

Data mining v energetickém průmyslu

Data mining in energetic industry

Ing. Jaromír Špico

Bakalářská práce
2009



Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky
Ústav aplikované informatiky
akademický rok: 2008/2009

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení: **Ing. Jaromír ŠPICO**
Studijní program: **B 3902 Inženýrská informatika**
Studijní obor: **Informační technologie**

Téma práce: **Data mining v energetickém průmyslu**

Zásady pro vypracování:

1. Popište problematiku dobývání znalostí z rozsáhlých databází vybraných energetických provozů.
2. Vyberte vhodné metody pro praktickou aplikaci data miningu.
3. Proveďte přípravu a statistickou analýzu získaných dat.
4. Analyzujte připravená data za účelem vytěžení nových informací a poznatků o konkrétním provozu.
5. Vhodně formulujte získané výsledky a navrhněte postup dalšího výzkumu v této oblasti.

Rozsah práce:

Rozsah příloh:

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

1. EDELSTEIN, H. A. Introduction to Data Mining and Knowledge Discovery. Two Crows Corporation, 1999. ISBN 1892095025
2. HAN, J., KAMBER, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2000. ISBN 1558604898
3. HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Verlag, 2001. ISBN 038799952845
4. MARŠÍK, R. Dolování dat – nastupující technologie na poli IT. ComputerWorld, příloha Business World, 1998, r.4, č.46. ISSN 1210-8790
5. PARR RUD, O. Data mining – Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM). Praha: Computerpress, 2002. ISBN 80-7226-577-6
6. ŠARMANOVÁ, J. Teorie a praxe dolování znalostí z dat. Informační bulletin ČStS, 2000, r.11, č.2, s. 16-26. ISSN 1210-8022
7. WEISS, S. M. Predictive Datamining: A practical guide. Morgan Kaufmann Publishers, 1997. ISBN 1558604030

Vedoucí bakalářské práce:

Ing. Pavel Vařacha

Ústav aplikované informatiky

Datum zadání bakalářské práce:

20. února 2009

Termín odevzdání bakalářské práce:

1. června 2009

Ve Zlíně dne 13. února 2009



prof. Ing. Vladimír Vašek, CSc.

děkan



doc. Ing. Ivan Zelinka, Ph.D.

ředitel ústavu

ABSTRAKT

Bakalářská práce se zabývá možnostmi aplikace data miningu v oblasti energetiky. V teoretické části je výčet a popis vhodných metod pro aplikaci data miningu. Z mnoha metod je podrobněji popsána predikce pomocí neuronových sítí. V praktické části je návod na predikci pomocí výpočetního prostředí Matlab a jeho toolboxu pro neuronové sítě. Konkrétní výsledky predikce jsou prezentovány na rozsáhlých datech získaných ve spolupráci s teplárnou Most - Komořany.

Klíčová slova: data mining, energetický průmysl, umělé neuronové sítě, Matlab, predikce, neural network toolbox, průměrná absolutní procentuelní chyba, teplárna

ABSTRACT

The bachelor thesis deals with possibilities of application of data mining in the sphere of energetic industry. In the theory there is an enumeration and description of appropriate methods for data mining application. From many methods the prediction using neural networks is described in more detail. In the practical part there are instructions for prediction using computing environment Matlab and its toolbox for neural networks. The specific results of prediction are presented on large data obtained in co-operation with the heating plant Most – Komořany.

Keywords: data mining, energetic industry, artificial neural networks, Matlab, prediction, neural network toolbox, mean absolute percentage error, heating plant

PODĚKOVÁNÍ

„Predikce je velmi těžká, zejména o budoucnosti.“

Niels Bohr, dánský fyzik

Na tomto místě bych chtěl vyjádřit své poděkování vedoucímu bakalářské práce, kterým byl pan Ing. Bc. Pavel Vařacha z Ústavu aplikované informatiky na Fakultě aplikované informatiky ve Zlíně. Děkuji mu za zodpovědné vedení a za mnohé praktické rady a připomínky, které byly cennými podněty pro vznik této práce.

Prohlašuji, že

- beru na vědomí, že odevzdáním bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby;
- beru na vědomí, že bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk bakalářské práce bude uložen v příruční knihovně Fakulty aplikované informatiky Univerzity Tomáše Bati ve Zlíně a jeden výtisk bude uložen u vedoucího práce;
- byl/a jsem seznámen/a s tím, že na moji bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- beru na vědomí, že podle § 60 odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- beru na vědomí, že podle § 60 odst. 2 a 3 autorského zákona mohu užít své dílo – bakalářskou práci nebo poskytnout licenci k jejímu využití jen s předchozím písemným souhlasem Univerzity Tomáše Bati ve Zlíně, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše);
- beru na vědomí, že pokud bylo k vypracování bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tedy pouze k nekomerčnímu využití), nelze výsledky bakalářské práce využít ke komerčním účelům;
- beru na vědomí, že pokud je výstupem bakalářské práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá. Neodevzdání této součásti může být důvodem k neobhájení práce.

Prohlašuji,

že jsem na bakalářské práci pracoval samostatně a použitou literaturu jsem citoval.

V případě publikace výsledků budu uveden jako spoluautor.

Ve Zlíně 1. června 2009

.....
podpis diplomanta

OBSAH

ÚVOD	9
I TEORETICKÁ ČÁST	10
1 PROČ JE DOBRÉ VYUŽÍT METODY DATA MININGU PRO ŘÍZENÍ ENERGETICKÉHO SYSTÉMU	11
1.1 DATA MINING – METODY DOBÝVÁNÍ ZNALOSTÍ Z DATABÁZÍ.....	12
1.1.1 Algoritmus k nejbližších sousedů.....	13
1.1.2 Analýza časových řad.....	14
1.1.3 Analýza sekvencí.....	14
1.1.4 ANOVA, ANCOVA	14
1.1.5 Bayesovské sítě	15
1.1.6 Diskriminační analýza.....	16
1.1.7 Evoluční (genetické) algoritmy	16
1.1.8 Faktorová analýza.....	17
1.1.9 Fuzzy logika	17
1.1.10 Kohonenovy mapy.....	18
1.1.11 Kontingenční tabulky	19
1.1.12 Korelace	20
1.1.13 Lineární regrese	20
1.1.14 Logistická regrese.....	20
1.1.15 Metody exploratorní analýzy dat.....	20
1.1.16 Naivní Bayesovský klasifikátor.....	21
1.1.17 Neuronové sítě.....	21
1.1.18 Rozhodovací pravidla.....	22
1.1.19 Rozhodovací stromy.....	22
1.1.20 Shluková analýza.....	23
1.1.21 Support vector machines	23
1.2 VÝBĚR VHODNÉ METODY	24
2 NEURONOVÉ SÍTĚ A PREDIKCE	26
2.1 MODEL NEURONU.....	26
2.2 ARCHITEKTURA NEURONOVÉ SÍTĚ	28
2.3 PRINCIPY PREDIKCE S NEURONOVÝMI SÍTĚMI.....	30
2.3.1 Vícenásobná predikce a autopredikce	31
2.3.2 Predikční chyba	31
II PRAKTICKÁ ČÁST	33
3 ANALÝZA POUŽITÝCH DAT A VÝPOČETNÍCH PROSTŘEDKŮ	34
3.1 ELEKTRÁRNA MOST - KOMOŘANY	34
3.1.1 Historie elektrárny Komořany.....	34
3.1.2 Teplo, elektrina a vedlejší produkty.....	35
3.1.3 Analýza získaných a použitých dat	36

3.2	MICROSOFT OFFICE EXCEL 2007 – PŘÍPRAVA DAT	38
3.3	MATLAB R2009A A NEURAL NETWORK TOOLBOX – POUŽITÍ DAT	39
3.3.1	Import dat do neuronové sítě.....	39
3.3.2	Učení sítě na historických datech (<i>nftool</i>).....	41
3.3.3	Postup predikce (<i>nntool</i>)	44
4	VÝSLEDKY PREDIKCE V SYSTÉMU ŘÍZENÍ DISTRIBUCE TEPLA MĚSTSKÉ AGLOMERACE POUŽITÍM NEURONOVÝCH SÍTÍ	46
4.1	PREDIKCE TEPLoty TOPNÉ VODY.....	46
4.2	SPECIALIZOVANÉ SÍTĚ	48
4.3	ZMĚNA POČTU NEURONŮ V SÍTI	49
	ZÁVĚR	53
	ZÁVĚR V ANGLIČTINĚ.....	54
	SEZNAM POUŽITÉ LITERATURY	55
	SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK	59
	SEZNAM OBRÁZKŮ	60
	SEZNAM TABULEK.....	62
	SEZNAM PŘÍLOH.....	63

ÚVOD

V dnešní době pocítujeme naléhavost otázek spojených s ekologií a ekonomikou. Už dávno platí výrok Benjamina Franklina *A Penny Saved Is A Penny Earned* – v překladu *Ušetřená koruna je koruna vydělaná*.

Snad v žádné oblasti neexistuje tak úzké spojení ekologie a ekonomiky, jako právě v oblasti energetiky. Každá firma z oboru se snaží zvýšit svoji efektivnost a hospodárnost. V oblasti provozů tepláren a elektráren je důležité znát vývoj spotřeby energetických veličin tak, aby nedocházelo ke zbytečnému plýtvání vzácných surovin, ale ani k podcenění vývoje a nedostatečné obsluze požadavků spotřebitelů. Z každodenního a neustálého energetického provozu se hromadí miliardy cenných dat, pro které existuje klíč k jejich pochopení – data mining. Firma, která bude používat sofistikované metody data miningu, bude mít větší finanční úspěch a bude schopna financovat investice do budoucnosti k prospěchu celé společnosti.

V této bakalářské práci se snažím popsat dostupné metody data miningu a využít je k predikci na skutečných datech.

V teoretické části je jednoduchý a stručný výčet známých metod data miningu. Podrobnější teoretický a praktický popis je možné získat v citované a jiné literatuře. Pro aplikaci jsem si vybral metodu neuronových sítí, které tvoří druhou část teorie. Zaměření je na aplikaci predikce pomocí neuronových sítí.

Praktická část poskytuje návod pro predikci pomocí neuronových sítí ve výpočetním prostředí Matlab. Druhou část tvoří prezentace výsledků vlastní predikce. Analyzovány byly data z teplárny Most – Komořany.

V závěru navrhuji možný postup dalšího výzkumu v této oblasti – použití metod data miningu a vyšších algoritmů umělé inteligence pro řízení teplárenských soustav městské aglomerace.

I. TEORETICKÁ ČÁST

1 PROČ JE DOBRÉ VYUŽÍT METODY DATA MININGU PRO ŘÍZENÍ ENERGETICKÉHO SYSTÉMU

Energetické distribuční sítě jsou komplexní a dynamické systémy. Pro jejich efektivní řízení je třeba důkladně porozumět základním mechanismům jejich chování. Správné zodpovězení otázek typu kolik? kdy? kam? přináší distributorům mnoho finančních výhod. Obecně existuje několik způsobů, jak stále nenahraditelné lidské porozumění doplnit porozuměním strojovým.

Distribuční sítě komodit jako elektřina, voda, plyn nebo teplo jsou prostředím vyžadujícím rychlé a efektivní řízení a rozhodování. Lidský operátor na základě své zkušenosti dobře chápe základní principy určující chování sítě. Současně ale jde o prostředí dynamická a hierarchická, generující velké objemy časových dat. Jde tedy o prostředí, kde může docházet k chybným nebo pozdním reakcím na konkrétní, potenciálně ne zcela obvyklý vývoj událostí. Adaptivní systémy pro podporu rozhodování založené na automaticky vytvářených datových modelech mohou významně napomoci kvalitě rozhodování a řízení. Obvykle jde o minimalizaci distribučních ztrát včasnou reakcí na změnu vstupních, vnitřních či výstupních parametrů sítě. Důsledkem včasných a přesných řídicích zásahů je také optimalizace nákladů odvozených ze smluvních vztahů, a tím i snížení koncové ceny předmětu spotřeby, zvýšení provozního zisku či komfortu koncových odběratelů.

Moderní metody strojového učení, tzv. dolování dat (získávání informací ze souborů dat) a matematické statistiky nabízejí množství možností, jak zmíněné modely vytvářet. Z obecného pohledu jde o induktivní postupy založené na zobecnění konkrétních, systematicky a dlouhodobě sledovaných údajů. Základními údaji jsou data zjišťovaná na různých místech distribuční sítě – typicky jde o odběr nebo průtok v daném kontrolním bodě. Doplnkovými údaji mohou být data zjišťovaná mimo distribuční síť a mající úzký vztah k objemu spotřeby sledované komodity – nejčastějšími doplňkovými veličinami jsou údaje o počasí, dále lze uvažovat např. cenu distribuovaného předmětu spotřeby či vstupních surovin.

Sledované veličiny jsou vzorkovány po určitých, nejčastěji pevně stanovených časových intervalech. Modelování v distribučních sítích lze převádět na analýzu, resp. „dolování“ z časových řad více proměnných. Teoretickým metodám zpracování časových dat budou věnovány další odstavce. [1]

1.1 Data mining – metody dobývání znalostí z databází

Data mining ([dejta majnyn], angl. dolování z dat či vytěžování dat) je analytická metodologie získávání netriviálních skrytých a potenciálně užitečných informací z dat. Někdy se chápe jako analytická součást dobývání znalostí z databází (Knowledge Discovery in Databases, KDD), jindy se tato dvě označení chápou jako souznačná. [2]

V dnešní době se používá velké množství metod dobývání znalostí z dat:

- Algoritmus k nejbližších sousedů
- Analýza časových řad
- Analýza sekvencí
- ANOVA, ANCOVA
- Bayesovské sítě
- Diskriminační analýza
- Evoluční (genetické) algoritmy
- Faktorová analýza
- Fuzzy logika
- Kohonenovy mapy
- Kontingenční tabulky
- Korelace
- Lineární regrese
- Logistická regrese
- Metody exploratorní analýzy dat
- Naivní bayesovský klasifikátor
- Neuronové sítě
- Rozhodovací pravidla
- Rozhodovací stromy
- Shluková analýza

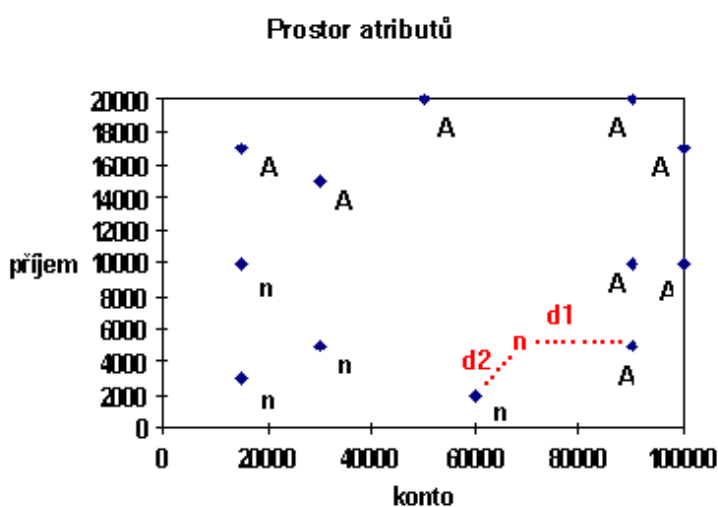
- Support vector machines
- a mnohé další [3]

