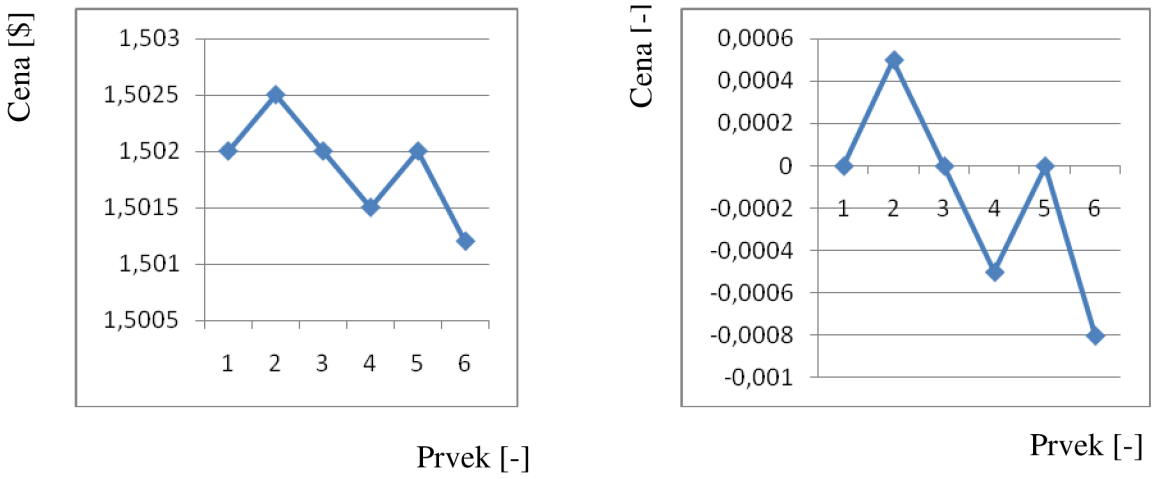
Cílem normalizace je převedení vzoru, který se vyskytne v libovolné cenové hladině do takového tvaru, aby byl srovnatelný s dalším vzorem, který se taktéž vyskytne v libovolné cenové hladině.

5.2.1 Offset normalizace

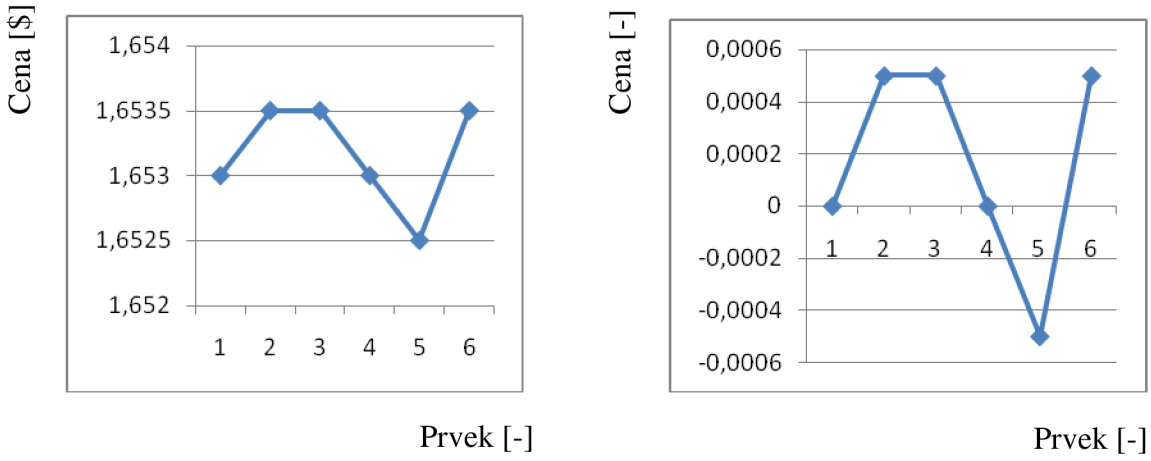
Úprava hypotetických vzorů pomocí metody offset normalizace je výpočetně méně náročnější než metoda min/max normalizace. Základem je určení referenční hodnoty, která bude po offset úpravě představovat nulu. Zvolenou referenční hodnotu odečteme od všech ostatních hodnot vzoru dle rovnice 8. Příklad je uveden na obr.č 14 a obr.č. 15. Jako referenční hodnota je u obr.č.14 volena hodnota 1,5020 a u obr.č.15 je volena hodnota 1,6530. Po této úpravě lze vidět, že na obou obrázcích vpravo jsou vzory ve srovnatelné cenové hladině.

$$offset = aktuální\ hodnota - referenční\ hdonota$$

Rovnice 8: Výpočet offset hodnoty



Obrázek 14: Vzor v hodnotách ceny a vzor v normalizovaném tvaru



Obrázek 15: Vzor v hodnotách ceny a vzor v normalizovaném tvaru

Na příkladu na obr.č. 14 a obr.č.15 je vidět, že dva vzory, vyskytlé v jiné cenové hladině, lze snadno převést do hladiny, ve které lze provést výpočet podobnosti těchto dvou vzorů.

5.2.2 Min/max normalizace

Základní myšlenkou tohoto typu normalizace je převedení do intervalu o určité velikosti, takže maximální hodnota vzoru v nenormalizovaném tvaru bude reprezentovat maximální hodnotu normalizovaného tvaru a minimální hodnota bude mít hodnotu minimální v normalizovaném tvaru. Výpočet této metody je však časově náročnější, jelikož musíme použít algoritmy pro výpočet maximální a minimální hodnoty v datech, která chceme normalizovat.

Obecný tvar lineární transformace můžeme vyjádřit rovnicí 9, kde „a“ je koeficient zmenšení (zvětšení), „x“ je transformovaná hodnota a „b“ je velikost posuvu nové hodnoty.

$$x_i' = ax_i + b$$

Rovnice 9: Obecný tvar lineární rovnice [17]

Nejčastější úprava je transformace dat do intervalu <0;1>, pro kterou můžeme použít výpočet rovnicí 10, kde funkce min() a max() vrací hodnotu nejmenšího a největšího prvku dat určených k normalizaci.

$$x_i' = \frac{x - min(x_1..x_i)}{max(x_1..x_i) - min(x_1..x_i)}$$

Rovnice 10: Výpočet normalizovaných hodnot v rozmezí <0;1> [17]

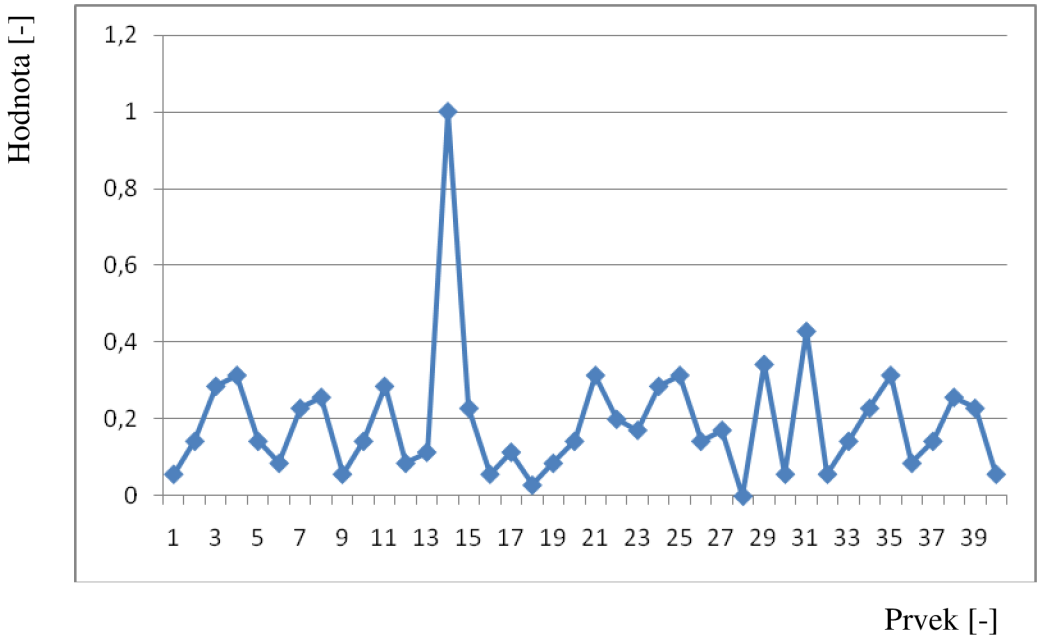
Převod z intervalu <0;1> do jakéhokoli jiného lze provést rozšíření předchozí rovnice na upravenou rovnici 11, kde O_{max} a O_{min} jsou hranice výstupního intervalu.

$$x_i' = \frac{x_i - min(x_1..x_i)}{max(x_1..x_i) - min(x_1..x_i)} \times (O_{max} - O_{min}) + O_{min}$$

Rovnice 11: Výpočet normalizovaných hodnot v libovolném rozmezí [17]

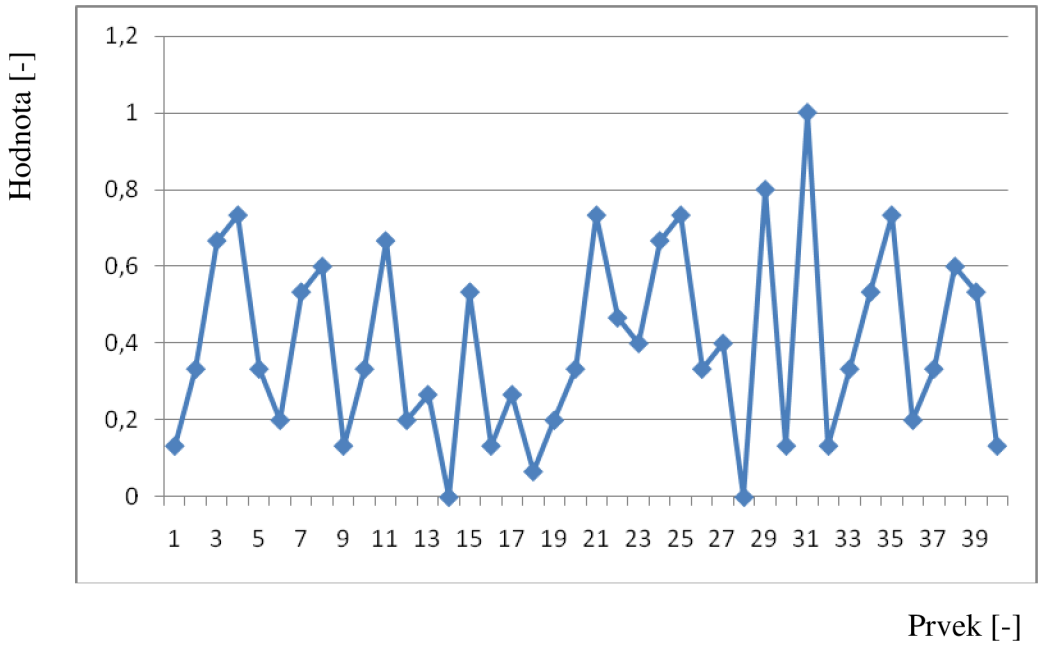
Nevýhoda lineární normalizace je v tom, že když vstupní data obsahují výrazně větší nebo výrazně menší hodnoty než je velikost střední hodnoty zbývajících dat, dojde k nevyužití velké většiny rozsahu.

Tyto výkyvy v datech jsou důsledkem anomálií, špiček či chyb. Ovlivnění využití rozsahu je zobrazeno na obr.č. 16.