V praxi rozlišujeme dva typy modelů, a to

- 1) **prediktivní, regresní** (jejich cílem je předpovědět hodnoty nějakých atributů na základě již známých hodnot jiných atributů),
- 2) **deskriptivní, klasifikační** (zde se popisují vzory v existujících datech, které mohou ovlivňovat rozhodování). [4]

1.1.1 Algoritmus k nejbližších sousedů

Patří mezi metody založené na analogii. V případě nejbližšího souseda (k-nearest neighbour rule) jsou koncepty (třídy) reprezentovány svými typickými představiteli. V procesu klasifikace se pak nový příklad zařadí do třídy na základě podobnosti (nejmenší vzdálenosti k reprezentantovi nějaké třídy – viz obr. 1). Jde tedy o metodu, která vychází ze shlukové analýzy. Klíčovým pojmem je koncept podobnosti, resp. vzdálenosti dvou příkladů. [5] Každý případ se posuzuje jako vektor o n komponentách, přičemž n je počet atributů nebo charakteristik. Metoda nepotřebuje učící fázi. K předpovědi třídy řešeného případu porovná algoritmus řešený případ se všemi případy trénovací množiny nebo s paměti a vypočte vzdálenost mezi nimi. Potom je většinová třída pro K nejpodobnějších trénovacích případů předpovědí pro řešený případ. Vzdálenost použitá v případech je eukleidovská vzdálenost mezi vektory. [6]



Obr. 1. Klasifikace dle nejbližšího souseda [5]

1.1.2 Analýza časových řad

Časová řada je chronologicky uspořádaná posloupnost hodnot určitého statistického ukazatele. Tento ukazatel musí být v čase vymezen věcně a prostorově shodně. Prakticky to znamená, že časová řada je řada čísel; tuto řadu tvoří hodnoty nějaké (např. ekonomické) veličiny, které jsou uspořádány od nejstarších po nejmladší nebo naopak. Z formálního hlediska je časová řada realizací náhodného procesu. Mezi tyto hlavní metody patří především:

- expertní (kvalitativní) metody - Delphi metoda, metoda historické analogie, dotazování
- grafická analýza – vyhledávání trendů a obrazců (patterns) v grafech
- dekompozice časových řad - časovou řadu tvoří trend, sezónní, cyklická a náhodná složka
- Box – Jenkinsovská analýza – náhodná složka je tvořena korelovanými (závislými) veličinami
- spektrální analýza - časová řada je nekonečnou směsí sinusových a kosinusových křivek s různými frekvencemi a amplitudami
- lineární dynamické (ekonometrické) modely - kauzální modely, kde je vysvětlovaná proměnná vysvětlována pomocí jedné nebo více vysvětlujících proměnných
- další kvantitativní metody [7]

1.1.3 Analýza sekvencí

Vysoce specializovaná metoda, v současnosti se používá hlavně v bioinformatice při výzkumu DNA. Používá se i pro predikci, např. v marketingu na tzv. NPTB model (Next Product To Buy – další produkt, který si zákazník koupí). [8]

1.1.4 ANOVA, ANCOVA

Analýza rozptylu (Analysis of variance – ANOVA) je metodou matematické statistiky, která umožňuje ověřit, zda na hodnotu náhodné veličiny pro určitého jedince má statisticky významný vliv hodnota některého znaku, který se u jedince dá pozorovat. Tento znak musí nabývat jen konečného počtu možných hodnot (nejméně dvou) a slouží k rozdělení jedinců

do vzájemně porovnávaných skupin. Analýza kovariance (Analysis of covariance – ANCOVA) je spojení metody ANOVA s regresí pro spojité proměnné. [9,10]

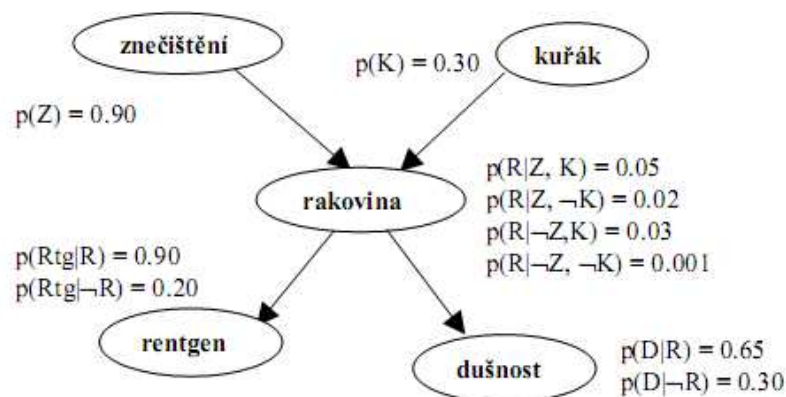
1.1.5 Bayesovské sítě

Bayesovské sítě spolu s bayesovskými metodami ve statistice a pravděpodobnosti představují velmi propracovaný a principiální způsob, jak uchopit a dále pracovat se znalostmi a informacemi zatíženými neurčitostí. Bayesovské sítě jsou grafické modely, schopné reprezentovat vztahy mezi proměnnými z určité problémové domény. Tyto vztahy mohou mít jak kauzální, tak i pravděpodobnostní interpretaci; jeví se tedy jako vhodná reprezentace pro kombinování apriorních expertních znalostí (které jsou často v pro člověka přirozenější kauzální formě) a dat. Bayesovské metody ve statistice a pravděpodobnosti představují způsob, jak zamezit „přílišné přiléhavosti“ (overfittingu) automatizovaně naučeného grafického modelu vzhledem k datům (šum, chyby v měření). Jelikož model reflektuje závislosti mezi všemi proměnnými v rámci problémové domény, je schopen se vyrovnat i se situacemi, kdy některé datové položky chybějí (tzv. missing values). Na rozdíl od pravidlových systémů umožňují Bayesovské sítě zachycení širších souvislostí.

Bayesovské uvažování slouží k aktualizaci našeho mínění o určitých hypotézách na základě nově přijatých informací (pozorování). K této aktualizaci používáme Bayesův vzorec:

$$p(H_i|E) = \frac{P(E|H_i)P(H_i)}{P(E)} = \frac{P(E|H_i)P(H_i)}{\sum_{j=1}^n P(E|H_j)P(H_j)} \quad (1)$$

Konvencí je značit jednotlivé hypotézy jako H_i a pozorování (evidenci) jako E .



Obr. 2. Rozdělení pravděpodobnosti v uzlech Bayesovské sítě [11]

Bezpochyby nejoblíbenější oblastí aplikace Bayesovských sítí je medicína. Důvodů může být několik: jedná se o velice komplexní doménu s množstvím nashromážděných expertních znalostí nejrůznějšího charakteru, Bayesovské sítě jsou schopny explicitně modelovat kauzální intervence, usuzovat diagnosticky i prediktivně a výhodou je také jejich vizuální povaha, která usnadňuje jejich použití při vysvětlování. [11]

1.1.6 Diskriminační analýza

Diskriminační analýza patří mezi metody zkoumání závislosti mezi skupinou p nezávisle proměnných, nazvaných diskriminátory, tj. sloupců zdrojové matice na jedné straně a jednou kvalitativní závisle proměnnou na druhé straně. Umožňuje zařazení objektu do jedné z již existujících tříd. Ve vstupních datech jsou svými hodnotami diskriminátorů u všech objektů dány zařazené objekty do primárních tříd. Dále jsou dány nezařazené objekty, pro které budeme hledat zařazení do třídy. Objekt zařadíme do třídy na základě jeho největší míry podobnosti, např. nejmenší Mahalanobisovy vzdálenosti. Předpokladem pro provedení diskriminační analýzy je především dostatečný počet objektů (nejméně 30) a normalita rozložení hodnot. [12] Diskriminační analýza:

1. Přiřazuje subjekty do předem definovaných skupin.
2. Předpokládá, že část populace je rozdělena do skupin a část ne - pro nezařazené vybere skupinu, která je jim nejbližší.
3. Vytváří nové skupiny na základě podobnosti s již existujícími. [13]

1.1.7 Evoluční (genetické) algoritmy

Genetické algoritmy patří do třídy evolučních algoritmů, které mimo ně zahrnují také evoluční programování, evoluční strategii a genetické programování. Jsou to vyhledávací algoritmy založené na mechanismu přirozeného výběru a principech genetiky. Jejich velkou výhodou je poměrná jednoduchost. Ideovým vzorem pro genetické algoritmy byly principy vývoje, které se uplatňují v přírodě. Zde existují populace jednotlivých živočišných druhů, složených z jedinců různých vlastností. Tyto vlastnosti jsou prvotně zakódovány v jejich genech, které tvoří větší celky, chromozómy. Při křížení vznikají noví jedinci, kteří mají zpravidla náhodně část genů od jednoho rodiče a část genů od rodiče druhého. Přitom ve zvlášť výjimečném případě může dojít k náhodné změně některého

genů v chromozómu, tzv. mutaci, která může být pro další vývoj druhu příznivá nebo ne. Podle svých vlastností má každý z potomků větší nebo menší schopnost obstát v přirozeném výběru a vytvořit další generaci. Proces výběru se stále opakuje a v jeho průběhu se zlepšují genetické vlastnosti daného druhu. Tak probíhala celá evoluce v přírodě. [14]

Princip práce genetického algoritmu je postupná tvorba generací různých řešení daného problému. Při řešení se uchovává tzv. populace, jejíž každý jedinec představuje jedno řešení daného problému. Jak populace probíhá evolucí, řešení se zlepšují. Tradičně je řešení reprezentováno binárními čísly, řetězci nul a jedniček, nicméně používají se i jiné reprezentace (strom, pole, matice, ...). Typicky je na začátku simulace (v první generaci) populace složena z naprosto náhodných členů. V přechodu do nové generace je pro každého jedince spočtena tzv. fitness funkce, která vyjadřuje kvalitu řešení reprezentovaného tímto jedincem. Podle této kvality jsou stochasticky vybráni jedinci, kteří jsou modifikováni (pomocí mutací a křížení), čímž vznikne nová populace. Tento postup se iterativně opakuje, čímž se kvalita řešení v populaci postupně vylepšuje. Algoritmus se obvykle zastaví při dosažení postačující kvality řešení, případně po předem dané době. [15]

1.1.8 Faktorová analýza

Faktorová analýza patří mezi metody redukce počtu původních proměnných. Ve faktorové analýze předpokládáme, že každou vstupující proměnnou můžeme vyjádřit jako lineární kombinaci nevelkého počtu společných skrytých faktorů a jediného chybového faktoru. Snažíme se vysvětlit závislost proměnných. K nevýhodám metody patří zejména nutnost zadat počet společných faktorů ještě před prováděním vlastní analýzy. [16]

1.1.9 Fuzzy logika

Fuzzy logika se poprvé objevila v roce 1965 v článku, jehož autorem byl profesor Lotfi A. Zadeh. Tehdy byl definován základní pojem fuzzy logiky a to fuzzy množina. Slovo fuzzy znamená neostrý, matný, mlhavý, neurčitý, vágní. Odpovídá tomu i to, čím se fuzzy teorie zabývá: snaží se pokrýt realitu v její nepřesnosti a neurčitosti. V klasické teorii množin prvek do množiny buďto patří (úplné členství v množině) nebo nepatří (žádné členství v množině). Fuzzy množina je množina, která kromě úplného nebo žádného členství připouští i členství částečné. To znamená, že prvek patří do množiny s jistou

pravděpodobností (stupeň příslušnosti). Funkce, která každému prvku universa přiřadí stupeň příslušnosti, se nazývá funkce příslušnosti.

Použití fuzzy technologie:

- Fuzzy regulace v japonském metru — automatické řízení metra — zvýšená přesnost zastavování, plynulejší brzdění a hlavně nižší spotřeba energie.
- Fotoaparát s automatickým vyhledáváním centrálního bodu pro zaostření (Minolta)
- ABS, řízení motoru, volnoběhu a klimatizace (Honda, Nissan, Subaru)
- Řízení výtahů (Mitsubishi)
- Korekce chyb ve slévárenských zařízeních na plastické výrobky (Omron)
- 3.5" disketové mechaniky (zlepšení doby vystavení hlaviček až o 30 %)
- palmtop Kanji určený pro rozpoznávání ručně psaných textů
- rozpoznávání řeči
- Fuzzy SQL (Omron)
- Pomoc při hledání identifikačních a profilových systémů pachatele (velký, ne příliš těžký, víceméně starý, ...)
- Analýza portfolia při investování na kapitálovém trhu [17]

1.1.10 Kohonenovy mapy

Samoorganizující neuronové sítě s učením bez učitele jsou stále více využívány pro rozlišení, rozpoznávání a třídění neznámých číslcových signálů a dat. Hlavním představitelem jsou Kohonenovy mapy. Ty sami rozpoznávají shodné prvky nebo naopak rozdíl mezi signály a je tak s nimi možné zpracovat úplně neznámé signály a data.

Dnes již mezi základní typy neuronových sítí a také mezi nejpobulárnější, patří tzv. SOM = Self-Organizing Maps (Samoorganizující se mapy), častěji známé po svém "stvořiteli" jako Kohonenovy mapy. Ty patří do skupiny samoučících se neuronových sítí, tzn. sítí s učením bez učitele, které ke svému nastavování nepotřebují ideální vzory. To znamená, že k učení sítě stačí jen velká skupina reálných signálů, z nichž některé mají určitou společnou vlastnost nebo naopak výrazné odlišnosti a již k nim nemusí být přiřazeny žádné ideální učící signály nebo informace (target = cílové hodnoty). Ty v případě tzv. učení s učitelem

udávají konečný cílový stav, do kterého se má síť učením dostat. A právě jejich získání bývá často velkým problémem. Naopak u SOM (Kohonenovy mapy) nám například stačí jen skupina nahranych řečových signálů a během učení si síť již sama nalezne společné znaky a odlišnosti, podle kterých se bude ve své aktivní činnosti rozhodovat. A to je ta výhoda, která za cca 20 let od vzniku Kohonenových map z nich udělala velmi často využívanou a velmi oblíbenou neuronovou síť. [18]

1.1.11 Kontingenční tabulky

Kontingenční tabulka se užívá k přehledné vizualizaci vzájemného vztahu dvou statistických znaků. Řádky kontingenční tabulky odpovídají možným hodnotám prvního znaku, sloupce pak možným hodnotám druhého znaku. V příslušné buňce kontingenční tabulky je pak zařazen počet případů, kdy zároveň měl první znak hodnotu odpovídající příslušnému řádku a druhý znak hodnotu odpovídající příslušnému sloupci. Například prvním znakem může být pohlaví člověka a druhým znakem měsíc jeho narození. Kontingenční tabulka o 2 řádcích (žena, muž) a 12 sloupcích (leden, únor, ..., prosinec) pak popisuje počty výskytů všech kombinací pohlaví a měsíce v nějakém souboru sledovaných jedinců.

Je možné, aby jeden řádek či sloupec odpovídal více možným hodnotám znaku. To se děje v případě, kdy znak nabývá některých hodnot příliš zřídka, takže je vhodné spojit více možných hodnot.

Součty (mezisoučty) všech hodnot v každém řádku, resp. sloupci nesou informaci o počtu výskytů jevů, při nichž nabyl první (resp. druhý znak) příslušné hodnoty bez ohledu na hodnotu druhého (resp. prvního) znaku.