Obrázek 16: Normalizace typu min/max se špičkovou hodnotou

Na obrázku č. 16 jsou kvůli jedné špičkové hodnotě všechny ostatní hodnoty transformovány do 60% výstupního rozsahu. Vynecháme-li tuto hodnotu, je rozsah využit mnohonásobně více, viz. obr.č. 17.



Obrázek 17: Normalizace typu min/max bez výskytu špičkové hodnoty

5.3 FORMÁT VSTUPNÍCH DAT PRO STATISTICKOU ANALÝZU

Možnosti transformací vstupních dat existují tisíce. Základní možnost vstupních dat je cena specifikovaná pomocí výše zmíněných čtyř základních informací o průběhu ceny v dané periodě (Open, High, Low, Close).

Druhá možnost je využití tzv. „typické“ ceny, kdy danou časovou periodu reprezentuje jedna číselná hodnota vypočtena z rovnice 12.

$$Typical\ price = \frac{High + Low + Close}{3}$$

Rovnice 12: Výpočet hodnoty ceny typu „Typical price“

Jako třetí možnost je implementováno použití vstupních dat pro rozeznávání vzorů pomocí transformací, které se nazývají „technické indikátory“. V aplikaci je zpřístupněná možnost využít kombinací 5-ti až indikátorů jako vstupních dat současně a pomocí vytvořeného programu najít mezi nimi vzory, které budou vykazovat predikční schopnosti. Pro výpočet těchto indikátorů je použito knihovny „ta_libc.h“, jejíž funkce je popsána níže. Na obr.č. 18 jsou graficky zobrazeny všechny tři druhy možných vstupních dat.

Obrázek 18: Typy vstupních dat [15]

5.4 KNIHOVNA „TA_LIBC.H“

Knihovna „ta_libc.h“ je volně šiřitelná a obsahuje speciální výpočty, indikátory, pro transformaci ceny z Open, High, Low a Close informace na určitou hodnotu. Většinou se však pro jednotlivé výpočty používá pouze cena koncová – Close. Knihovna obsahuje 150 základních nástrojů technické analýzy od sofistikovanějších a ne příliš známých (Polychromatic Momentum, RAVI TrendIndicator …) až po indikátory známé a využívané v různorodých odvětvích (klouzavé průměry, CCI …).

Práce s touto knihovnou ušetří mnoho času strávených nad tvorbou indikátorů a jejich ladění.

Příklad výpočtu jednoduchého klouzavého průměru:

```
TA_MA( 0, 399,&closePrice[0],30,TA_MAType_SMA,&outBeg, &outNbElement, &out[0] )
```

Popis parametrů funkce:

0 – ukazatel na počátek v datech, od kdy se má začít počítat
399 – ukazatel na konec v datech
&closePrice[0] – ukazatel na vstupní data pro výpočet
30 – perioda klouzavého průměru
TA_MAType_SMA – typ klouzavého průměru (jednoduchý klouzavý průměr)
&outBeg – ukazatel na místo zápisu do výstupních dat
&outNbElement – počet prvku k vypočtení indikátoru
&out[0] – ukazatel na výstupní soubor po provedení výpočtu

5.5 ZÁKLADNÍ ČÁSTI PROGRAMU

Pro komunikaci programu s koncovým uživatelem je navrženo ovládací prostředí, které umožňuje nastavení parametrů a specifikací procesu nalezení smysluplných vzorů a testování nalezených vzorů.

5.5.1 Načtení dat – záložka Input data

První část programu slouží k načtení dat, která budou použita jako vstupní data pro proces učení. Jelikož programovací prostředí c++ builder nemá implementovanou funkci pro načtení dat, která je potřebná, je nutné vytvořit algoritmus pro tento proces.

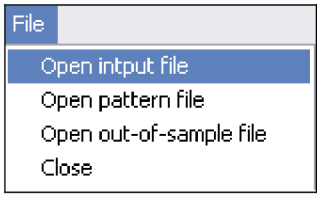
Jeden záznam průběhu ceny s sebou nese několik informací. Jsou to informace o datu, čase, otevírací ceně, maximální ceně, minimální ceně, zavírací ceně a množství provedených obchodů viz. tabulka 1.

Datum	Čas	Otevírací cena[$]	Maximální cena[$]	Minimální cena[$]	Zavírací cena[$]	Počet obchodů[-]
5.6.2008	16:05	105.950	106.000	105.930	105.980	128
5.6.2008	16:10	105.990	106.040	105.890	105.900	134
5.6.2008	16:15	105.910	106.040	105.890	106.030	169
5.6.2008	16:20	106.040	106.040	105.960	106.020	112
5.6.2008	16:25	106.010	106.050	105.990	106.040	98

Tabulka 1: Informace o záznamu ceny pro definovaný časový úsek

Pro proces vyhledávání vzorů nejsou důležité všechny informace. Důležitá je pouze informace o čase a cenách. Informace o čase je podstatná z hlediska filtrování vyhledávání vzorů na čase. Cenové informace slouží k hlavním výpočtům.

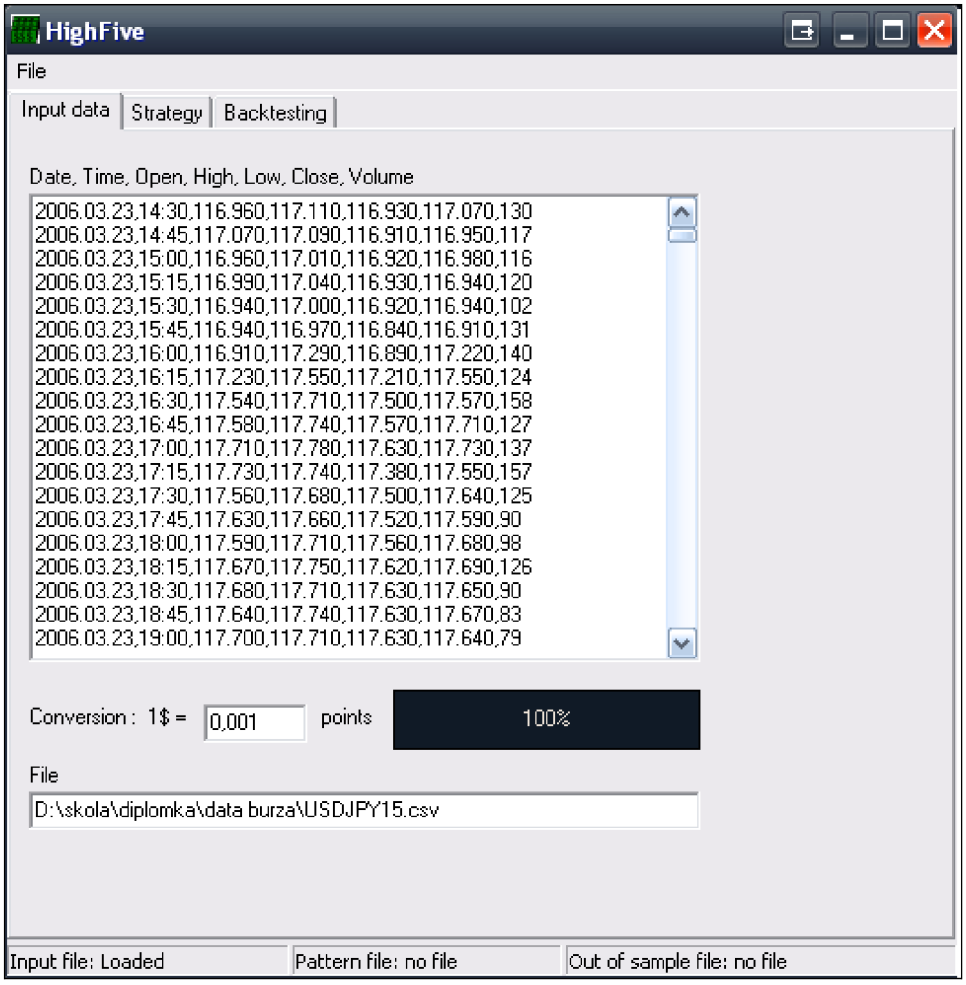
Obr.č. 20 zobrazuje grafické prostředí části programu sloužící k načítaní dat. Výběr dat pro načtení se provede pomocí záložky File->Open input file , viz. obr.č. 19.



Obrázek 19: Menu pro načtení vstupních dat

Po výběru dojde automaticky k načtení dat do programu. Signalizace správného načtení je realizována pomocí stavového řádku, kde je signalizován stav: „Input file: Loaded“. V této části programu je klíčové nastavení hodnoty „Conversion“, která dá programu informaci o velikosti pohybu, který odpovídá 1$.

Záznam dat v tabulce 1 je pořízen sledováním měnového páru USD/JPY(vztah amerického dolaru vůči japonskému jenu). Je patrné, že minimální pohyb je řádově 0,01. Tento pohyb odpovídá hodnotě 10$, čili hodnota 1$ odpovídá 0,001 bodu pohybu. Pro pár EUR/USD(vztah eura vůči americkému dolaru) je hodnota rovna 0,00001.



Obrázek 20: Záložka Input data

5.5.2 Hledání vzorů – záložka Strategy

Druhá obrazovka slouží k nastavení specifikací pro vyhledávání vzorů. Zde je vysoký stupeň volnosti nastavení.

Strategy: volba typu strategie mezi long (nákup) a short (prodej).

Data source: volba typu vstupních dat. Při volbě „OHLC price“ jsou jako vstupní data volena Open, High, Low a Close ceny jednotlivých úseček. Zde je implementován určitý stupeň volnosti a to funkcí váhování v položce „OHLC weigh setting“, kde je možno přidělit jednotlivé části ceny určitou váhu.

Při volbě „Typical Price“ jsou vstupní data transformována funkcí, která je popsána výše. Volba „Indicators“ umožňuje jako vstupní data využít až pěti kombinací technických indikátorů, viz. výše.

Pattern types: volba typu použité normalizace. Doporučeno je použít normalizovaný tvar.

Search pattern from: slouží k nastavení délky hledaného vzoru po určitém kroku.

Use Step: krok délky hypotetického vzoru při procesu učení.

Error methods: volba mezi dvěma možnostmi pro výpočet chyby, která vyjadřuje míru podobnosti vzorů. Vysvětlení rozdílu mezi těmito metodami bude vysvětleno níže. Pří volbě „2nd method“ je důležité nastavení hodnoty „Error range“, která definuje míru podobnosti.

Error range: vyjadřuje míru podobnosti vzorů při užití druhé metody výpočtu chyby. Zadává se v rozsahu 0-1, kdy tyto hodnoty vyjadřují mez podobnosti 0-100 %

Generate data to file: povolí nebo zakáže záznam jednotlivých průběhů výpočtů.

Commisions: vyjadřuje poplatek za jeden provedený obchod. Je to klíčový prvek při vyhodnocování. Průměrná hodnota poplatku je 5$.

Profit Target: velikost potencionálního zisku, který chceme predikovat.

Stop Loss: maximální definovaná ztráta při vstupu do pozice, čili půjde-li pohyb proti nám, dojde k automatickému výstupu z pozice na přesně definované hodnotě.

# of bars: maximální délka požadované situace (délka trvání obchodu). Použijeme-li 5-ti minutový časový záznam, hodnota maximální délky trvání obchodu o velikost 12 bude znamenat, že maximální doba setrvání v obchodu bude jednu hodinu, po dosažení této hodnoty dojde k neprodlenému výstupu z pozice.