Kromě prostého popisu četností kombinací hodnot dvou znaků nabízí kontingenční tabulka možnost testovat, zda mezi oběma znaky existuje nějaký vztah. K tomu lze užít např. test dobré shody. Znaky použité k zobrazení v kontingenční tabulce pak musí představovat diskrétní hodnoty (je možné tedy využít kvalitativní, diskrétně kvantitativní či spojitě kvantitativní znaky, v posledním případě však pouze s rozdělením jednotlivých znaků do skupin – tzv. skupinové třídění). [19]

1.1.12 Korelace

Korelace je ve statistice vzájemný vztah mezi znaky či veličinami. Korelační koeficient může nabývat hodnot od -1 až po $+1$. Hodnota korelačního koeficientu -1 značí zcela nepřímou závislost, tedy čím více se zvětší hodnoty v první skupině znaků, tím více se zmenší hodnoty v druhé skupině znaků, např. vztah mezi uplynulým a zbývajícím časem. Hodnota korelačního koeficientu $+1$ značí zcela přímou závislost, např. vztah mezi rychlostí bicyklu a frekvencí otáček kola bicyklu. Pokud je korelační koeficient roven 0 , pak mezi znaky není žádná statisticky zjiřitelná závislost, např. vztah mezi hodnotami porodnosti v Křemílkově a počtem čápů v Křemílkově. [20]

1.1.13 Lineární regrese

Regresní analýza lineární závislosti má za úkol určit odhady koeficientů a (posunutí) a b (směrnice), které charakterizují regresní přímku, vyjádřenou rovnicí $y = a + bx$. Předpokládá se, že nezávisle proměnná x je prakticky bez chyby nebo aspoň s chybou podstatně menší než je chyba závisle proměnné y . Regresní analýza se uskutečňuje "metodou nejmenších čtverců". Pro odhady regresních koeficientů platí známé sumační vztahy. [21]

1.1.14 Logistická regrese

Logistická regrese je označení metody matematické statistiky zabývající se problematikou odhadu pravděpodobnosti nějakého jevu (závisle proměnné) na základě určitých známých skutečností (nezávisle proměnných), které mohou ovlivnit výskyt jevu. Událost, zda zkoumaný jev nastal, se modeluje pomocí náhodné veličiny, která nabývá hodnoty 0 , pokud jev nenastal, nebo 1 , pokud jev nastal. O náhodné veličině, která nabývá dvou hodnot 0 a 1 se říká, že má alternativní rozdělení. [22]

1.1.15 Metody exploratorní analýzy dat

Tabulka rozdělení četností podává informaci o počtu (četnosti) výskytu jednotlivých variant znaku v souboru. Chceme-li mezi sebou porovnávat různá rozdělení četností lišící se svým rozsahem a dospět také ke snazší interpretaci výsledků, je vhodné převést *absolutní četnosti* na *relativní četnosti*.

(a) spojnicové a sloupkové grafy

Pro grafické znázornění prostého rozdělení četností se využívá *polygon četností*. Na ose x jsou hodnoty znaku (x_i) a na ose y jim odpovídající četnosti (n_i).

(b) bodové grafy

Bodové grafy používají jako grafické prostředky body umístěvané v souřadnicové soustavě.

(c) výsečové grafy

U výsečových grafů relativní četnosti obměn znaku znázorňujeme pomocí výsečí kruhu, které získáme rozdělením středového úhlu úměrně k podílu jednotlivých částí zobrazovaného jevu vyjádřených v procentech.

(d) graf STEM-and-LEAF

Tento graf je dalším znázorněním rozdělení četností obvyklým ve statistickém softwaru. První sloupec udává kumulativní absolutní četnosti od nejmenší hodnoty k mediánu (hodnota v závorce) a od největší hodnoty k mediánu. Počet číslic za čarou udává četnost příslušné obměny tarifní třídy.

(e) krabičkový graf

Tento graf se nejčastěji používá pro zobrazení kvartilů. Přehledně znázorňuje charakter analyzované proměnné pomocí kvartilů, vnitřních a vnějších hradeb a extrémů (minimum, maximum). Slouží k identifikaci odlehlých pozorování. Základním prvkem grafu je obdélník, jehož hrany tvoří hodnoty dolního a horního kvartilu, tzn., že uvnitř obdélníku je 50 % hodnot proměnné. Uvnitř je svislou čarou vyznačen medián a popř. tečkou aritmetický průměr. [23]

1.1.16 Naivní Bayesovský klasifikátor

Takovémuto bayesovskému klasifikátoru, který namísto skutečných hodnot využívá pouze jejich odhady, se říká "naive Bayes" (naivní Bayes). [24]

1.1.17 Neuronové sítě

Umělé neuronové sítě vycházejí z analogie s lidským mozkem. Podobně jako mozek jsou tvořeny množstvím navzájem propojených elementů; neuronů. V umělých neuronových sítích je neuron chápán jako buňka, která přijímá podněty od jiných neuronů, které jsou k

ní připojeny „na vstupu“. Pokud souhrnný účinek těchto vstupních podnětů překročí určitý práh, neuron se aktivuje a sám začne svým výstupem působit na další neurony. První modely neuronů a neuronových sítí se zkoumaly v rámci umělé inteligence již v 50. letech. Důležitá (z hlediska dobývání znalostí) je schopnost těchto modelů učit se z příkladů. Na rozdíl od stromů nebo pravidel, kde jsou nalezené znalosti srozumitelné uživateli, v neuronové síti jsou znalosti „rozprostřeny“ v podobě vah jednotlivých vazeb mezi neurony. Neuronová síť se vlastně chová jako černá skříňka; není příliš zřejmé, co se uvnitř děje. Složitější umělé neuronové sítě bývají tvořeny množstvím různě navzájem propojených neuronů. K nejznámějším typům umělých neuronových sítí (používaných pro klasifikaci) patří vícevrstvá síť. [5]

1.1.18 Rozhodovací pravidla

If-then konstrukce nalezneme ve všech programovacích jazycích, používají se i v běžné mluvě. Není tedy divu, že pravidla s touto syntaxí patří, vedle stromů k nejčastěji používaným prostředkům pro reprezentaci znalostí, ať už získaných od expertů, nebo vytvořených automatizovaně z dat. Jedním z nejznámějších algoritmu pro tvorbu pravidel je algoritmus pokrývání množin pracující metodou odděl a panuj (separate and conquer). Při pokrývání množin jde totiž o to nalézt pravidla, která pokrývají příklady téže třídy a oddělit je od příkladů třídy jiné. Pro naše data bychom našli pravidla uvedená na obr. 4. Použití těchto pravidel pro rozhodování o novém klientovi je opět velice jednoduché. Nalezneme první pravidlo, jehož předpokladům klient vyhovuje. Závěr tohoto pravidla pak určí, zda půjčit nebo ne. [5]

Tab. 1. Rozhodovací pravidla [5]

If konto=vysoké	then úvěr=ano
If příjem=vysoký	then úvěr=ano
If příjem=nízký & konto=střední & pohlaví=muž & nezaměstnaný=ne	then úvěr=ano

1.1.19 Rozhodovací stromy

Způsob reprezentování znalostí v podobě rozhodovacích stromů je dobře znám z řady oblastí. Vzpomeňme jen nejruznějších „klíčů k určování“ různých živočichů nebo rostlin

známých z biologie. Indukce rozhodovacích stromů patří k nejznámějším algoritmům z oblasti symbolických metod strojového učení. Při tvorbě rozhodovacího stromu se postupuje metodou „rozděl a panuj“ (separate and conquer). Trénovací data se postupně rozdělují na menší a menší podmnožiny tak, aby v těchto podmnožinách převládaly příklady jedné třídy. Použití rozhodovacích stromů pro klasifikaci odpovídá analogii s klíči k určování rostlin nebo živočichů. Od kořene stromu se na základě odpovědí na otázky (umístěné v nelistových uzlech) postupuje příslušnou větví stále hlouběji, až do listového uzlu, který odpovídá zařazení příkladu do třídy. [5]

1.1.20 Shluková analýza

Shluková analýza patří mezi metody učení bez učitele. Jejím cílem je v dané množině objektů nalézt její podmnožiny – shluky objektů – tak, aby si členové shluku byli navzájem podobní, ale nebyli si příliš podobní s objekty mimo tento shluk. [25, 26]

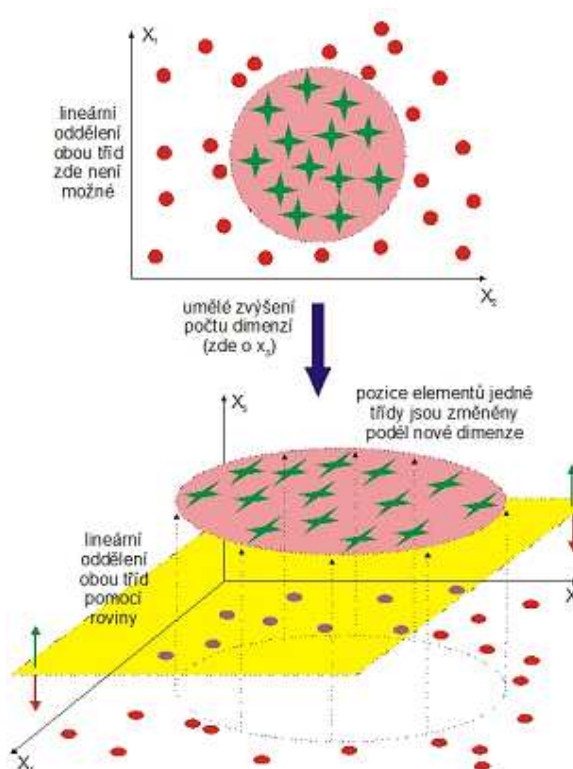


Obr. 3. Objekty ve dvojrozměrném prostoru: jedná se o 2 nebo 3 shluky? [26]

1.1.21 Support vector machines

K alternativním, relativně novým metodám patří podpůrné vektory (support vector machines, SVM), které tvoří určitou kategorii tzv. jádrových algoritmů (kernel machines). Tyto metody se snaží využít výhody poskytované efektivními algoritmy pro nalezení lineární hranice a zároveň jsou schopny representovat vysoce složité nelineární funkce. Jedním ze základních principů je převod daného původního vstupního prostoru do jiného, vícedimensionálního, kde již lze od sebe oddělit třídy lineárně.

Tato myšlenka je v podstatě jednoduchá, jak ukazuje obrázek obr. 6. V původním dvourozměrném prostoru jsou dvě třídy, oddělené nelineární kružnicí. Přidáním další dimenze vznikne možnost prvkům třídy uvnitř kružnice přidat další souřadnici, která je posune např. nahoru podél nové osy x_3 , takže pro oddělení obou tříd již lze použít rovinu rovnoběžnou s rovinou danou osami x_1 a x_2 . [27]



Obr. 4. Princip vzniku možnosti lineárního oddělení dvou tříd s nelineárními hranicemi pomocí přidané dimenze [27]

1.2 Výběr vhodné metody

Cílem této části práce bylo poskytnout přehled metod používaných pro data mining. Mnohé z nich se již úspěšně používají v oblasti inteligentního řízení energetických systémů, zejména neuronové sítě, metoda podpůrných vektorů (support vector machines), rozhodovací stromy, genetické algoritmy a kombinace prediktorů. Statisticky lze přínos kombinace více samostatných metod zdůvodnit tím, že konečné chyby jednotlivých metod lze rozdělit na chyby způsobené daty a chyby způsobené samotnou metodou.

Do první kategorie chyb patří např. odchylka predikčního systému způsobená náhlou netypickou změnou vstupních hodnot, popř. jejich zcela chybným zadáním. Tyto chyby se ve výsledné predikci odrazí vždy, nezávisle na typu či kvalitě použité metody.

Druhou kategorií jsou chyby způsobené metodou samou; cílem každého systému je tuto chybu minimalizovat. Vhodným způsobem minimalizace se jeví právě kombinace více metod. Zde lze uplatnit techniku lokální komparativní výhody některé z metod. Vyplývá-li např. ze statistické analýzy úspěšnosti jednotlivých metod, že některá z nich dosahuje nejlepších výsledků za specifických podmínek (v daném měsíci, za extrémních teplot nebo v odpoledních hodinách), bude tato metoda použita přednostně právě za již zmíněných podmínek. [1]

Zdůvodnění výběru vhodné metody je uvedeno na začátku praktické části, která pojednává o analýze získaných dat. Ze všech zmíněných metod data miningu v energetickém průmyslu jsem pro analýzu získaných dat využil neuronové sítě. Proto je potřebné znát aspoň teoretické minimum k pochopení principů, na jakých neuronové sítě pracují.

2 NEURONOVÉ SÍTĚ A PREDIKCE

Predikce neboli předpověď budoucích sledovaných veličin má obrovský význam nejen v energetickém průmyslu. Neuronové sítě mohou být použity pro predikci s různou mírou úspěchu. Jejich výhoda spočívá v automatickém učení závislostí jenom z naměřených dat bez toho, aby bylo zapotřebí přidávat další informace (jako typ závislosti u regrese apod.). Neuronová síť se trénuje na historických datech s cílem odhalit skryté závislosti a využít je pro predikování budoucnosti. Jinými slovy, neuronová síť nepředstavuje explicitně daný model. Je to spíše černá skříňka, která je schopna se něco naučit z dat. [28]

Modely neuronových sítí se často označují jako umělé neuronové sítě (ANN) na rozlišení od biologických neuronových sítí a taktéž na zdůraznění toho, že tyto modely mají vlastnosti umělé inteligence. Tabulka 2 ukazuje, jak se neuronová síť liší od počítače a počítačových programů.