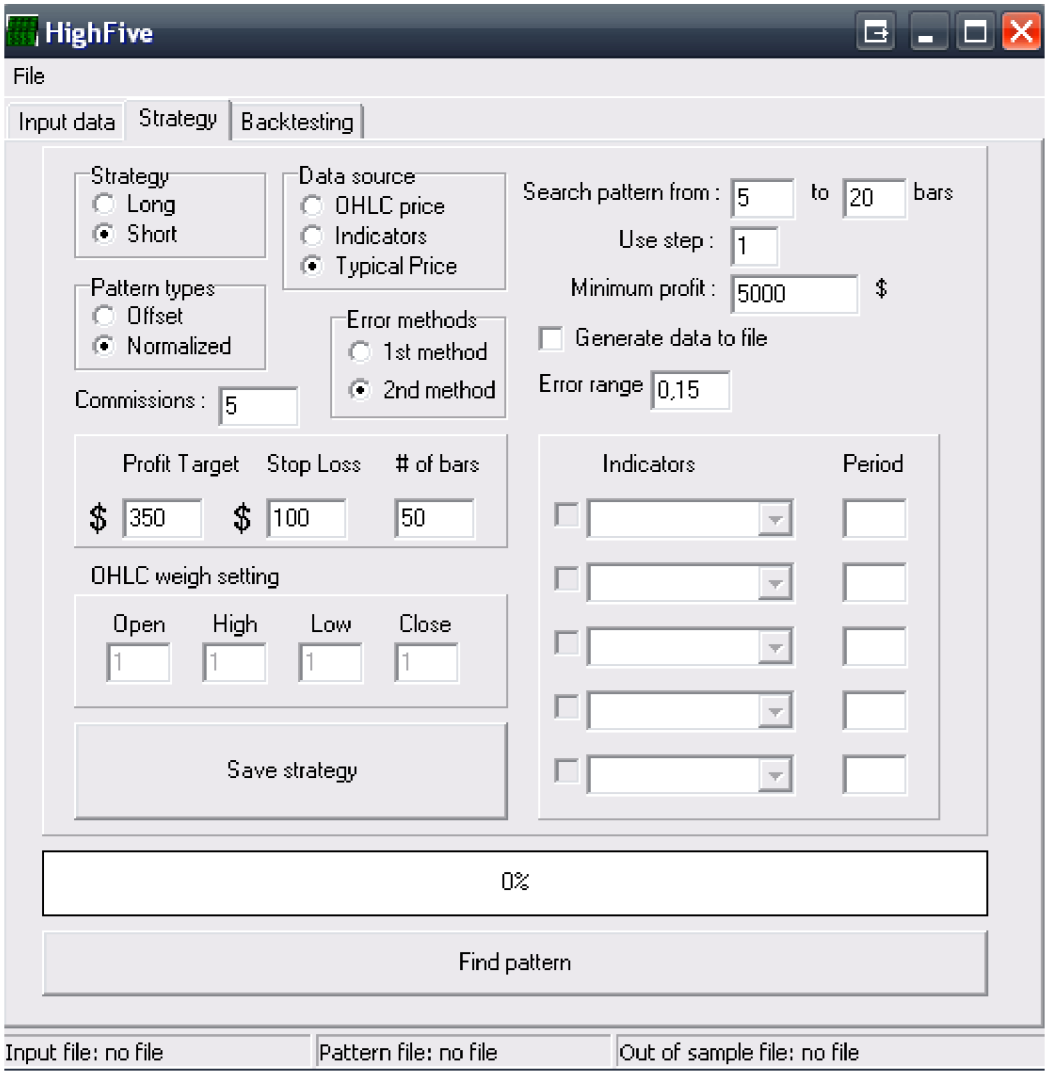
Definice „Profit Target“, „Stop Loss“ a „# of bars” je ve své podstatě přesná definice situace v časové řadě, kterou se budeme pokoušet predikovat.

Minimum profit: vyjadřuje hodnotu profitu, při které začíná být vzor statisticky zajímavý.

Indicators: umožňuje nastavení transformací ceny pro vstupní data při volbě „Data source->Indicators“. Pro jednoduchost jsou zpřístupněny pouze dva typy. Klouzavé průměry a indikátor CCI(speciální transformace z ceny). U každého indikátoru je třeba volit jeho periodu.

Po nastavení všech výše vyjmenovaných parametrů je třeba vše uložit tlačítkem „Save strategy“. Tento proces je třeba udělat pokaždé při změně jakéhokoliv nastavení strategie. Poslední krok je spuštění procesu vyhledávání vzorů pomocí tlačítka „Find pattern“. Poté začne samotný proces učení. Jde o výpočetně náročný proces, proto je třeba počítat s časovou prodlevou řádově desítky minut.

Za běhu je generován soubor „patterns.txt“, který je implicitně vytvořen v adresáři, odkud se načítají vstupní data. Soubor obsahuje záznam nejdůležitějších parametrů procesu. Pro urychlení výpočtů je lepší vypnout volbu „Generate data to file“.

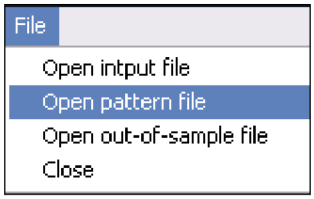


Obrázek 21: Záložka Strategy

5.5.3 Ověření metody – záložka backtesting

Proces ověření funkčnosti nalezených vzorů je jedna z nejdůležitějších částí celého procesu. Dochází zde k aplikaci nalezených vzorů v učících datech na datech nových a dochází k vyhodnocení úspěšnosti.

Vstupem jsou dva soubory. První soubor je vygenerovaný předešlou části programu, textový soubor s názvem „patterns.txt“ a druhý soubor je databáze dat, na kterých bude proveden test. Načtení těchto souborů se provádí pomocí menu File viz. Obr.č. 22. Je nutné jako první načíst soubor s testovacími daty a až potom soubor „patterns.txt“.



Obrázek 22: Menu pro načtení vstupních dat

Nastavení tohoto procesu obsahuje určitý stupeň volnosti nastavení.

Strategy: se volí typ vstupu do obchodu (nákup nebo prodej). Volba typu strategie by měla být totožná s volbou strategie při procesu učení.

Trade setup: umožňuje volbu ignorování nalezeného vzoru, je-li aktuálně již jedna pozice otevřena (Ignore pattern in trade) nebo naopak povoluje vstup do trhu při právě probíhajícím obchodu (Open new position in trade).

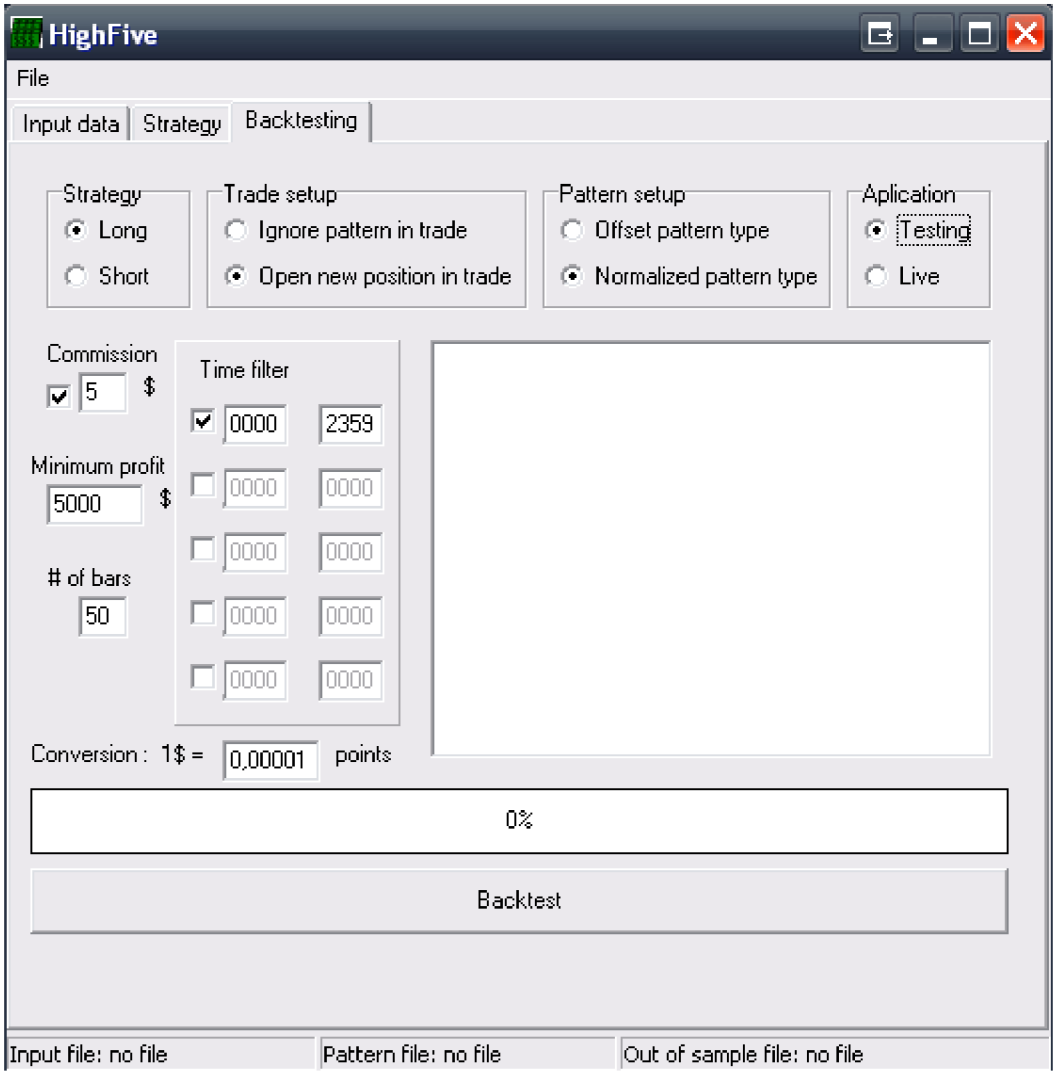
Pattern setup: umožňuje volbu typu normalizace. Je nutné volit stejný typ normalizace jako při procesu učení.

Jelikož lze program použít i jako pouze testovací, bez nutnosti předešlého procesu vyhledávání vzorů, je nutno znova nastavit parametry „Commision“,“# of bars“, „Conversion“ .

Time filter: slouží k nastavení časových filtrů, kdy bude program testovat pouze definované časové intervaly.

Application: slouží k definování, zda-li šlo o aplikaci vzorů v testovacích datech nebo na reálném trhu.

Po ukončení testování se zobrazí výsledky v textovém boxu záložky „ Backtesting“.



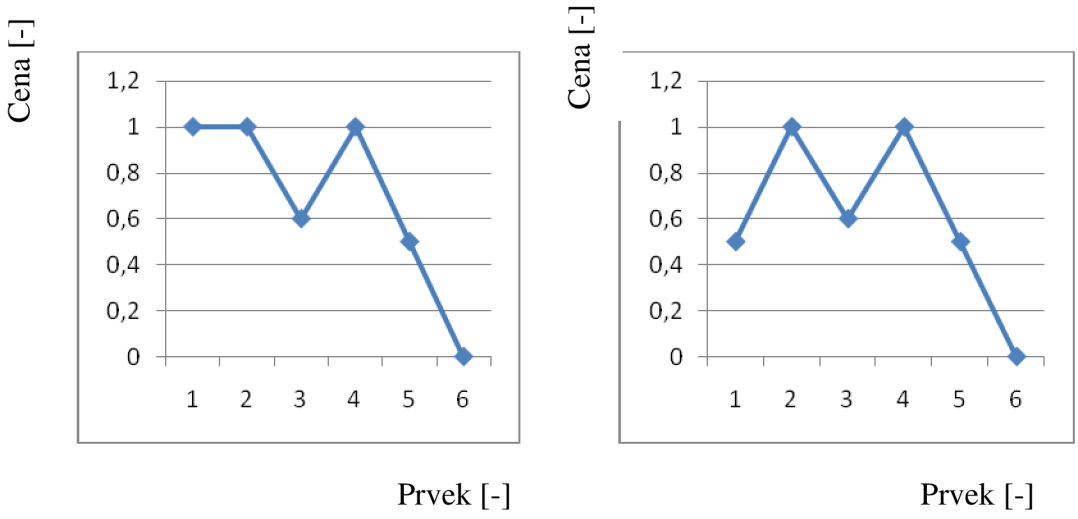
Obrázek 23: Záložka Backtest

5.6 METODY VÝPOČTŮ CHYBOVÝCH FUNKCÍ

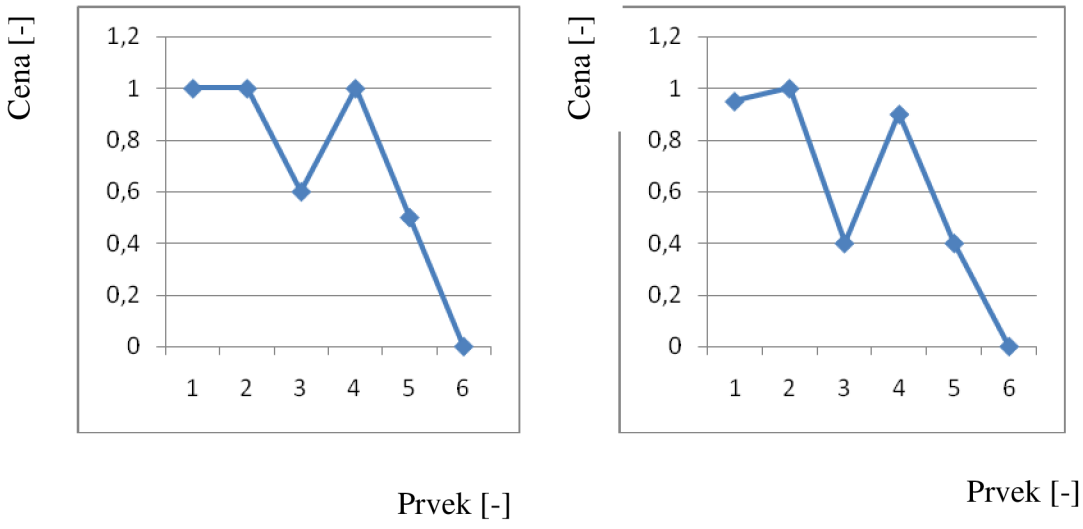
Pro porovnávání míry podobnosti dvou vzorů je využito dvou metod. První metoda je založena na celkové sumě chyb a druhá metoda je založena na splnění definované meze odlišnosti dílčích prvků vzorů.

5.6.1 První metoda výpočtu chyby

Výpočet míry podobnosti u této metody je založen na součtu všech odchylek vzoru od porovnávané situace. Nezáleží na rozdílu jednotlivých částí, ale v úvahu se bere až suma všech chyb. Tuto metodu chyby lze aplikovat u všech typů vstupních dat i zvoleného typu normalizace.

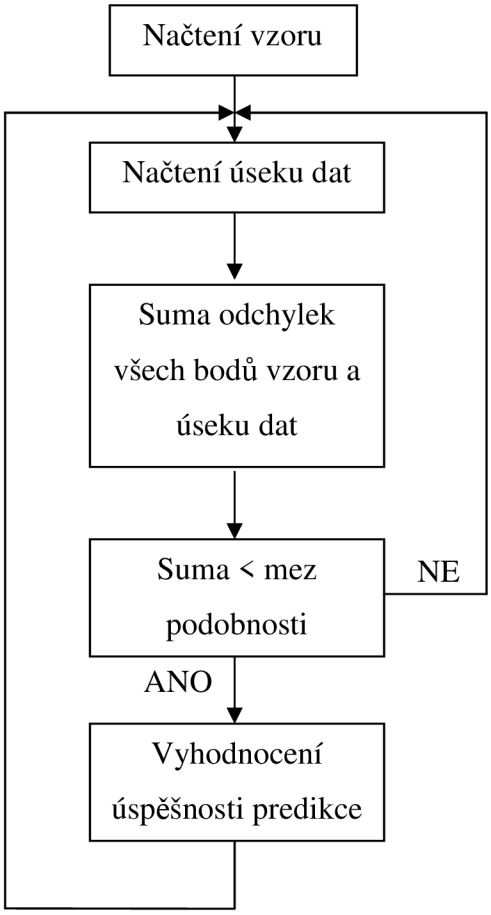


Obrázek 24: Dva rozdílné vzory k porovnání první metodou



Obrázek 25: Dva rozdílné vzory k porovnání první metodou

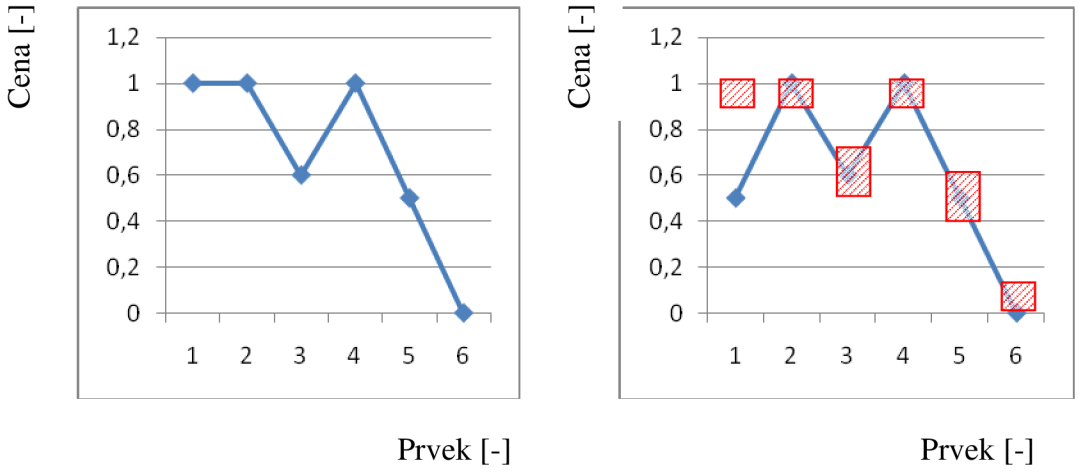
Velikost chyby na obou obrázcích dosahuje stejné hodnoty. U prvního případu je však velký rozdíl pouze u první hodnoty a zbytek vzoru je zcela stejný. Celková odchylka je tedy dána pouze první odchylkou. U druhého případu se liší téměř všechny hodnoty, avšak ne o tak výraznou hodnotu, jako u prvního případu. Součet všech odchylek dává přibližně stejnou hodnotu. U prvního případu lze rozpoznat, že jde už o velmi odlišný vzor a naopak druhý případ je výrazně podobnější, avšak oba vykazují stejnou míru podobnosti. Rozhodovací proces klasifikace je zobrazen na obr. č. 26.



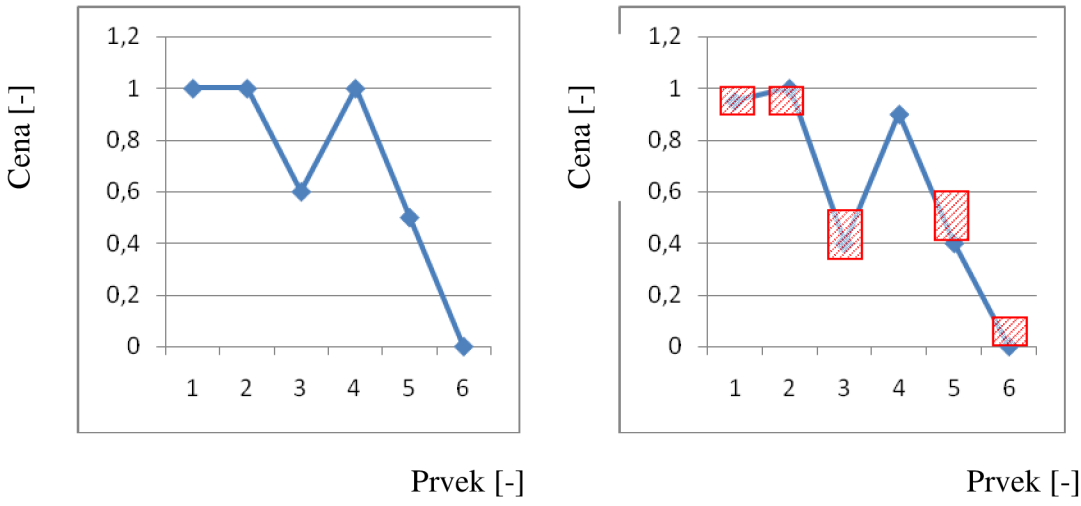
Obrázek 26: Algoritmus výpočtu chyby u první metody

5.6.2 Druhá metoda výpočtu chyby

Druhá metoda není založena na sumě všech hodnot, ale na míře podobnosti jednotlivých hodnot vzoru od nových dat. Tato metoda výpočtu chyby je zpřístupněna pouze u volby vstupních dat „Typical Price“ a typ normalizace min/max. Nejdůležitější hodnotou je „Error range“, kterou uživatel definuje maximální mez odlišnosti jednotlivých hodnot vzoru a dat. Hodnota 0,1 představuje odlišnost 10%. Na obr.č. 27 a 28 je zobrazen princip této metody a na obr.č. 29 je rozhodovací algoritmus.



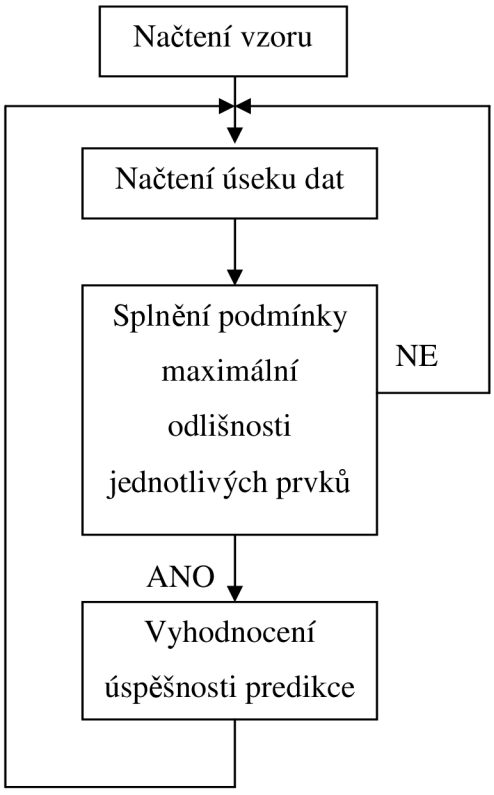
Obrázek 27: Dva rozdílné vzory k porovnání druhou metodou



Obrázek 28: Dva rozdílné vzory k porovnání druhou metodou

Při volbě maximální odchylky 0,20, čili 20% první případ na obr.č. 27 nesplňuje tuto podmínku již v prvním bodě. Ostatní body mají sice 100% podobnost, ale vzor je vyřazen hned po prvním porovnání a dále se nebere jako potencionálně zajímavý. U druhého případu na obr.č. 28 je více hodnot odlišných od referenčního vzoru, ale všechny splňují toleranci 20% a tento vzor je klasifikován jako dostatečně podobný. Červené plochy na obrázcích definují maximální odlišnost.