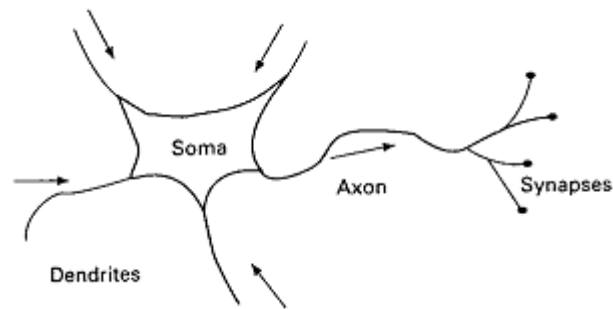
Tab. 2. Rozdíly mezi PC a neuronovou sítí [29]

Neuronová síť	Počítač
Je učena nastavováním vah, prahů a struktury	Je programován instrukcemi, (if, then, go to,...)
Paměťové a výkonné prvky jsou uspořádány spolu	Proces a paměť pro něj jsou separovány
Paralelismus	Sekvenčnost
Tolerují odchylky od originálních informací	Netolerují odchylky
Samoorganizace během učení	Neměnnost programu

Využití neuronových sítí je opravdu široké a nabývá čím dál tím více na významu. Lze je použít například na identifikaci radarových či sonarových signálů, predikci chování, klasifikaci, optimalizaci, filtraci a v mnoha dalších úkolech. [29]

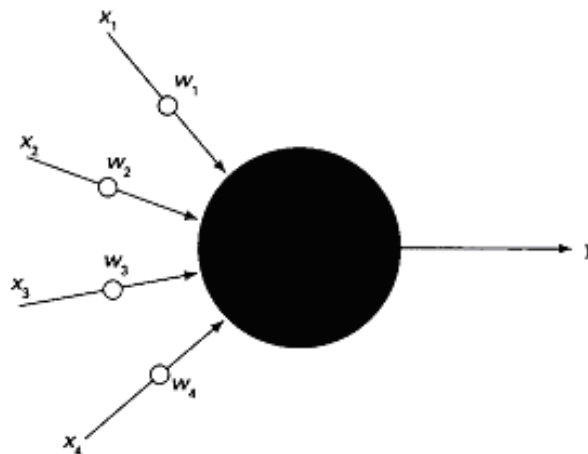
2.1 Model neuronu

Stavební jednotkou neuronové sítě je zjednodušený model organického neuronu. Lidský mozek obsahuje více než 10^{11} neuronů. Pro téměř všechny organické neurony se dají rozeznat anatomicky tři rozdílné části: řada přicházejících vláken (dendritů), buněčné tělo (soma) a jedno vycházející vlákno (axon). Axony se rozdělují na rozličná zakončení, z kterých každé tvoří kontakt s dalšími neurony. Neuron může přijímat až 10 000 vstupů od ostatních neuronů. Struktury, kde se vlákna spojují, se nazývají synapse. [30]



Obr. 5. Zjednodušený biologický neuron [30]

Model umělého neuronu je zobrazen na obr. 6. Skalární vstupy x se násobí skalárními váhami w a vytvoří $w.x$ a jsou zaslány do sumátoru. Odchylka b je taktéž vložena do sumátoru. Výstup sumátoru a , taktéž nazývaný jako síťový vstup, jde do transformační funkce f , která vytvoří skalární neuronový výstup y . [31]



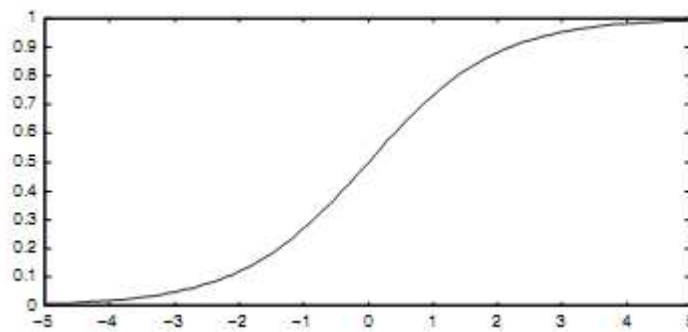
Obr. 6. Umělý model neuronu [30]

$$a = \sum_{i=1}^R w_i \cdot x + b \quad (2)$$

$$y = f(a) \quad (3)$$

Nejdůležitější transformační funkcí je log-sigmoid. Funkce logsig má hodnoty v intervalu mezi 0 a 1. [32, 33]

$$y = \frac{1}{1 + e^{-a}} \quad (4)$$



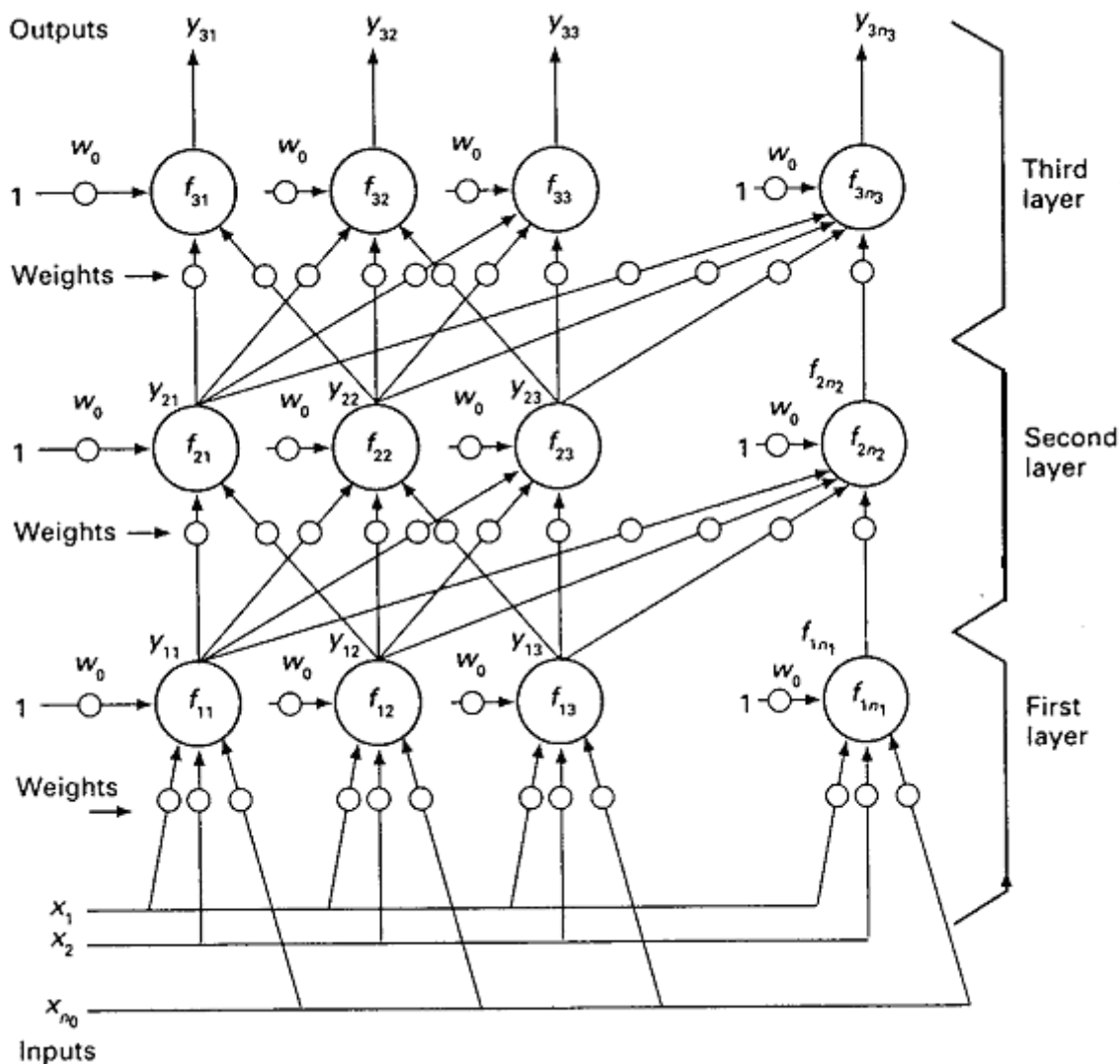
Obr. 7. Funkce log sigmoid [34]

2.2 Architektura neuronové sítě

Typická síť s dopředním šířením má neurony uspořádané v zřetelné vrstvé topologii. Vstupní vrstva slouží k zařazení hodnot vstupních proměnných. Neurony ve skryté a výstupní vrstvě jsou navzájem spojeny se všemi složkami v předchozí vrstvě. Je možné charakterizovat sítě, které jsou částečně spojeny jenom k některým jednotkám v předchozí vrstvě; ačkoliv pro většinu aplikací jsou lepší plně propojeny sítě.

Když se síť uvede v činnost, hodnoty vstupních proměnných jsou umístěny do vstupních jednotek a pak jednotky ve skryté a výstupní vrstvě postupně vykonávají svou činnost. Každá z nich spočítá svojí aktivační hodnotu tak, že od vážené sumy výstupů jednotek v předchozí vrstvě odečte práh (prahovou hodnotu). Aktivační hodnota projde skrz aktivační funkci a vytvoří se výstup neuronu.

Když se vykoná celá síť, výstupy z výstupní vrstvy slouží jako výstup celé sítě. Nejznámějším příkladem učícího algoritmu neuronových sítí je backpropagation. [35]



Obr. 8. Struktura vícevrstvé sítě se vstupní, skrytou a výstupní vrstvou [30]

Metoda backpropagation se používá pro vypočítání vah w . Skládá se ze dvou kroků. Nejdříve je potřebné vypočítat výstupy na základě vstupů a vah (dopředný krok). Dále se spočte chyba E jako suma čtverců rozdílů mezi výstupem y a očekávaným výstupem o pro všechny výstupy.

$$E = \sum (n_i - o_i)^2 \rightarrow \min \quad (5)$$

Hodnota chyby E je použita ve zpětné kalkulaci vah (zpětný krok). Proces se opakuje tak dlouho, dokud se hodnota E nepřibližuje požadované hodnotě. Problém učení je optimalizační úloha, kde funkce E musí být minimalizována. [32, 33]

2.3 Principy predikce s neuronovými sítěmi

Velkou výhodou neuronových sítí je schopnost učit se z příkladů a schopnost zachytit nelineární závislosti. Nevýhodou je, že obecně je nemožné odhadnout rozsah chyb anebo stanovit intervaly spolehlivosti předem. Teorie neuronových sítí neposkytuje žádné vodítko na tyto problémy a proto je většina z těchto odhadů výsledkem heuristických procedur. [36]

Doporučeným typem sítě pro predikci je vícevrstvá síť s algoritmem backpropagation. Neuronová síť s třemi vrstvami, která transformuje vstup na výstup v jednoduchém příkladě 2 vstupů a jednoho výstupu, obsahuje v první skryté vrstvě $2n+1$ neuronů a v další vrstvě $n \cdot (2n+1)$ neuronů, to znamená topologie je například 2 vstupy – 5 neuronů – 10 neuronů – 1 výstupní hodnota. Dalšími úpravami bylo zjištěno, že aplikace Kolmogorovova teorému na problematiku neuronových sítí vede pouze k existenčnímu důkazu, že k řešení libovolného problému stačí síť o třech vrstvách.

Když mluvíme o predikci, je to stochastická (pravděpodobná) predikce, protože nikdo nedokáže zaručit, že se systém bude vyvíjet tak, jak jsme předpověděli. To, zda daný systém bude předpověditelný či ne, záleží na několika kritériích, např. na tom, jestli je systém chaotický anebo deterministický. Dalším problémem je integrita dat. Pokud nejsou trénovací data připravena dobře, výsledky budou k ničemu. Neuronové sítě mají oproti standardním technikám jako AR, MA, ARIMA tu výhodu, že nepotřebují ke své činnosti model a navíc jsou „tolerantní“ k šumu v dané časové řadě. Pojmeme tolerantní zde musíme rozumět to, že i při zašuměné řadě jsou schopny dát rozumné výsledky v porovnání s klasickými metodami.

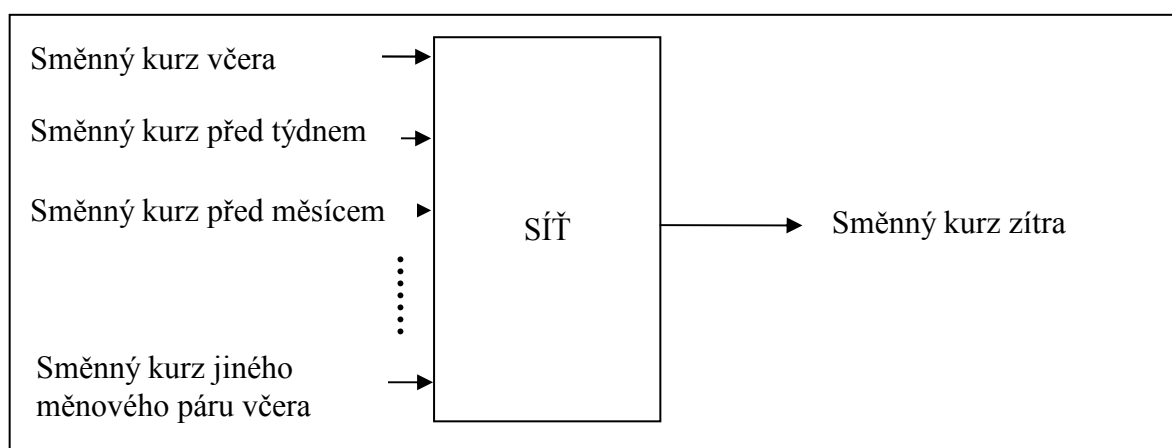
Trénovací množina dat by měla obsahovat vstupní a výstupní vektory s historickými daty pro predikci. Data v tréninkovém výstupním vektoru jsou posunuta oproti vstupům o tolik časových jednotek, o kolik chceme predikovat z historie do budoucnosti (např. 3 dny dopředu). Predikční interval je délka sekce jednoho výstupního vektoru. To znamená délka predikce pro jeden podnět. Pokud máme např. jako vstup do sítě vektor o délce dvaceti neuronů-dnů a na výstupu sítě je pouze jeden neuron-den posunut vzhledem k poslednímu vstupnímu prvku o tři dny, pak mluvíme o jednodenní predikci s předstihem o tři dny. Pokud je na výstupu např. pět neuronů, pak mluvíme o tzv. pětidenní predikci, atd.

Z předchozího popisu je zřejmé, že existuje mnoho kombinací jak predikovat z časové řady. Počet výstupných vektorů určuje taktéž posun vektorů mezi sebou navzájem. Pokud je první predikovaný vektor v intervalu 1.1. – 5.1., druhý bude 6.1. - 10.1 a tak dále.

Někdy se pro predikci používají také diference, které jsou rozdílem mezi hodnotou poslední a hodnotou nynější. V takovém případě odpadají starosti s eliminací trendu a sezónními variacemi.

2.3.1 Vícenásobná predikce a autopredikce

Predikce chování na základě historie jednoho údaje je možná, ale dosahuje nejhorších výsledků. Mnohem lepší výsledek dostaneme, pokud predikujeme na základě historie více údajů o daném systému. Obecně platí, že čím více informací o historickém chování systému pro predikci máme, tím lepší predikci můžeme obdržet. Příklad vícenásobné predikce je na obr. 9.



Obr. 9. Vícenásobná predikce kurzů měn [29]

Existují dva druhy predikce – klasická predikce a autopredikce. Autopredikce nevyužívá nové naměřené nebo pozorované skutečné vstupní hodnoty, ale predikuje hodnoty z naší sítě jako nové vstupy. Další predikce je tak zatížená predikční chybou, která negativně ovlivňuje výsledky. Je vhodné nepoužívat autopredikci na vzdálenou predikci. [29]

2.3.2 Predikční chyba

Každá predikce se musí vyhodnotit z pohledu úspěšnosti. V praktické části jsem vyhodnocoval výsledky pomocí kritéria MAPE (průměrná absolutní procentuelní chyba):

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right| \quad (6)$$

kde A_t je skutečná hodnota, F_t předpovězená hodnota, n je počet predikovaných hodnot. Výsledné číslo ukazuje průměrnou absolutní procentuální chybu, kterou je zatížen každý bod predikce. [37]

II. PRAKTICKÁ ČÁST

3 ANALÝZA POUŽITÝCH DAT A VÝPOČETNÍCH PROSTŘEDKŮ

V této části popíšu použitý software a data, s kterými jsem pracoval, rovněž uvedu pár informací o vybraném energetickém provozu - elektrárně Komořany.

Neuronové sítě jsem zvolil jako metodu data miningu z více důvodů. Prvním z nich byl, že jsem už něco o neuronových sítích věděl a chtěl jsem tyto poznatky použít v praxi. Druhým důvodem bylo doporučení vedoucího bakalářské práce. V teoretické části jsem rovněž ověřil informaci, podle které se neuronové sítě v energetických provozech používají zcela běžně. Dostupnost vhodného softwaru byla taky rozhodující – Matlab v sobě obsahuje mocný nástroj pro práci s neuronovými sítěmi. Posledním důvodem je zvědavost – zda se podaří na rozsáhlých datech využít software efektivně a úspěšně.