Algoritmus je demonstrován na obr.č. 29



Obrázek 29: Algoritmus výpočtu chyby u druhé metody

5.7 NALEZENÍ VZORU VYKAZUJÍCÍHO PREDIKČNÍ CHOPNOSTI

Výstupem procesu tohoto programu je soubor informací, obsahující seznam vzorů, které prošly procesem učení a ověřování a byly shledány významné v oblasti predikce.

5.7.1 Nalezení u první metody výpočtu chyby

Pro každý hypotetický vzor, tedy data o definované velikosti před definovanou strategií, je vypočtena chybová funkce. Hypotetický vzor se porovná se všemi daty a vygenerují se hodnoty pro každé porovnání. Tyto dvě hodnoty nesou informaci o tom, jak velká byla míra podobnosti a zda-li u tohoto případu došlo k úspěšné predikci či naopak. Příklad této dvojice je uveden v tabulce č.2.

Porovnání	Chyba[-]	Predikce[$]
1	0,15	1000
2	0,6	-300
3	0,59	-300
4	0,1	1000
5	0,9	-300
6	1,5	-300
7	0,6	-300
8	0,3	1000
9	0,12	-300
10	2	-300
11	0,3	-300
12	0,5	-300

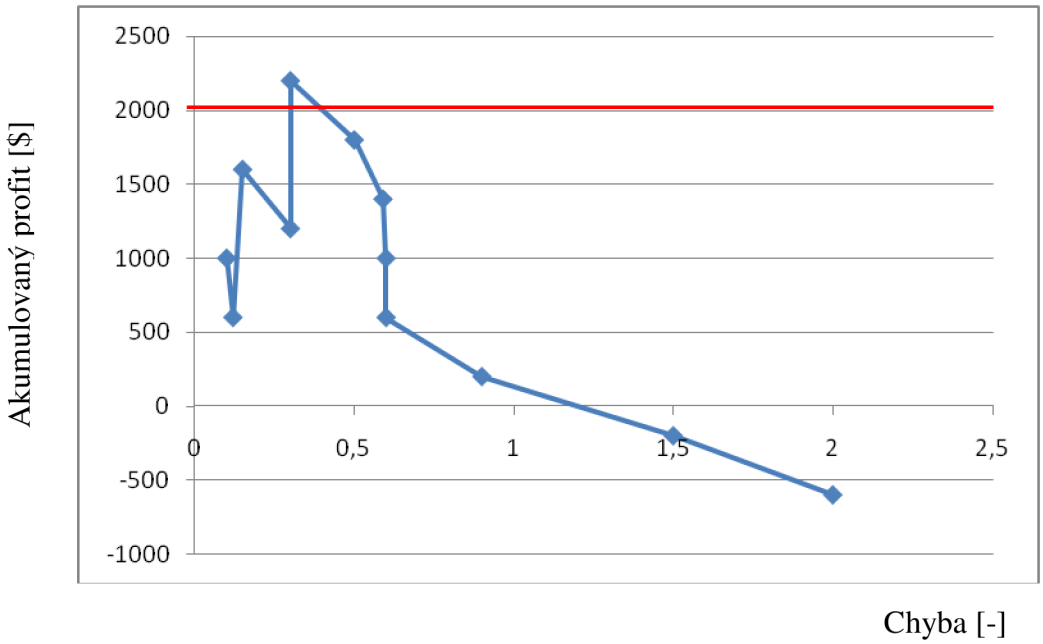
Tabulka 2: Velikost chyby a úspěšnost predikce

Hodnota -300 ve sloupci Predikce znamená, že nedošlo ke správné predikci a hodnota 1000, znamená, že predikce proběhla úspěšně. Pro stanovení optimální míry podobnosti je třeba provést seřazení těchto hodnot podle velikosti chyby a provést sumu úspěšných a neúspěšných vzorů viz. tabulka 3.

Porovnání	Chyba[-]	Predikce[$]	Suma[$]
4	0,1	1000	1000
9	0,12	-400	600
1	0,15	1000	1600
8	0,3	-400	1200
11	0,3	1000	2200
12	0,5	-400	1800
3	0,59	-400	1400
2	0,6	-400	1000
7	0,6	-400	600
5	0,9	-400	200
6	1,5	-400	-200
10	2	-400	-600

Tabulka 3: Seřazení porovnávání dle chyby od největší po nejmenší

Graf sumy v závislosti na odchylce je zobrazen na obr.č. 30 a z tohoto grafu lze snadno odvodit optimální mez podobnosti.



Obrázek 30: Graf závislosti velikosti chyby na sumě predikce

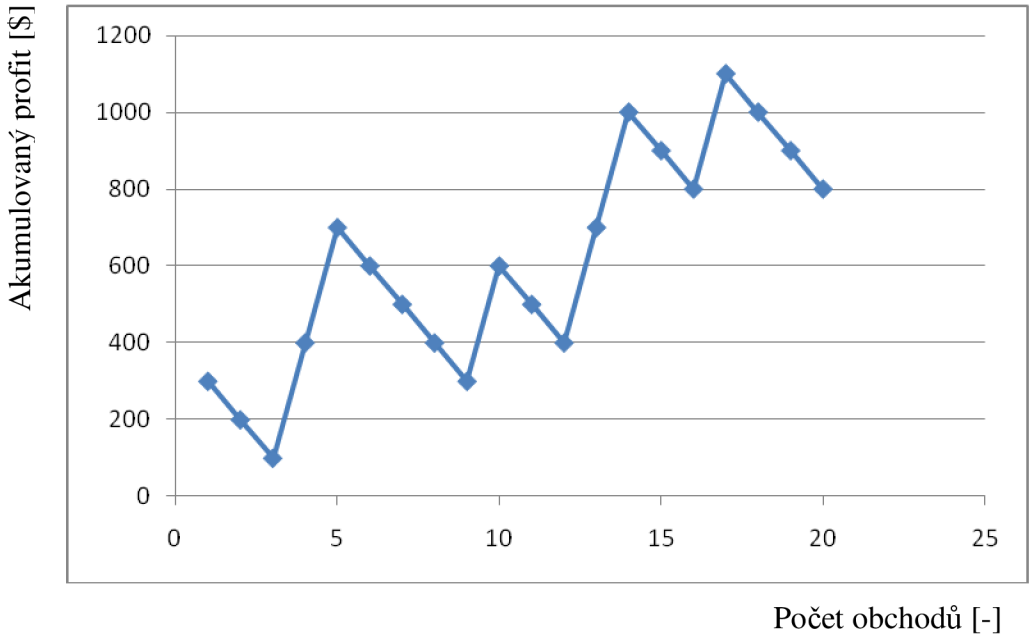
Z grafu je patrné, že maximálního naakumulovaného profitu bylo dosaženo při odchylce 0,3. Tato hodnota se volí jako maximální mez podobnosti a je to informace, která se, zároveň s normalizovanou podobou vzoru, přenáší do ověřovacího procesu. Jelikož nastane hodně situací, kdy dojde k nízkému naakumulovanému profitu a daný vzor není nijak výrazně zajímavý, je zavedena proměnná „Minimum profit“, která definuje hranici, od které začíná být vzor zajímavý z hlediska výkonnosti predikce. Zvolili bychom tuto hodnotu v tomto případě rovnou 3000$, tento vzor by neprošel filtrem a nebyl by brán jako potencionálně zajímavý. Volili bychom hodnotu 2000$, vzor se zaznamená do souboru patterns.txt a může být dále otestován.

5.7.2 Nalezení u druhé metody výpočtu chyby

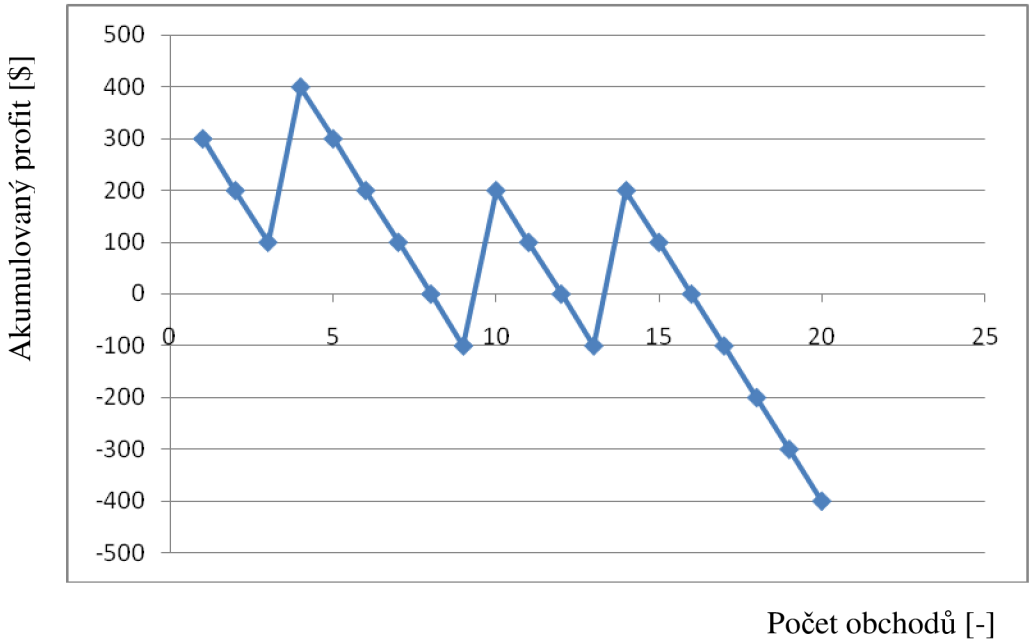
Princip této metody není založen na automatickém vyhledání optimální chyby, ale uživatel volí sám mez podobnosti pro každý porovnávaný bod vzoru hodnotou proměnnou „Error range“. Splní-li všechny hodnoty porovnávaného vzoru tuto podmínku, dojde k zaznamenání úspěšnosti či neúspěšnosti predikce.

Úspěšný vzor k zaznamenání do souboru je ten, který splní podmínku většího naakumulovaného profitu, než je právě definována hodnota „Minimum profit“.

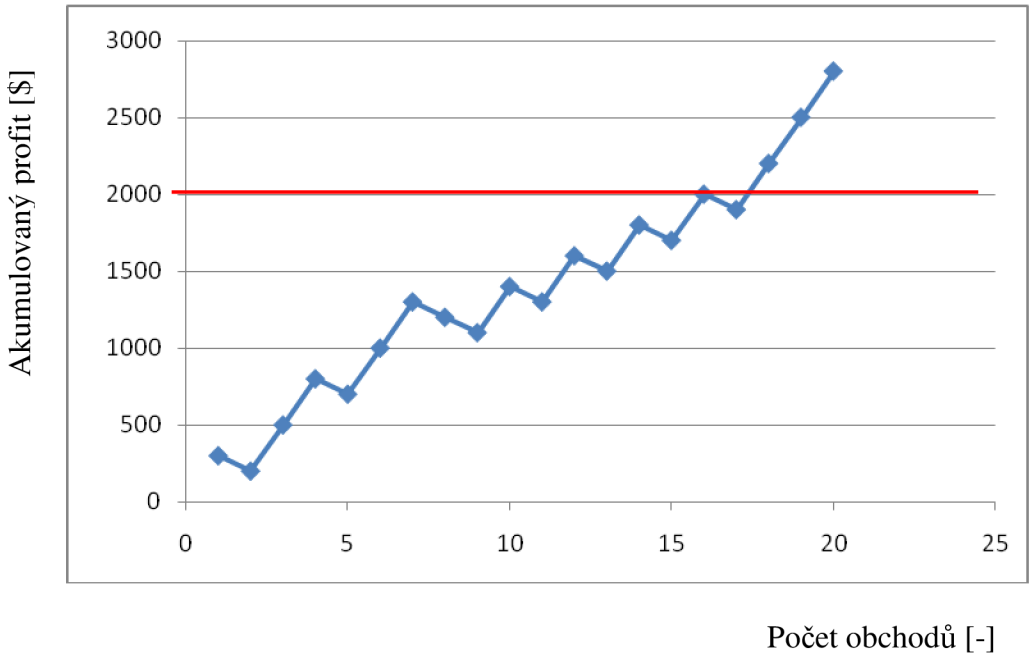
Příklad: Hodnota minimálního naakumulovaného profitu je 2000$. Vzor na obr.č.31 nesplňuje podmínku. Vykazuje sice kladné hodnoty akumulovaného profitu, ale ne dostatečně velké. Obrázek č.32 zobrazuje vzor, který nemá ani kladné hodnoty, až obrázek č. 33 představuje vzor, který splnil podmínku druhého typu výpočtu chyby a je proveden jeho záznam do soubor patterns.txt pro další ověřování.



Obrázek 31: Závislost počtu obchodů na akumulovaném profitu



Obrázek 32: Závislost počtu obchodů na akumulovaném profitu



Obrázek 33: Závislost počtu obchodů na akumulovaném profitu

5.8 DOSAŽENÉ VÝSLEDKY

K ověření predikčních schopností vytvořené aplikace je třeba provést několikanásobné ověřování. Aplikace byla učena na datech měnového páru EUR\USD v období 2.1.2008 – 30.6.2008 a data mají periodu záznamu 5 minut. Tato data jsou v přiloženém souboru „1st half e_u.csv“. Nastavení parametrů učení bylo následující:

Záložka Input data : Conversion = 0,00001

Záložka Strategy: Strategy = Long

Data source = Typical Price

Pattern types = Normalized

Error methods = 2nd method

Search pattern from = 5 to 15

Use step = 1

Minimum profit = 10000

Generate data to file = false

Error range = 0,15

Commissions = 5

Profit Target = 350

Stop Loss = 100

# of bars = 50

Po spuštění procesu učení s tímto nastavením bylo zjištěno, že v oblasti učících dat existuje 46 vzorů, které splňují výše uvedená kritéria. První ověření výkonnosti je založeno na aplikaci těchto nalezených vzorů v nové oblasti dat. Data pro testování jsou použita v časovém rozmezí 1.7.2008 – 31.12.2008. Ověření bylo provedeno s následujícím nastavením:

Záložka Backtesting: Strategy = Long

Trade setup = Open new position in trade

Pattern setup = Normalized pattern type

Appliaction = Test

Minimum accumulation = 5000

# of bars = 50

Commissions = 5

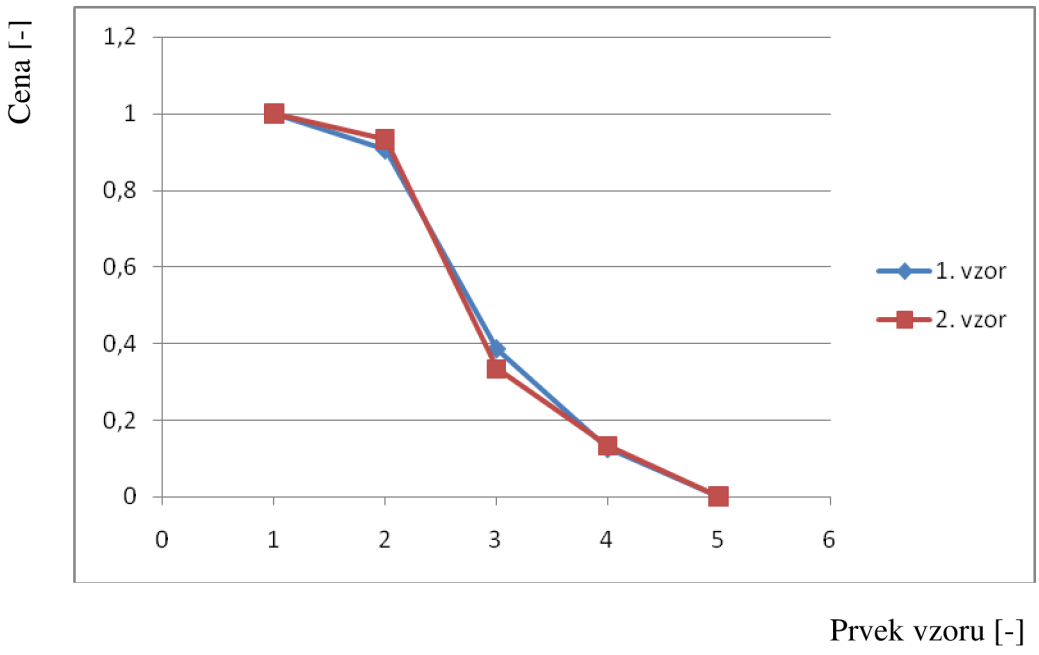
Po spuštění testu s tímto nastavením došlo k ověření výkonnosti vzorů na nových datech a bylo shledáno, že pouze 2 vzory vykazují predikční schopnost dle zvolených požadavků, i mimo oblast učících dat. Tyto vzory jsou aplikovány v reálném prostředí v období 2.1.2009 – 18.5.2009.

Po testu (1.7.2008 - 31.12.2008) byla výkonnost predikce, která je reprezentována akumulací distribuce úspěšných (profit) a neúspěšných (ztráta) obchodů, pro jednotlivé vzory následující:

#	normalizovaný tvar vzoru [-]	výkonnost [$]
1	1; 0,905; 0,368; 0,126; 0	6150
2	1; 0,933; 0,333; 0,133; 0	9420

Tabulka 4: Výsledky a tvar úspěšných vzorů po prvním testu

Z obr.č. 34 je vidět, že oba vzory mají podobný průběh, čemuž odpovídá i úspěšná predikce obou vzorů mimo oblast učících dat.



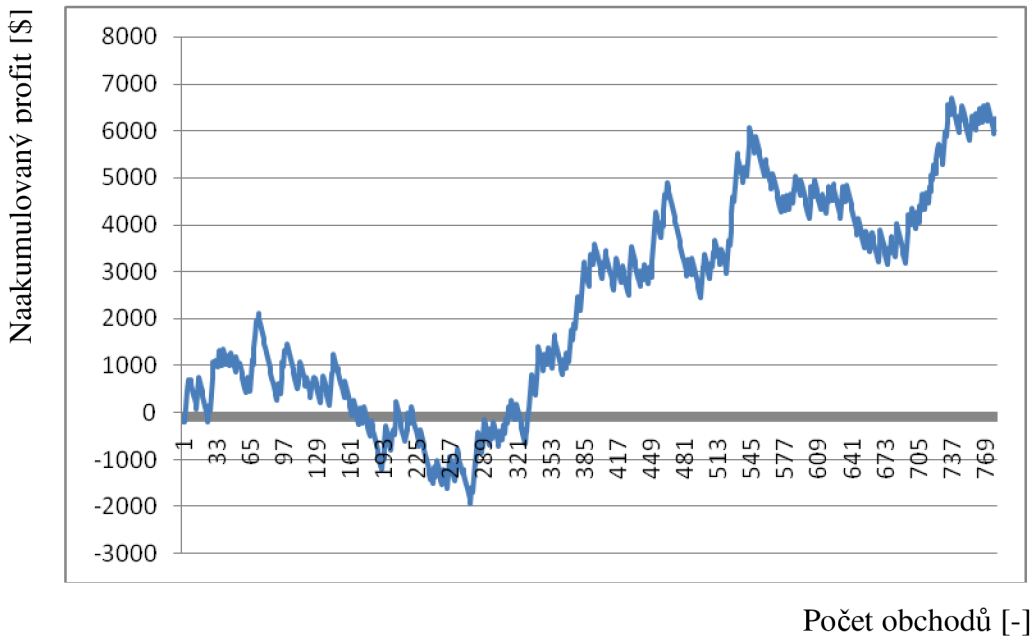
Obrázek 34: Vzor v normalizovaném tvaru

Oba tyto vzory byly aplikovány v reálném prostředí za období 2.1.2009-18.5.2009 a jejich výkonnost v této oblasti dat je vyjádřena v tabulce 5.

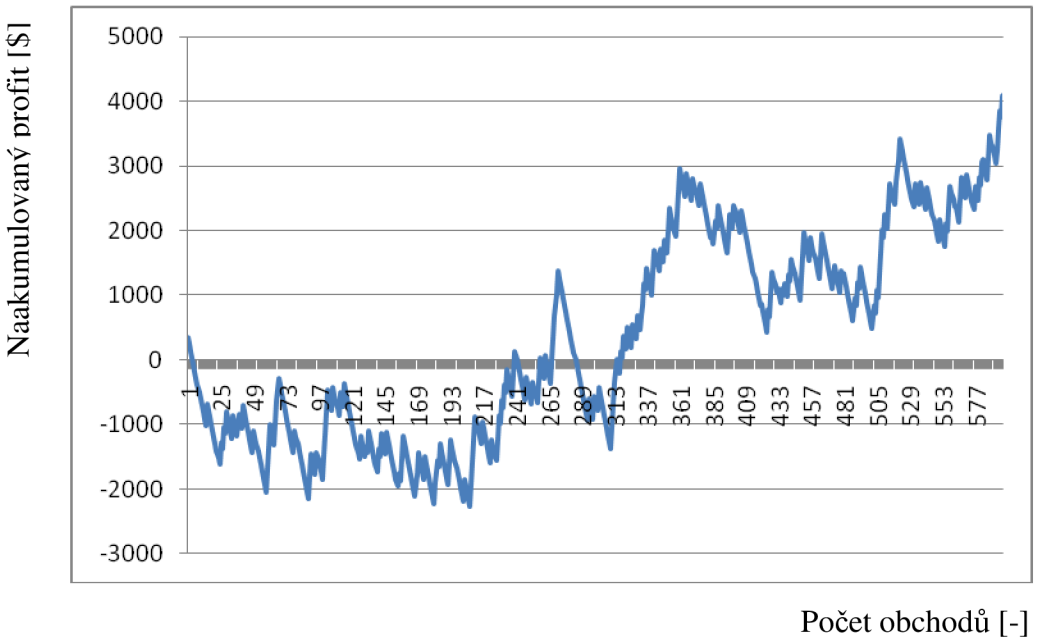
#	normalizovaný tvar vzoru [-]	výkonnost [$]
1	1; 0,905; 0,368; 0,126; 0	4084
2	1; 0,933; 0,333; 0,133; 0	1382,9

Tabulka 5: Výkonnost a tvar vzorů po aplikaci v reálném prostředí

Při porovnání výkonnosti dojdeme k závěru, že dochází k poklesu, ale i přesto vykáže aplikace těchto vzorů na konci období zisk 5466,9$. Na obrázku č. 35 je zobrazen průběh akumulace distribuce profitů a ztrát v průběhu test v druhé polovině roku 2008 a na obrázku č. 36 je zobrazen test na prvním čtvrtletí roku 2009. Obě křivky jsou pro vzor 1.



Obrázek 35: Akumulace distribuce profitů a ztrát za druhé pololetí roku 2008



Obrázek 36: Akumulace distribuce profitů a ztrát za první čtvrtletí roku 2009

Test pro měnový pár USD/JPY:

Záložka Input data : Conversion = 0,001

Záložka Strategy: Strategy = Short

Data source = Typical Price

Pattern types = Normalized

Error methods = 2nd method

Search pattern from = 5 to 15

Use step = 1

Minimum profit = 10000

Generate data to file = false

Error range = 0,15

Commissions = 6

Profit Target = 400

Stop Loss = 120

# of bars = 50

Při tomto nastavení bylo nalezeno 97 vzorů, které splňují tyto požadavky.