3.1 Elektrárna Most - Komořany

Data pro praktickou část byly získány v úzké spolupráci s elektrárnou v Komořanech. V současné době probíhá v ČR vědeckovýzkumná činnost pro několik energetických společností na více pracovištích v rámci grantu „GAČR č. 101/06/0920 Vývoj a využití řídicích algoritmů vyšší úrovně pro řízení teplotenských soustav jako nástroje pro snižování cen energií a zlepšování životního prostředí“.

3.1.1 Historie elektrárny Komořany

Komořanská elektrárna, největší a nejdůležitější část současné společnosti United Energy, a.s., existuje od 50. let minulého století. Od té doby několikrát změnila majitele a stala se součástí různých společností. Od svého vzniku je ale významným nezávislým producentem tepla a elektrické energie na severu Čech.

Elektrárna Komořany se začala stavět ještě během druhé světové války v roce 1943 a stavba pak pokračovala v rámci národního podniku Mostecko-sokolovských elektráren i po roce 1945. První kotel a turbogenerátor byly uvedeny do provozu v roce 1951. V padesátých letech bylo uvedeno do provozu pět kotlů o výkonu 90 tun páry za hodinu a tři kondenzační turbosoustrojí po 32 MW. Od počátku byla elektrárna zásobována hnědým uhlím přímo ze sousední úpravny. Od roku 1952 byla elektrárna Komořany součástí národního podniku Mostecké elektrárny, později jako samostatný podnik Elektrárna Komořany.

Investiční akce pod zkratkou EKY III započala v roce 1963 a jejím cílem bylo zahájení teplárenské éry Komořan. Byl postaven nový teplárenský protitlaký stroj TG 8, výměňková stanice, horkovod do Mostu a také 180 metrů vysoký komín. První teplo začalo proudit do Mostu v roce 1964 a o dva roky později byla dokončena stavba komínu, který zlepšil ovzduší v těsné blízkosti elektrárny.

V sedmdesátých letech přechází pod hlavičku tehdejších Severočeských elektráren, jejich hlavní součástí jsou Komořany, provoz veškerých rozvodů tepla v celém kraji. K elektrárně jsou přiřčeny výtopny v Lounech, Teplicích a Bílině. Současně se pracuje na výstavbě dalších horkovodů z Komořan - do Chomutova (uveden do provozu v roce 1977) a Litvínova (1978).

V roce 1982 nastává pro Komořany zlom - začíná převládat teplárenský provoz nad výrobou elektřiny, o devět let později se dokonce podle toho změnil i název společnosti na teplárnu.

Od března 2009 společnost mění název na United Energy, a.s. Nadále provozuje elektrárnu v Komořanech a navazující rozvody tepla pro Most a Litvínov.

3.1.2 Teplo, elektřina a vedlejší produkty

Zdrojem tepelné energie je teplárna v Komořanech u Mostu s kombinovanou výrobou tepla a elektřiny. Kotelna s deseti kotli, kterou zásobuje hnědým uhlím sousední úpravna uhlí Mostecké uhelné, a.s., má instalovaný výkon 974 MWt pro výrobu páry. Instalovaný výkon strojovny pro výrobu tepla je 505,89 MWt.

Teplem z Komořan jsou zásobovány průmyslové areály a města Most a Litvínov. Primární rozvody centrálního zásobování teplem (CZT) v majetku společnosti pro tyto lokality měří více než 77 kilometrů. United Energy zajišťuje zásobování 35 tisíc bytových jednotek v Mostě a Litvínově a dále školských a zdravotnických zařízení, úřadů, obchodů a průmyslových podniků.

Elektrická energie je vyráběna v teplárně Komořany na 8 turbogenerátorech o celkovém instalovaném elektrickém výkonu 239 MWe. United Energy v roce 2005 založila 100% dceřinou společnost United Energy Trading, a.s., prostřednictvím které je realizován prodej elektrické energie konečným zákazníkům.

Při výrobě tepla a elektrické energie v elektrárně Komořany vznikají vedlejší produkty. Původně odpadní látky jsou nyní využity pro výrobu certifikovaných výrobků, vhodných pro stavebnictví či k zahlazování důlních děl. Popílek vzniklý fluidním spalováním uhlí a zachycený látkovými filtry nebo směs popelovin, tzv. aditivovaný granulát je distribuován po železnici nebo nákladními auty. [38]

3.1.3 Analýza získaných a použitých dat

Data pro aplikaci data miningu jsem obdržel ve formě jednoho rozsáhlého souboru ve formátu Microsoft Excel 2003 (.xls). Soubor má velikost 12 MB a obsahuje tři časové řady, které jsou odděleny na samostatných listech – roky 2005, 2006 a 2007. List 2005 obsahuje celkem 2212 řádků a 58 sloupců. To je celkem 128 296 údajů včetně záhlaví tabulky. List 2006 je ještě rozsáhlejší – 8763 řádků, 58 sloupců, to je 508 254 údajů. Poslední list 2007 má stejný datový rozsah jako list 2006. Spolu je to za tři roky 1 144 804 údajů.

Několik desítek buněk lokálně neobsahuje data – v souboru je menší neúplnost dat. Po prvním náhledu na rozsah dat můžeme konstatovat, že se jedná o data mining v pravém slova smyslu. Časová řada obsahuje 58 zaznamenaných proměnných, z nichž jsou první čtyři časové – den, měsíc, hodina a časový údaj v plném rozsahu. Další proměnné se týkají deseti kotlů (t/h), sedmi turbogenerátorů (MW), výroby, vlastní spotřeby a dodávky (MW), atmosférických teplot ve městech Most a Litvínov, průtoku, teploty topné vody a tlaku topné a vratné vody na jednotlivých výměňkových stanicích.

Pokud sečteme počet řádků souboru, dostaneme přibližně 19 738 zaznamenaných stavů energetického systému. Data jsou zaznamenány s periodou jedné hodiny. To je přibližně 822 celých dnů po 24 hodinách.

V roce 2005 jsou data časově od 1.10.2005 do 31.12.2005. Další roky jsou již datově pokryty kompletně, tzn. od 1.1.2006 do 31.12.2007.

Z množství sloupců (veličin) a řádků (naměřených stavů s periodou jedna hodina) bylo potřeba vybrat ty, které se použijí pro predikci a pro trénování neuronové sítě.

Z naměřených veličin jsem vybral tyto 4 hlavní:

t_{ex} Most [°C]... teplota externí ve městě Most ve stupních Celsia

G Most [t/h]... množství dodané teplé vody v tunách za hodinu

Ttv Most [°C]... teplota topné vody dodané do města Most ve stupních Celsia

Tvv Most [°C]... teplota vratné vody z města Most ve stupních Celsia

Další veličiny byly časové – měsíc v roce (1-12), den (1-31 dle počtu v měsíci), hodina (1-24) a den v týdnu (pondělí-neděle).

Popis dat pro rok 2005 je uvedený v tabulce 3. Časově jsou data ohraničeny od 1. 10. 2005 do 31. 12. 2005. Celkem se jedná o 2208 naměřených stavů. Data byly určeny pro učení neuronové sítě.

Tab. 3. Popis použitých dat – rok 2005

2005	Minimum	Maximum	Průměr
t_ex Most [°C]	-10,1	20,2	5,3
G Most [t/h]	572,6	2 280,8	1 435,1
Ttv Most [°C]	88,2	131,1	109,7
Tvv Most [°C]	52,5	69,7	58,0

Popis dat pro rok 2006 je uvedený v tabulce 4. Časově jsou data ohraničeny od 1. 1. 2006 do 31. 3. 2006 a od 1. 10. 2006 do 31. 12. 2006 (topná sezóna). Celkem se jedná o 4368 naměřených stavů. Data byly určeny pro učení neuronové sítě.

Tab. 4. Popis použitých dat – rok 2006

2006	Minimum	Maximum	Průměr
t_ex Most [°C]	-13,9	20,4	3,5
G Most [t/h]	44,6	2 583,2	1 441,9
Ttv Most [°C]	77,4	137,6	116,2
Tvv Most [°C]	48,1	71,7	60,2

Popis dat pro rok 2007 je uvedený v tabulce 5. Časově jsou data ohraničeny od 1. 1. 2007 do 31. 3. 2007. Celkem se jedná o 2160 naměřených stavů. Data byly určeny pro učení neuronové sítě.

Tab. 5. Popis použitých dat – 1. část roku 2007

2007	Minimum	Maximum	Průměr
t_ex Most [°C]	-6,4	15,5	5,7
G Most [t/h]	698,6	2 031,2	1 291,6
Ttv Most [°C]	99,5	134,1	114,3
Tvv Most [°C]	56,1	68,3	60,8

Pro testování výkonnosti naučené neuronové sítě byly použity data vyčleněné z druhé části roku 2007. Jejich popis je uvedený v tabulce 6. Časově jsou data ohraničeny od 1. 10. 2007 do 31. 12. 2007. Celkem se jedná o 2208 naměřených stavů.

Tab. 6. Popis použitých dat – 2. část roku 2007

2007 testovací	Minimum	Maximum	Průměr
t_ex Most [°C]	-6,2	18,9	4,5
G Most [t/h]	551,1	1 975,3	1 286,0
Ttv Most [°C]	91,8	134,3	114,9
Tvv Most [°C]	55,3	68,9	61,0

Pro shrnutí uvádím, že pro učení neuronových sítí jsem vymezil základní soubor 8736 stavů s periodou jedné hodiny pro 4 hlavní proměnné. Pro testování výkonnosti predikce je k dispozici soubor obsahující 2208 naměřených stavů.

3.2 Microsoft Office Excel 2007 – příprava dat

Z předchozího popisu dat vyplývá nutnost připravit data tak, aby byly vhodné na predikci a import do programu pro využití neuronových sítí.

Prvním krokem bylo prozkoumání dat z pohledu úplnosti. Všechny data ve sloupcích, které jsem potřeboval pro predikci, byly úplné.

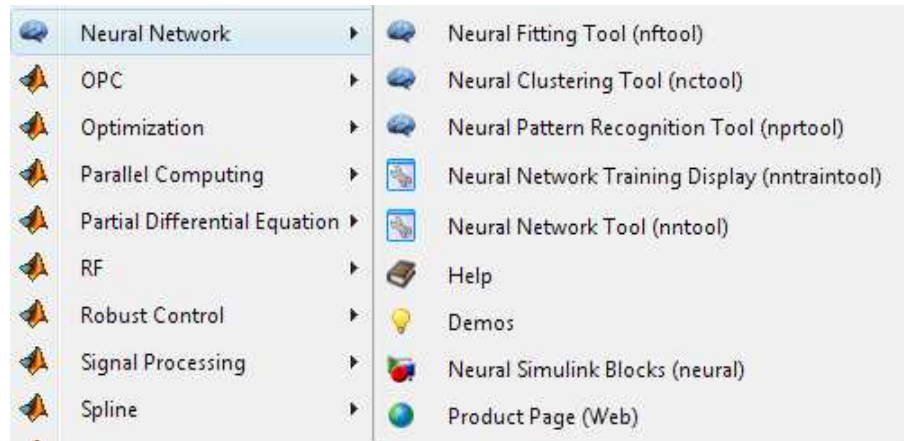
V řádcích nebyly data zcela korektní. Pro predikci jsem potřeboval doplnit sloupec den v týdnu pro každé měření. Při kopírování smyčky pondělí-neděle po řádcích se někdy nepotkal konec dne s danou hodinou. Problém byl v chybějící hodině (řádku), anebo naopak v opakování řádku s tím samým údajem. Nadbytečné řádky jsem smazal, chybějící jsem doplnil průměrem hodnot z předchozí a následné hodiny. Celkově bylo těchto úprav přibližně 20, co by nemělo v tak rozsáhlém souboru negativně ovlivnit predikci.

Přípravu dat pro import jsem prováděl kopírováním do samostatných souborů z původního. Využíval jsem filtr pro výběr rozsahu (např. úterý-pátek) a transformaci matic tak, abych získal data v potřebné formě a kvalitě. Pro zautomatizování práce s rozsáhlými daty jsem používal s filtrem i funkci CONCATENATE pro spojování rozsáhlých řetězců.

Vstupní soubor ve formátu Excel 2003 (přípona .xls) měl 2 listy: v prvním byly vstupy do sítě a v druhém požadované výstupy. Na nich se síť učila a trénovala.

3.3 Matlab R2009a a Neural Network Toolbox – použití dat

K predikci pomocí neuronových sítí jsem použil programové prostředí Matlab ve verzi 7.8.0.347 (R2009a). Neural Network Toolbox byl ve verzi 6.0.2.



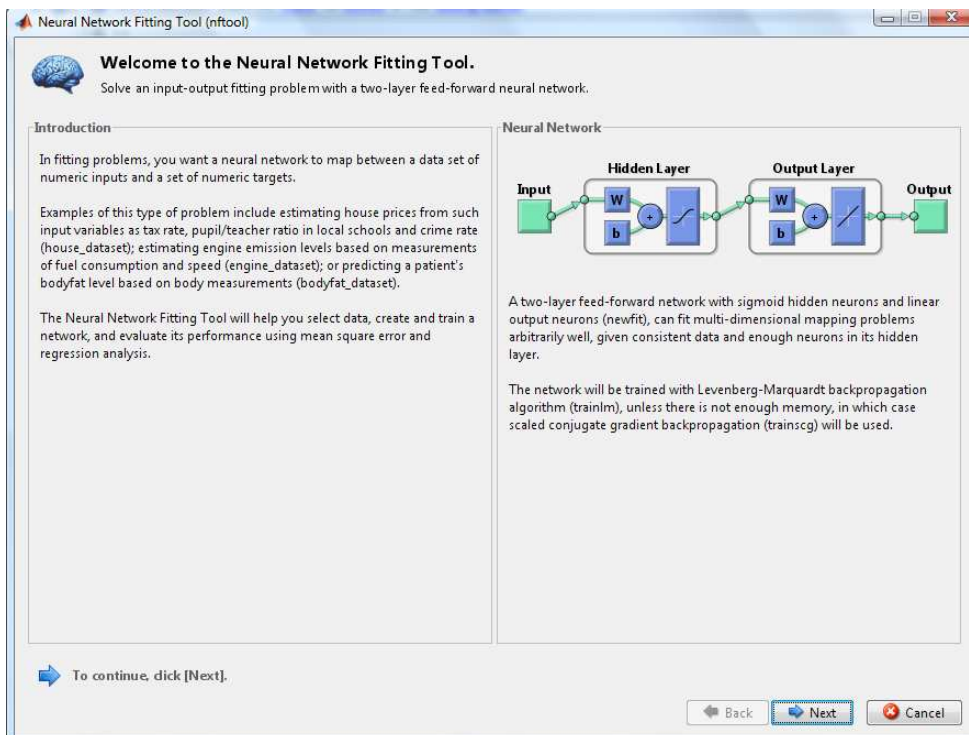
Obr. 10. Start – Toolboxes – Neural Network

Po spuštění Matlabu je toolbox pro práci s neuronovými sítěmi přístupný z menu *Start* (obr. 10). Toolbox obsahuje několik nástrojů, pro predikci jsem využil dva – *nftool* a *nntool*.

3.3.1 Import dat do neuronové sítě

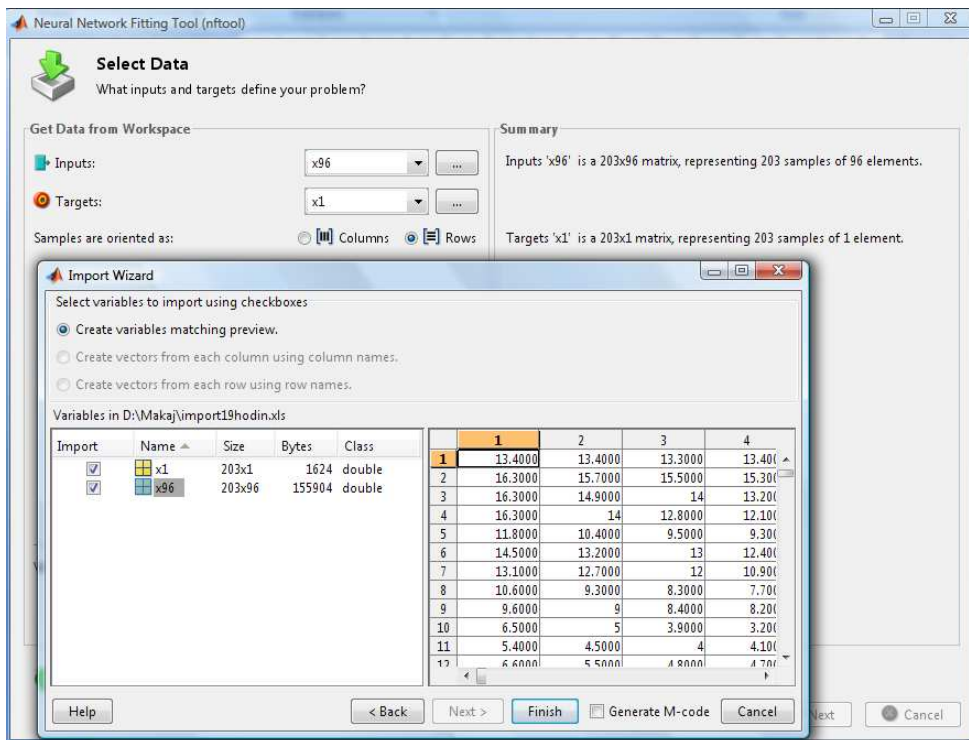
Pro import dat spustíme nástroj *nftool* (obr. 11). V okně se kromě jiného uvádí následující popis:

Neural Network Fitting Tool pomůže s výběrem dat, vytvoří a trénuje síť a ohodnotí její výkonnost pomocí střední kvadratické chyby a regresní analýzy. Dvouvrstvá síť s dopředním šířením se skrytými neurony pro funkci sigmoid a lineárními výstupními neurony může být vhodná pro mnohorozměrné problémy libovolně dobře, pokud do ní dodáme konzistentní data a dostatek neuronů v její skryté vrstvě. Síť bude učena algoritmem Levenberg-Marquardt backpropagation, dokud bude dostatek volné paměti.



Obr. 11. Nástroj nftool po spuštění

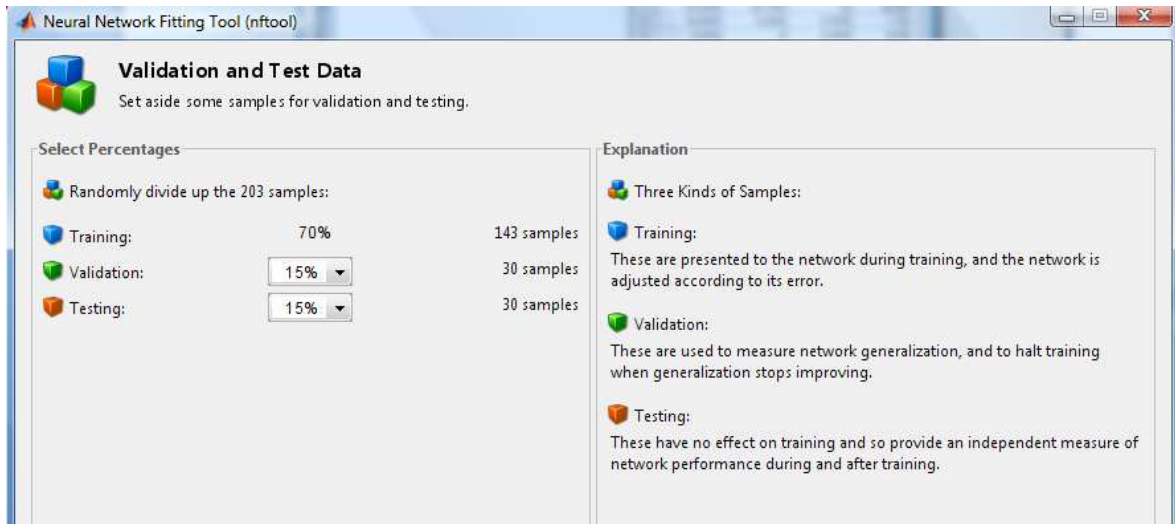
Následuje okno pro výběr dat (obr. 12). Data jsem měl nachystané v samostatném souboru ve formátu Excel 2003 (*.xls). Na jednom listu byly vstupné proměnné, na druhém žádoucí výstupní proměnná (target). Na těchto datech se síť učí a trénuje.



Obr. 12. Výběr dat pro import

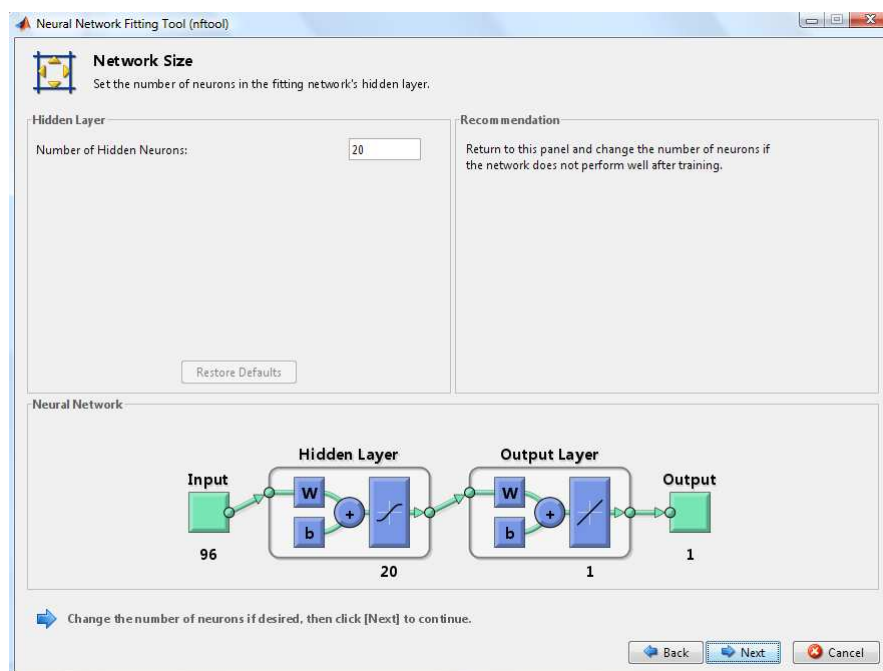
3.3.2 Učení sítě na historických datech (*nftool*)

V dalším okně můžeme data procentuelně rozdělit na trénovací, validační a testovací. Výchozí rozdělení je na obr. 13. Trénovací data dostane síť v průběhu tréninku a nastaví se podle nich. Validace data slouží k měření schopnosti sítě zobecňovat. Trénování se zastaví, když se zobecňování už nezlepšuje. Testovací data nemají vliv na trénink a tak poskytují nezávislé měřítko výkonnosti sítě v průběhu a po skončení tréninku.



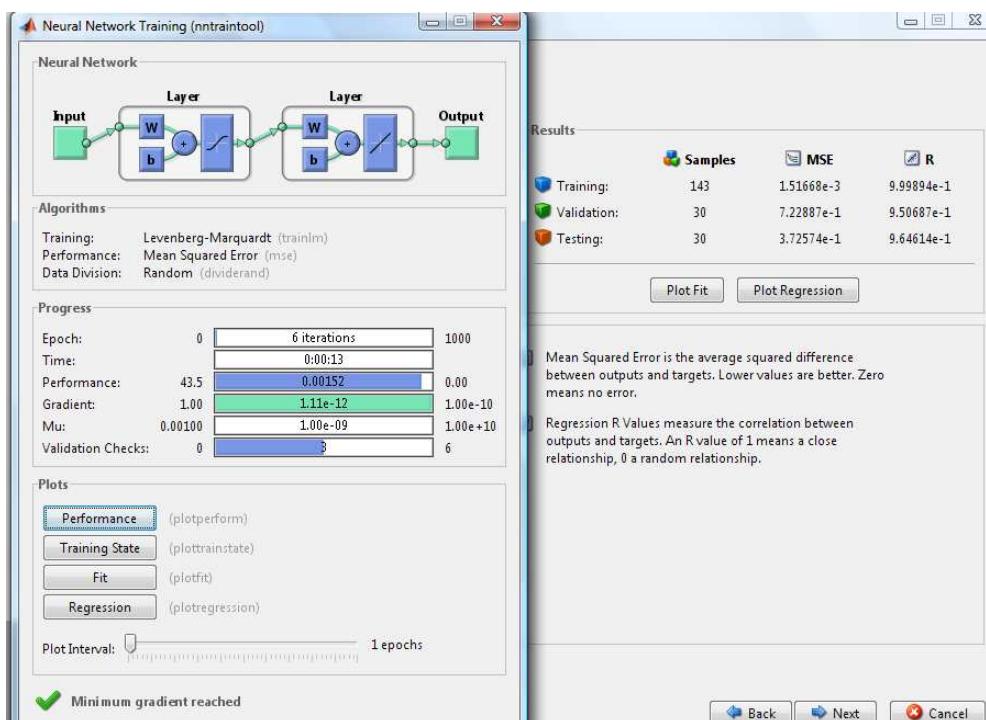
Obr. 13. Rozdělení dat na trénovací, validační a testovací

Po rozdělení dat definujeme velikost sítě. Výchozí nastavení je 20 neuronů (obr. 14).



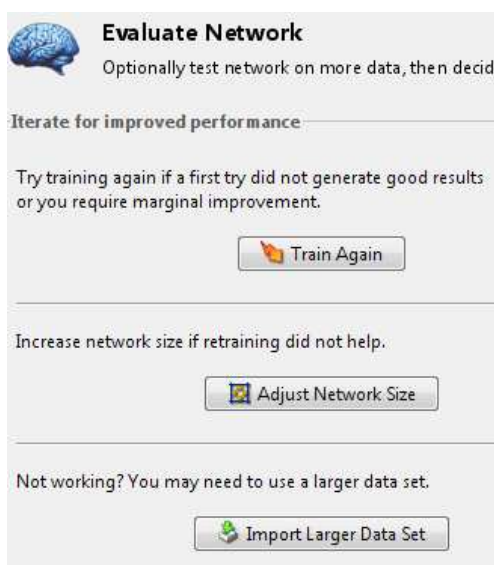
Obr. 14. Nastavení počtu neuronů v síti

Pak následuje samotný trénink neuronové sítě na datech. Probíhá v epochách a iteracích v nástroji Neural Network Training (nntraintool). Po skončení tréninku se automaticky zavře a v pravé části můžeme vidět výsledky tréninku v číslech – střední kvadratická chyba (MSE) a regrese (R) – obr. 15.



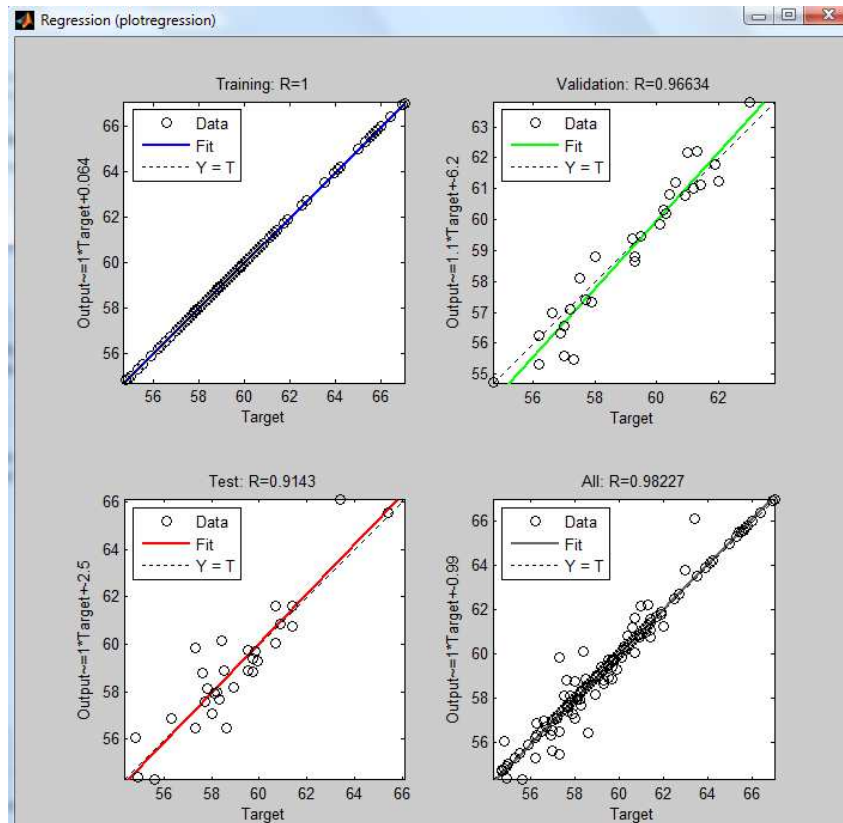
Obr. 15. Trénink sítě a výsledky tréninku

Pokud nejsme s výsledky tréninku spokojeni, můžeme opakovat trénink za nezměněných podmínek, změnit velikost sítě nebo importovat větší sadu dat (obr. 16).



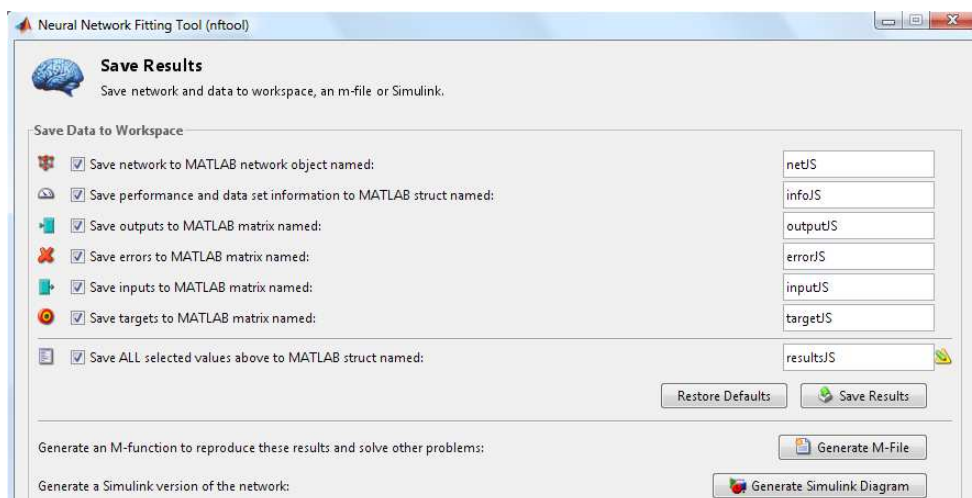
Obr. 16. Opakování tréninku

Ve výsledcích tréninku máme možnost vizualizace regrese (obr. 17). Zobrazuje se pro tři typy dat (trénovací, validační, testovací) a pro všechna data celkově. Čím blíže R dosahuje hodnoty 1, tím lépe je síť trénovaná na daná data.