Pro proces testování bylo použito nastavení:

Záložka Backtesting: Strategy = Short

Trade setup = Open new position in trade

Pattern setup = Normalized pattern type

Appliaction = Test

Minimum accumulation = 15000

# of bars = 50

Commissions = 6

Time filter = 0000 - 2359

Z původního počtu bylo vybráno 5 vzorů, které splňují požadavky predikčních schopností i v nové oblasti dat. Jejich výkonnost je zobrazena v tabulce 6.

#	normalizovaný tvar vzoru [-]	výkonnost [$]
1	0; 0,15; 0,35; 0,55; 1	16253,9
2	0; 095; 0,405; 0,548; 1	15001,9
3	0; 0,231; 0,4; 0,646; 1	15426
4	0; 0,17; 0,396; 0,66; 1	15913
5	0; 0,094; 0,312; 0,719; 1	15573,9

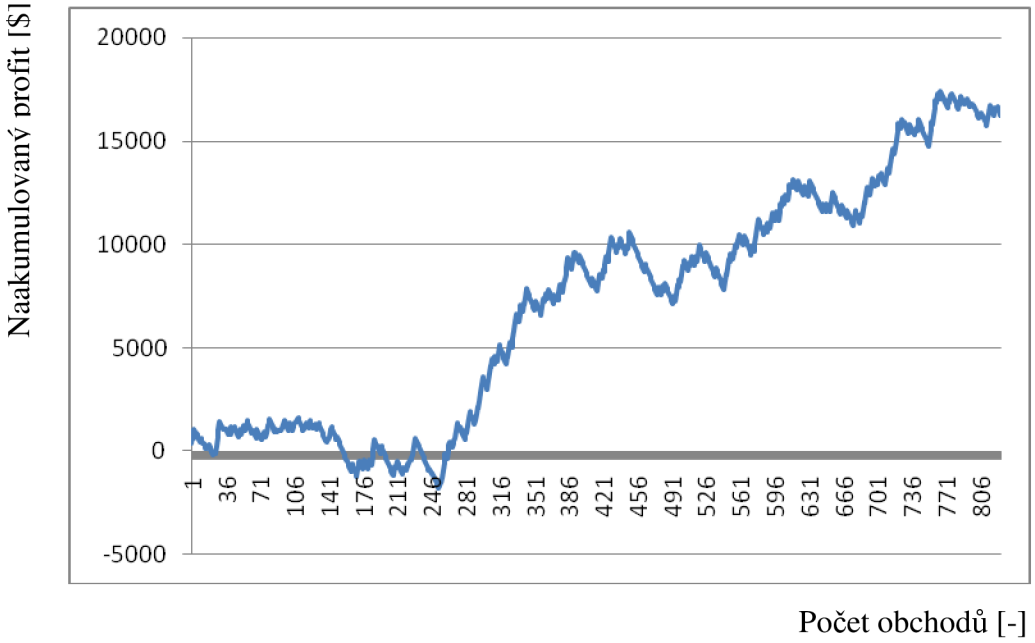
Tabulka 6: Výsledky a tvar úspěšných vzorů po prvním testu

Tyto vzory byly podrobeny testu za první čtvrtletí roku 2009 s výsledky, které jsou zobrazeny v tabulce 7.

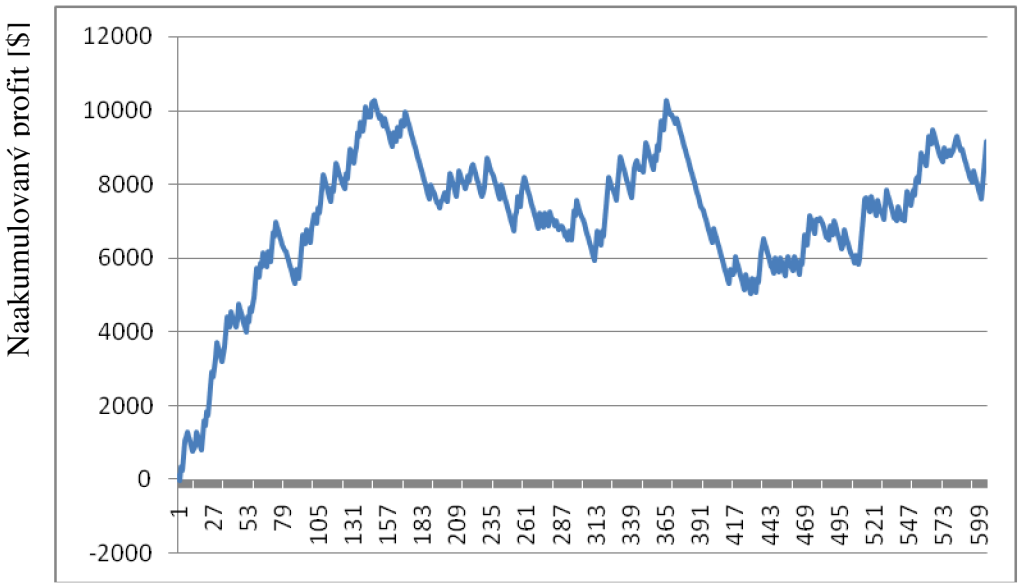
#	normalizovaný tvar vzoru [-]	výkonnost [$]
1	0; 0,15; 0,35; 0,55; 1	9138
2	0; 095; 0,405; 0,548; 1	5436
3	0; 0,231; 0,4; 0,646; 1	6710
4	0; 0,17; 0,396; 0,66; 1	8751
5	0; 0,094; 0,312; 0,719; 1	1115

Tabulka 7: Výkonnost a tvar vzorů po aplikaci v reálném prostředí

Na obrázcích 37 a 38 je zobrazen průběh akumulace distribuce v testovacích a reálných datech. Výsledky jsou pro vzor 1.

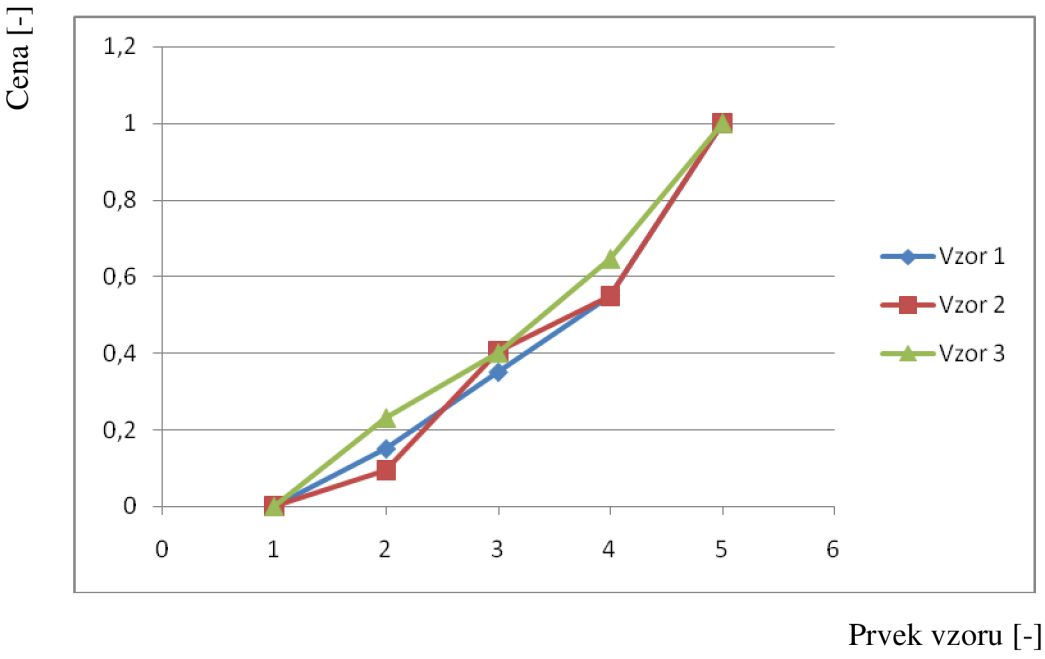


Obrázek 37: Akumulace distribuce profitů a ztrát za druhé pololetí roku 2008

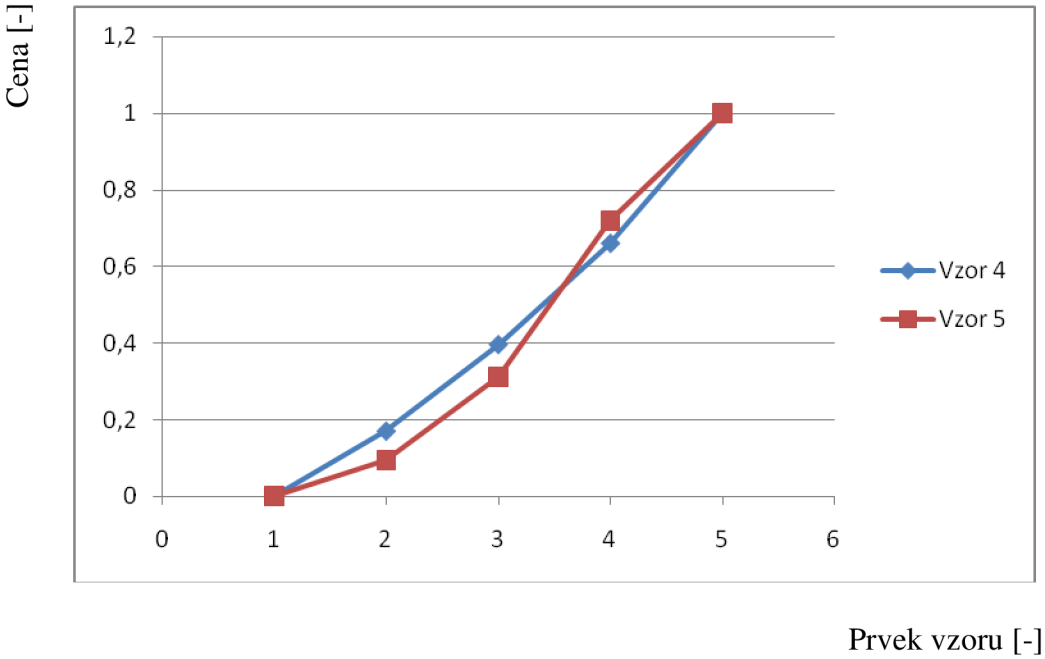


Obrázek 38: Akumulace distribuce profitů a ztrát za první čtvrtletí roku 2009

Na obrázku 39 a 40 je zobrazen normalizovaný tvar všech pěti vzorů. Z obrázků je patrné, že úspěšné vzory mají velmi podobný průběh.



Obrázek 39: Vzor 1, 2 a 3 v normalizovaném tvaru



Obrázek 40: Vzor 3 a 4 v normalizovaném tvaru

5.9 VÝSTUP PROGRAMU

Po procesu učení, který trvá řádově desítky minut při výše uvedeném nastavení, je vygenerován soubor „patterns.txt“, který je implicitně vytvořen v adresáři, odkud jsou načtena vstupní data pro proces učení. Tento soubor obsahuje informace důležité pro otestování vzorů nalezených v učících datech.

Proces testování, který je v záložce „Backtesting“ využívá právě souboru „patterns.txt“, které ověří na nových datech. Po procesu ověření je vygenerován soubor „patterns_test.txt“, který obsahuje informace o vzorech, které úspěšně prošly testem ověřování. Dále jsou vygenerovány jednotlivé průběhy ověřování do souborů ve formátu „BacktestXXX_OOS.txt“, kde hodnoty XXX nesou informaci o pozici nalezeného vzoru.

Výstupem programu je série vzorů, které vykazují predikční schopnosti. Podoba těchto vzorů v normalizovaném tvaru i jejich výkonnost je zobrazena v informačním poli v záložce „Backtesting“.

5.10 DOPORUČENÉ NASTAVENÍ

Ekonomické systémy vykazují velkou dávku chaotického chování a nestability, což znamená, že predikovat jakýmkoli způsobem vývoj trhu není snadný proces. Pro dobré výsledky doporučuji použít nastavení parametrů tak, jako ve výše uvedených příkladech.

Jelikož je predikce vývoje těchto systémů velmi náročná, je pravděpodobné, že při jiném nastavení vstupních hodnot parametrů nedojde k nalezení vzorů s predikční schopností. Nastavení jednotlivých parametrů procesu učení vyžaduje znalost uživatele problematiky obchodování na elektronických trzích.

Jako příklad lze uvést nastavení stejných parametrů procesu učení, kromě hodnoty „Strategy“, která byla zvolena „Short“. Systém, při této jediné změně parametru, nedokáže nalézt žádný vzor vykazující predikční schopnosti mimo oblast učících dat.

Velká obsáhlost vstupních dat znesnadňuje orientaci a ověřování funkčnosti, proto jsou na přiloženém CD dva soubory, které obsahují přijatelné množství dat pro

orientaci. Jde o soubory „test short.csv“ a „test long.csv“. Průběhy vykazují pohyb časové řady jedním směrem a průběhy jsou zobrazeny na obrázcích „test short.jpg“ a „test long.jpg“.

Struktura adresářů na CD je založena na měnovém páru, typu zvolené strategie a dále na jednotlivé části procesu (learning, test, live). V procesu učení otevřeme složku „learning“, ve které jsou historická data pro proces učení. Po tomto procesu se v adresáři „learning“ vytvoří soubor „pattern.txt“. Tento soubor překopírujeme do složky test, ve které je připravena druhá databáze dat, která bude sloužit pro testování. V aplikaci v záložce „ Backtesting“ je nutné u této fáze nastavit v poli „Application“ položku „Test“. Po procesu testování je v adresáři vygenerován soubor „patterns_test.txt“, který obsahuje pouze vzory, které splnily definované požadavky na predikční výkonnost. Chceme-li provést další test, zkopírujeme tento soubor do složky „live“, kde je připravena další sada dat. Soubor „patterns_test.txt“ je nutné přejmenovat, například na „patterns_t.txt“ a provedeme test. Zde je důležité nastavení položky „Application“, kterou je třeba nastavit na „Live“. Výsledkem je výkonnost vzorů při simulaci nasazení vzorů do reálného prostředí, čili v souboru budou i případné neúspěšné vzory. Součet výkonností vzorů v tomto souboru představuje výsledek predikce, kdyby tyto vzory byly nasazeny v reálném prostředí, čili na elektronické burze.

5.11 APLIKACE V JINÝCH ODVĚTVÍCH

Vyhledávání vzorů a klasifikace je často využíváno v mnoha dalších odvětvích. V oblasti biomedicíny lze rozeznávání vzorů využít například při detekci zvýšeného rizika infarktu z rozboru moči. Časté využití lze nalézt v rozpoznávání obrazových vzorů, kde lze například detekovat tvář v obraze či klasifikovat nasnímaný terén z video kamery. V akustice lze těchto aplikací využít k rozpoznávání řeči. Těchto metod lze také využít při rozpoznávání rukopisu.

Klasifikace a rozpoznávání v oblasti exaktních věd je snadnější, než aplikace u ekonomických systémů, kdy systém obsahuje velké množství vstupů, které nelze matematicky popsat. Mezi tyto vlivy patří emoce účastníků trhu, krize, teroristické útoky či vliv důležitých fundamentálních zpráv.

6. ZÁVĚR

Cílem diplomové práce bylo udělat rešerši v oblasti vyhledávání vzorů a vytvořit aplikaci, která dle zvolené metody vyhledávání nalezne vzory vykazující predikční schopnosti.

Pro vývoj aplikace byl volen jazyk C++. Při tvorbě aplikace pro vyhledávání vzorů je využito metody založené na podobnosti v datech, která byla použita pro proces učení. Jde o metodu založenou na instancích – IBL (Instance Based Learning). Metoda IBL je založena na velmi jednoduchém principu, ale je výpočetně náročná. Základ spočívá v načtení celých vstupních dat do paměti a při předložení nové situace, u níž má být provedena predikce, provést porovnání a vyhodnocení. Je-li aplikace této metody v prostředí, kde je požadavek rychlého rozhodování, je třeba provést úpravu. Hlavní rozdíl spočívá v množství dat, která budou ukládána do paměti.

Před aplikací v prostředí, kde požadujeme predikci, je proveden proces učení a je provedena selekce pouze těch vzorů, které v oblasti vstupních dat vykazují predikční schopnosti. Tímto procesem se radikálně zmenší velikost načítaných dat a dojde k urychlení výpočtů a je možno využívat predikce i v prostředích, která tuto rychlost vyžadují.

Pro predikci je voleno prostředí elektronických burz, které právě vyžaduje velmi rychlé rozhodování. Vstupní data při pětiminutovém časovém rámci po dobu jednoho roku obsahují přibližně 70 000 záznamů, což je velký objem dat pro zjišťování predikčních schopností nově příchozí situace v reálném čase. Procesem učení provedeme výběr vzorů s predikčními schopnostmi a při běhu rozhodovacího procesu už pouze porovnáváme vybrané vzory a nové situace, na které při splnění požadavku definované míry podobnosti provedeme predikci.

Nejdůležitější částí procesu je ověřování. Vzory vykazující predikční schopnosti v oblasti učících dat podrobíme analýze výkonnosti na nových datech. Tímto procesem neprojde drtivá většina úspěšných vzorů po fázi učení. Detailnější ověřování vzorů lze provést několikanásobným ověřením na nových datech.

V aplikaci jsou implementovány dvě funkce pro vyjádření chyby, čili zjištění míry podobnosti dvou vstupních prvků. První metoda je založena na součtu jednotlivých chyb a druhá metoda je založena na splnění podmínky podobnosti jednotlivých prvků porovnávaných dat. Lepší výsledky jsou dosaženy aplikací druhé metody výpočtu odlišnosti.

Dosažené výsledky jsou při použití aplikace, kde vstupní data pro učení tvoří historická data průběhu ceny finančního instrumentu(eur/usd a usd/jpy) za období, které začíná 2.1.2008 a končí 30.6.2008. Pro proces ověřování výkonnosti vzorů je použita databáze historických cen za období 1.7.2008 až 31.12.2008. Výstupem procesu ověřování je seznam vzorů vykazujících predikční schopnosti jak v datech učících, tak v datech nových. Tyto vzory byly použity pro obchodování v roce 2009 s kladným výsledkem, kdy průměrný zisk na jeden obchod u strategie použité pro měnový pár eur/usd představoval 0,4% a u strategie použité pro měnový pár usd/jpy je průměrný zisk na jeden obchod roven přibližně 1%. Vyhledávání vzorů bylo aplikování na měnových párech eur/usd a usd/jpy.

Tvar nalezených vzorů vypovídá o tom, že při spekulaci nad cenovým poklesem očekáváme nejdříve krátký cenový nárůst a naopak při spekulaci nad vzrůstem ceny vyhledáváme krátký cenový pokles, který predikuje následný pohyb nahoru. Jde o tzv. korekce pohybů.

Tuto práci budu dále rozvíjet v rámci disertační práce na Fakultě podnikatelské, kdy mám v plánu využít sofistikovanějších metod učení a predikce.

7. LITERATURA

[1] Ing. Petr Honzík, Ph.D.: Strojové učení. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2006. 85s

[2] Doc. Ing. Petr Dostál, CSc.: Pokročilé metody analýz a modelování v podnikatelství a veřejné správě, Brno: CERM, 2008. 340s. ISBN 978-80-7204-605-8

[3] Přispěvatelé Wikipedie, Data mining [online], Wikipedie: Otevřená encyklopedie, c2009, Datum poslední revize 19. 02. 2009, 19:58 UTC, [citováno 21.04.2009] http://cs.wikipedia.org/w/index.php?title=Data_mining&oldid=3643283

[4] Přispěvatelé Wikipedie, Statistika [online], Wikipedie: Otevřená encyklopedie, c2009, Datum poslední revize 26. 01. 2009, 19:04 UTC, [citováno 21. 04. 2009] http://cs.wikipedia.org/w/index.php?title=Statistika&oldid=3550458

[5] Přispěvatelé Wikipedie, Expertní systém [online], Wikipedie: Otevřená encyklopedie, c2009, Datum poslední revize 2. 04. 2009, 13:02 UTC, [citováno 21.04.2009] http://cs.wikipedia.org/w/index.php?title=Expertn%C3%AD_syst%C3%A9m&oldid=3803170

[6] Sang C. Suh, Dan Li, Jingmiao Gao: A novel chart pattern recognition approach: A case study on Cup with Handle. [citováno 21.10.2008] http://istg.tamu-commerce.edu/Suh-Papers/ANNIE2004.pdf http://istg.tamu-commerce.edu/Suh-Papers/IDPT2005.pdf

[7] [online] Silas N. Onyango: On the pattern recognition of Verhulst-logistic Ito Processes in Market Price Data. [citováno 20.04.2009] http://www.strathmore.edu/research/nairobi_stock_exchange_study.pdf

[8] [online] Robi Polikar: Pattern Recognition. [citováno 13.04.2009] http://users.rowan.edu/~polikar/RESEARCH/PUBLICATIONS/wiley06.pdf

[9] [online], Invariant PR using Bayesian Inference on Hierarchical Sequences. [citováno 10.3.2009] http://www.stanford.edu/~dil/RNI/DilJeffTechReport.pdf

[10] [online], Time Series Analysis, [citováno 10.9.2008] http://www.statsoft.com/textbook/sttimser.html

[11] John H. Cochrane, [online], Time Series for Macroeconomics and Finance, University of Chicago, Chicago, 1997, [citováno 10.9.2008] http://faculty.chicagobooth.edu/john.cochrane/research/Papers/time_series_book.pdf

[12] David R. Brillinger, [online], Time Series: General, University of California, Berkley, CA, USA, 17.11.2000 http://www.stat.berkeley.edu/~brill/Papers/encysbs.pdf

[13] [online], Analýza časových řad, [citováno 11.9.2008] http://iastat.vse.cz/casovky/casovky0.htm

[14] [online], Numerical Weather Forecast, [citováno 10.5.2008] http://new.meteo.pl/index_eng.php

[15] Software METATRADER, volný zdroj historických burzovních dat

[16] Přispěvatelé Wikipedie, Metoda nejmenších čtverců [online], Wikipedie: Otevřená encyklopedie, c2009, Datum poslední revize 5. 03. 2009, 16:53 UTC, [citováno 20. 05. 2009] http://cs.wikipedia.org/w/index.php?title=Metoda_nejmen%C5%A1%C3%ADch_%C4%8Dtverc%C5%AF&oldid=3692405

[17] Dezider Meško, Normalizace dat pro neuronovou síť GAME, České vysoké učení technické v Praze, Fakulta elektrotechnická, 2008.45s

SEZNAM PŘÍLOH

Na přiloženém CD jsou tyto složky:

Dokumentace:	Obsahuje diplomovou práci v elektronické podobě, ve formátu DOC a PDF.
Metadata:	Metadata.
Program:	Obsahuje aplikaci pro vyhledávání vzorů v dynamických datech, která byla vytvořena v prostředí C++ Builder.
Data:	Obsahuje datové soubory pro proces vyhledávání vzorů. K dispozici je měnový pár eur/usd, usd/jpy a testovací soubory.