Obr. 17. Vizualizace regrese

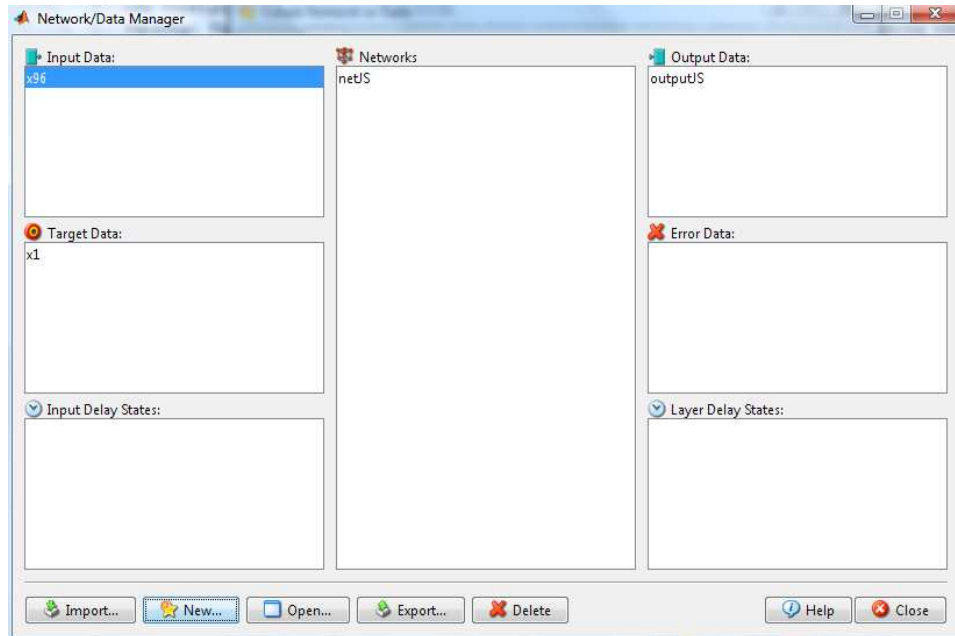
Poslední důležitou činností před samotnou predikcí je uložení vytrénované sítě do prostředí Matlab, aby byla k dispozici pro další práci.



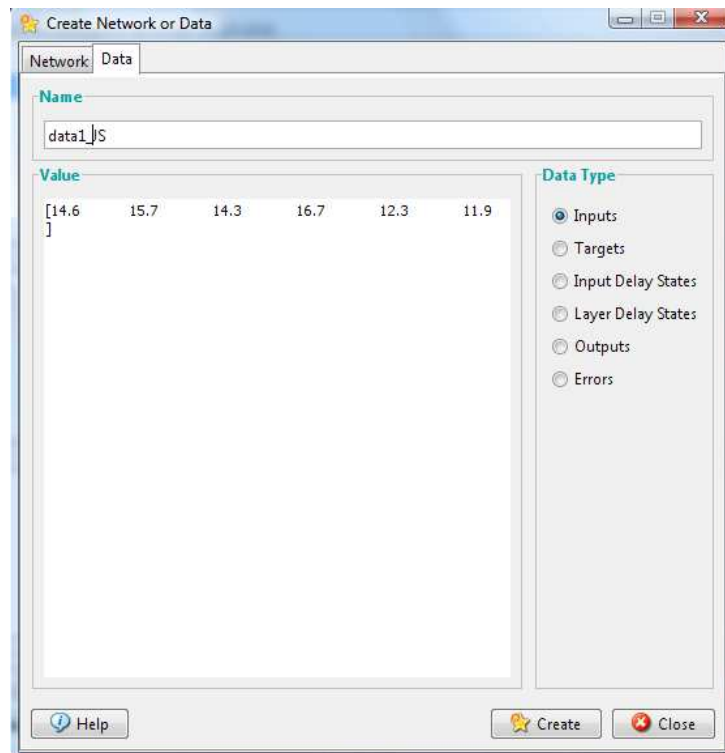
Obr. 18. Uložení sítě a výsledků tréninku

3.3.3 Postup predikce (*nntool*)

Pro predikci si spustíme nástroj *nntool* (obr. 19). Tlačítkem *Import* importujeme dříve vytvořenou síť. Můžeme importovat i další vektory – vstupy, výstupy, žádané výstupy (targety), chyby a zpoždění.



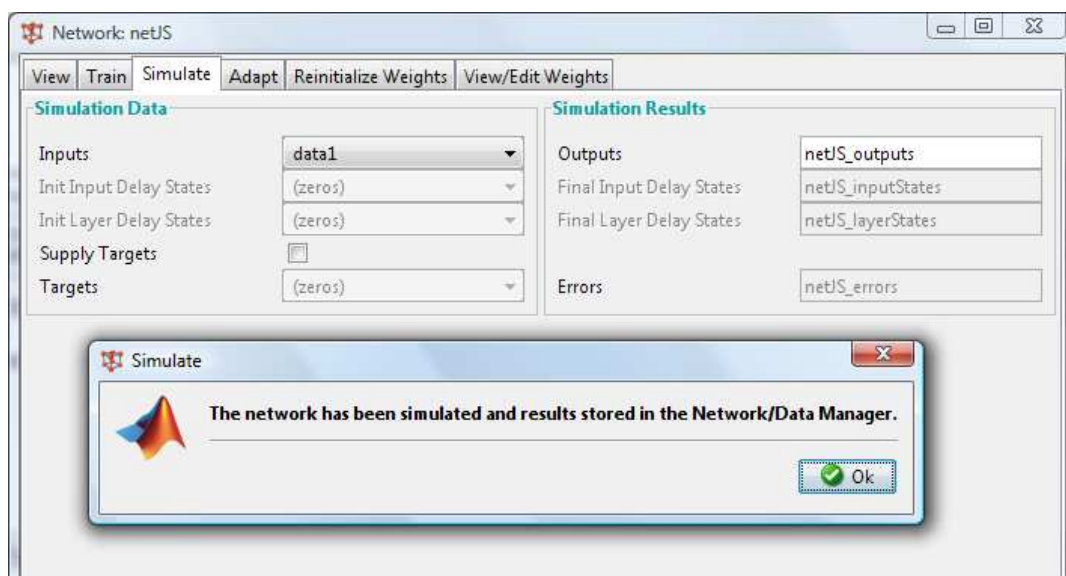
Obr. 19. Prostředí pro práci se sítěma



Obr. 20. Vytvoření nového vstupu do sítě

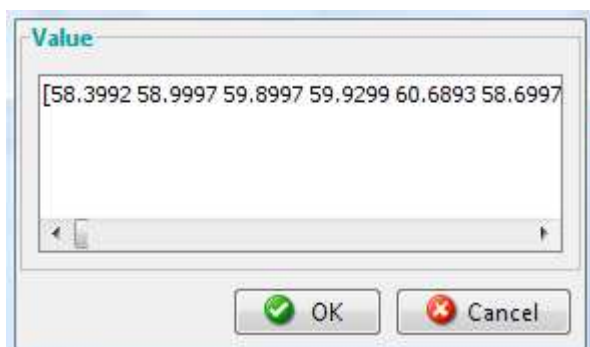
Tlačítkem *New* vytvoříme nový vstupní vektor, pro který požadujeme od sítě odpověď – v našem případě predikci. Vstupní data se zadávají do hranatých závorek, jednotlivé proměnné se oddělují středníkem (obr. 20). Čísla musí mít místo desetinné čárky desetinnou tečku. Lze to vyřešit například nahrazením v *Excelu* anebo v *Poznámkovém bloku*. Vstupní proměnnou můžeme pojmenovat.

Simulace sítě (predikce) se spustí kliknutím na tlačítko *Open* při označení příslušné sítě. V záložce *Simulate* vybereme *Inputs* (vstupy), pro které chceme predikovat. V kolonce *Outputs* pojmenujeme výstupní vektor (obr. 21). Pokud jsou data zadána správně, obdržíme informační hlášku, v opačném případě výstrahu a simulace se neprovede.



Obr. 21. Predikce v síti pomocí simulace

Výsledky simulace získáme tak, že otevřeme nový vektor, který jsme zadali v simulátoru (obr.22). Pokud je výstupní vektor příliš velký, doporučuji data zkopírovat zpátky do *Excelu*, zpětně nahradit desetinné tečky za čárky a dál s predikovanými daty pracovat.



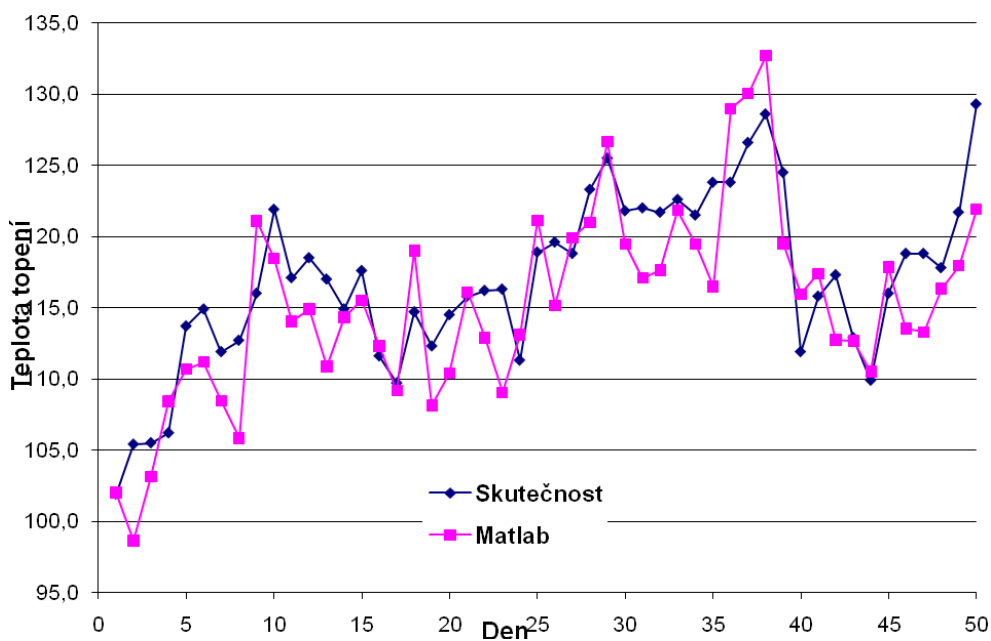
Obr. 22. Výslední vektor predikce

4 VÝSLEDKY PREDIKCE V SYSTÉMU ŘÍZENÍ DISTRIBUCE TEPLA MĚSTSKÉ AGLOMERACE POUŽITÍM NEURONOVÝCH SÍTÍ

4.1 Predikce teploty topné vody

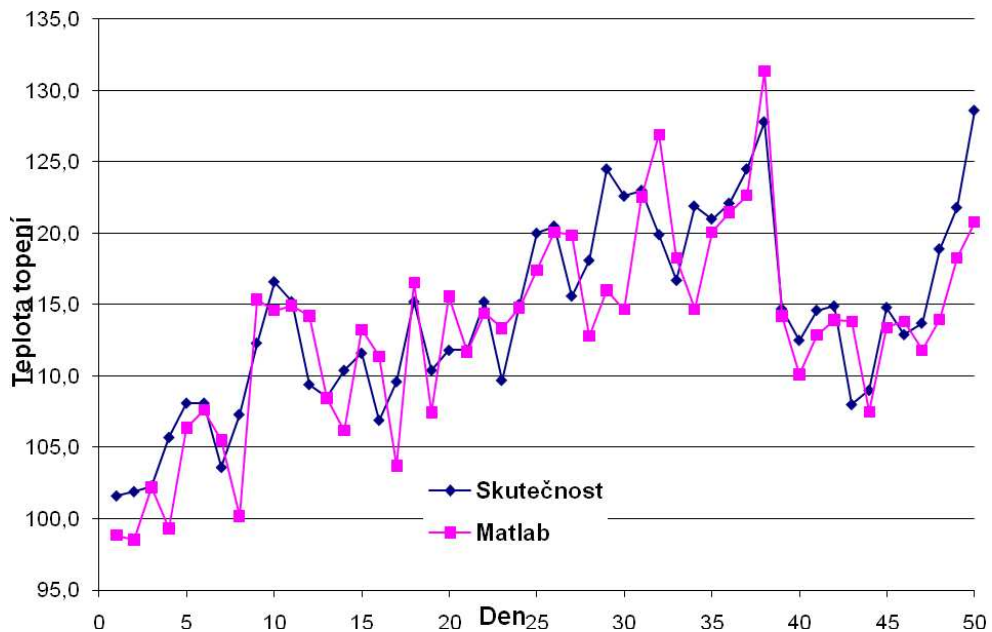
Z obrovského počtu možností nastavení predikce jsem vybral několik ukázek. V této první části se predikuje proměnná T_{tv} Most dnešní – teplota topné vody pro město Most dnes v °C.

První predikce (obr. 23) byla vykonána pomocí neuronové sítě, která se trénovala na datech od 1. 1. 2007 do 8. 10. 2007, na celkem 200 hodnotách (dní) naměřených proměnných. Proměnné byly měřeny vždy pouze v čase 1h v noci a pouze pro dny pracovní (pondělí-pátek). Nebyly zohledněny státem uznané svátky. Vstupem do sítě byla teplota topné vody včerejší v 1h, teplota externí ve městě Most v 1h, rozdíl teploty externí včera a dnes v 1h. Celkem tedy 3 vstupní proměnné. Cílem je předpověď dnešní teploty topné vody tak, jak by jí nastavil zkušený operátor. Graf zachycuje porovnání skutečně nastavené teploty operátorem a predikované teploty, kterou by nastavila neuronová síť. Rozdíl je průměrně 2,81%. Předpověď je okamžitá jednodenní (jeden krok dopředu) pro celkem 50 dnů.



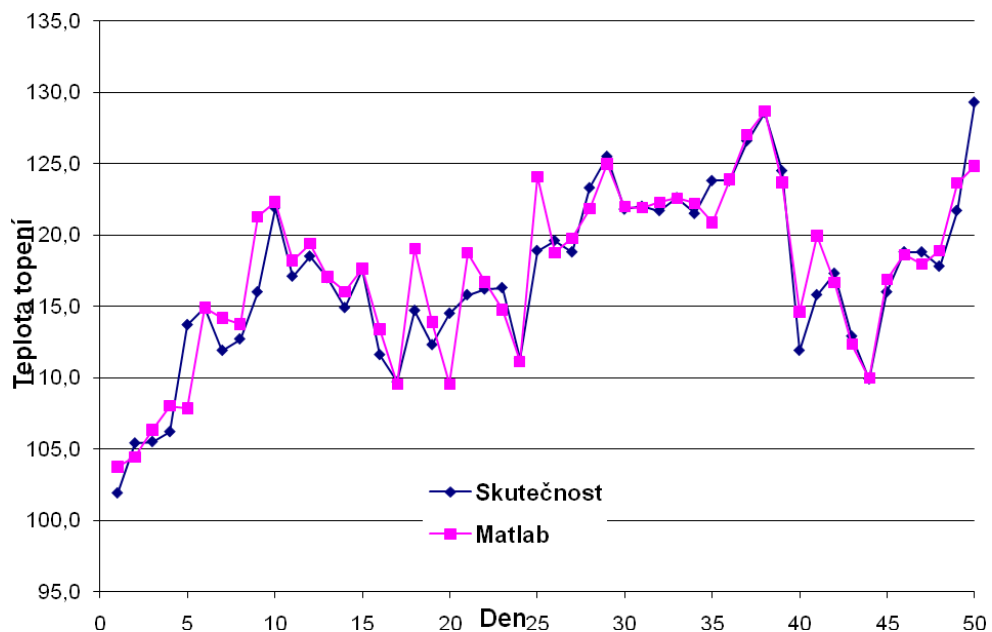
Obr. 23. Predikce teploty topné vody v 1h (MAPE=0,0281)

V další ukázce predikce (obr. 24) zůstali všechny parametry pro trénování a predikci stejné kromě jednoho – data se vztahují k hodině 11h dopoledne. Chyba je průměrně 2,59%.



Obr. 24. Predikce teploty topné vody v 11h (MAPE=0,0259)

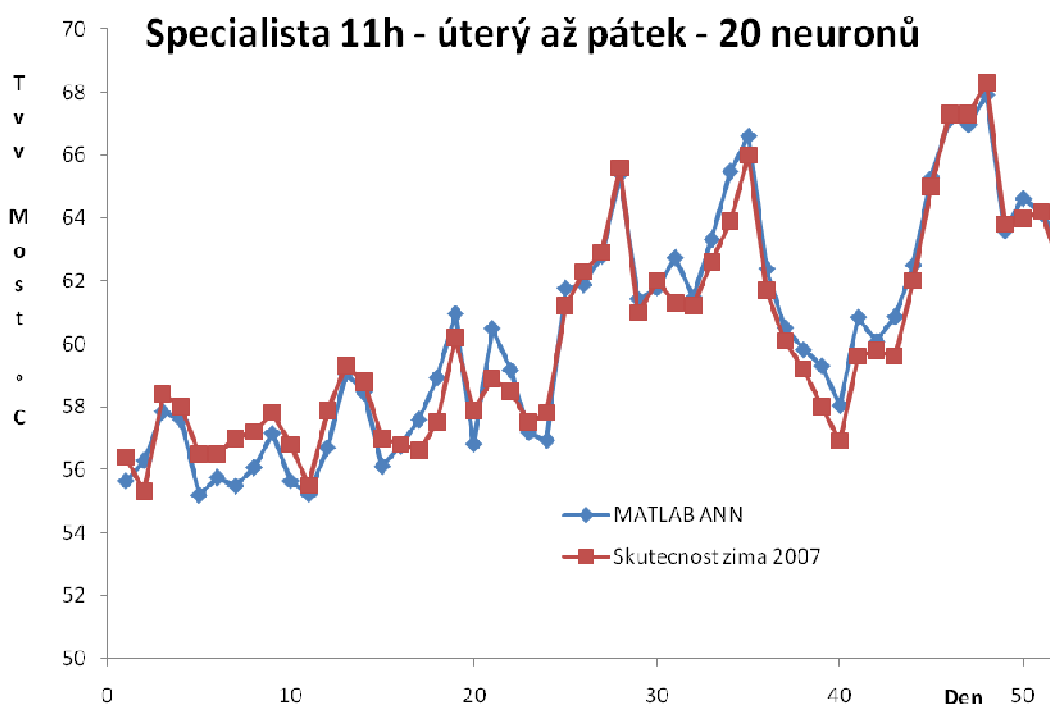
V poslední ukázce v této části byla opět predikována 1h v noci na stejných datech. Zvýšil se ale počet známých vstupných parametrů o teplotu vratné vody dnes a o teplotu topné vody před hodinou. Průměrná chyba predikce je 1,28%.



Obr. 25. Predikce teploty topné vody v 1h – 6 parametrů (MAPE=0,0128)

4.2 Specializované sítě

V této části jsem změnil predikovanou proměnnou. *T_{vv} Most* je teplota vratné vody z města Most, čili voda, která se vrací z aglomerace. Neměla by být příliš studená (málo se topí) ani příliš horká (zbytečně se topí). Skutečnost, která může negativně ovlivnit predikci je dopravní zpoždění. V různých částech dne proudí voda systémem vždy jinak a nedá se počítat s konstantním zpožděním.

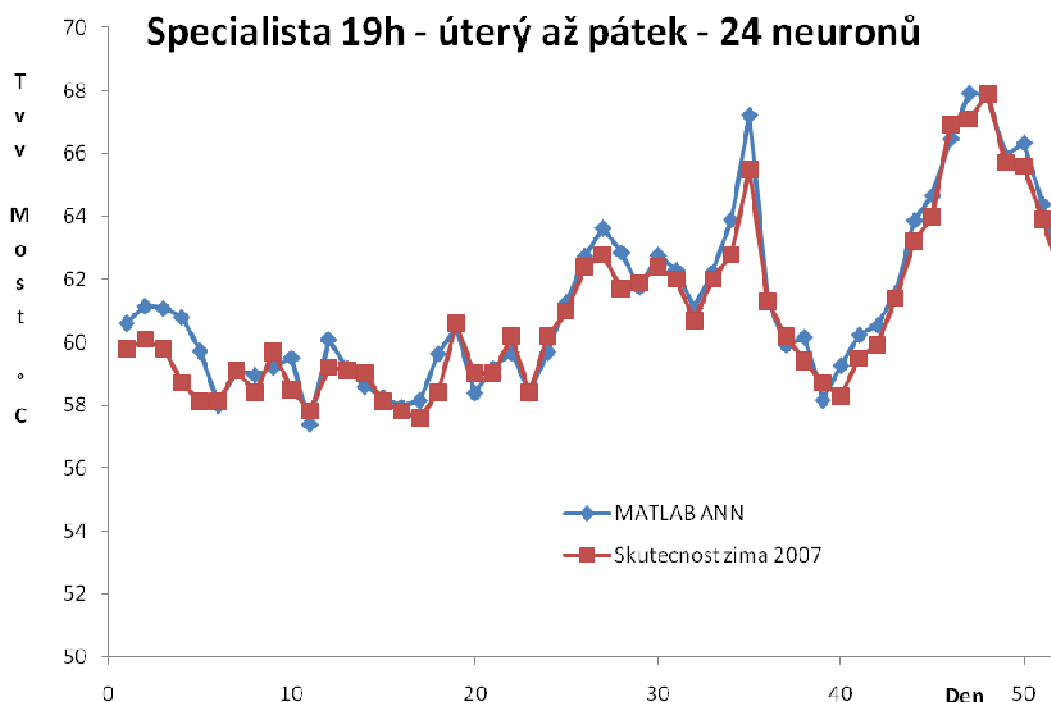


Obr. 26. Specialista na 11h (MAPE=0,0119)

Popis řešení: Pro dny úterý až pátek v 11h (měsíce leden-březen, říjen-prosinec 2005 a 2006). Neuronová síť má 20 neuronů, na vstupu jsou 4 druhy proměnných za 24 hodin zpátky = 96 vstupů. Výstup je 1 - teplota vratné vody Most za nejbližší další hodinu. Síť byla testována na 52 dnech v měsících říjen-prosinec 2007, v 11h, dny úterý až pátek. Ukazatel MAPE (% chybovost předpovědi oproti skutečnosti) dosáhl hodnotu 1,19%.

Vstupy do sítě byly teplota externí za posledních 24h zpátky, podobně teplota topné vody, teplota vratné vody a množství dodané vody v tunách za hodinu.

Na dalších obrázku (obr. 27) je ukázka predikce pro specialistu na 19h večer.



Obr. 27. Specialista na 19h (MAPE = 0,0097)

Popis řešení: Pro dny úterý až pátek v 19h (měsíce leden-březen, říjen-prosinec 2005 a 2006). Neuronová síť má 24 neuronů, na vstupu jsou 4 druhy proměnných za 24 hodin zpátky = 96 vstupů. Výstup je 1 - teplota vratné vody Most za nejbližší další hodinu. Síť byla testována na 52 dnech v měsících říjen-prosinec 2007, v 19h, dny úterý až pátek. Ukazatel MAPE (% chybovost předpovědi oproti skutečnosti) dosáhl hodnotu 0,97%.

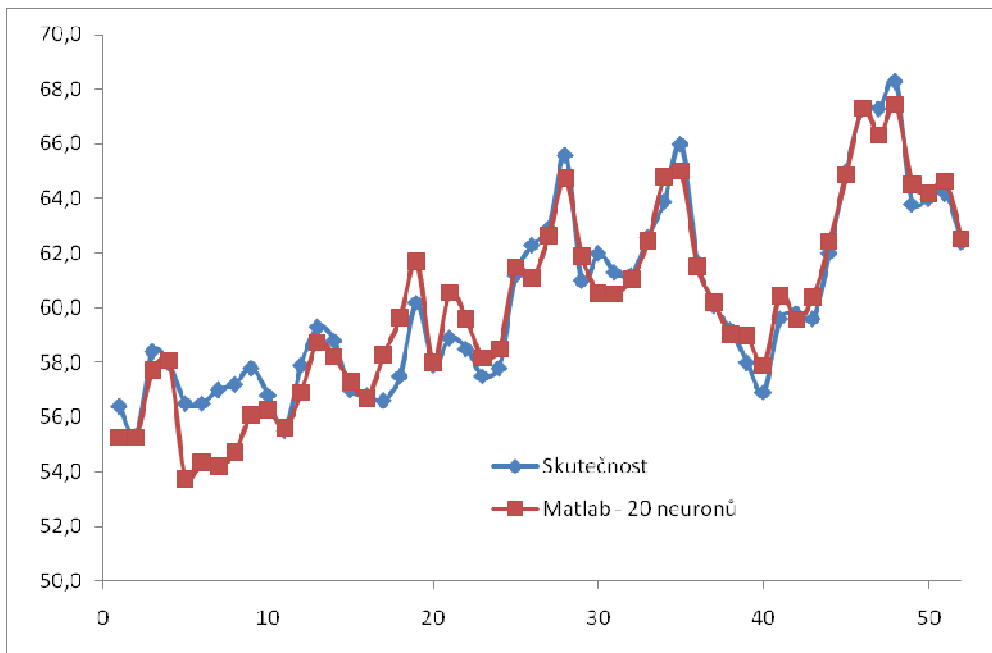
Zde se mírně projevilo zlepšení ve zvýšení počtu neuronů z 20 na 24. Další pokusy byly proto směřovány do odhadu, jak ovlivní predikci zvýšení počtu neuronů.

4.3 Změna počtu neuronů v síti

Mohlo by se zdát, že zvýšení počtu neuronů automaticky znamená zlepšení výkonnosti umělé neuronové sítě. Následující grafy dokazují, že tomu tak není.

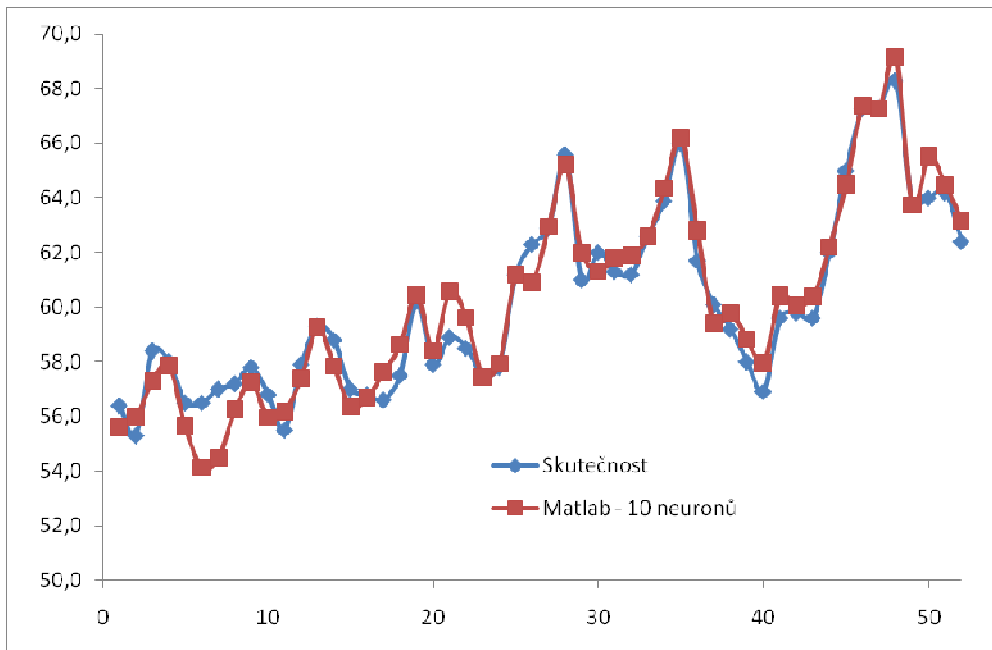
Vstupní podmínky byly stejné, jako v obr. 25. Bylo použito 6 vstupních proměnných, ale za 24h zpátky, tzn. jeden vstupní vektor měl 144 hodnot. Celkově se síť učila na 208 případech z historie a předpovídala 52 hodnot. Výstupní vektor měl pouze 1 proměnnou – teplotu vratné vody o hodinu později.

Na obr. 28 vidíme, že síť, která má 20 neuronů, si vedla celkem dobře a průměrná chyba byla pouze 1,38%.



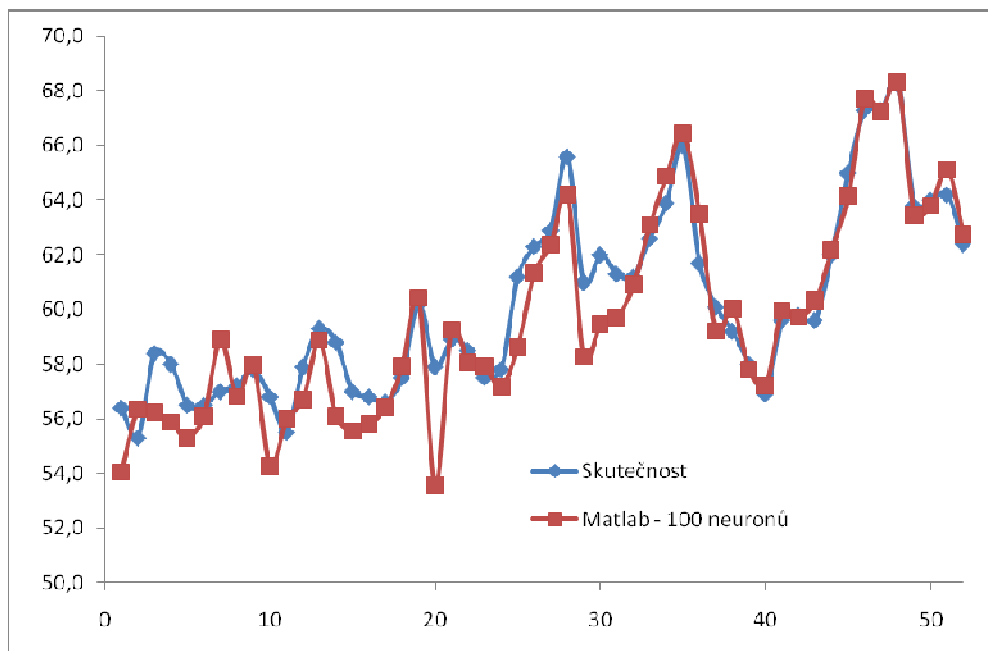
Obr. 28. 20 neuronů (MAPE=0,0138)

Snížení počtu neuronů na 10 (obr. 29) dokonce zlepšilo výkonnost neuronové sítě a chyba klesla na 1,15%.



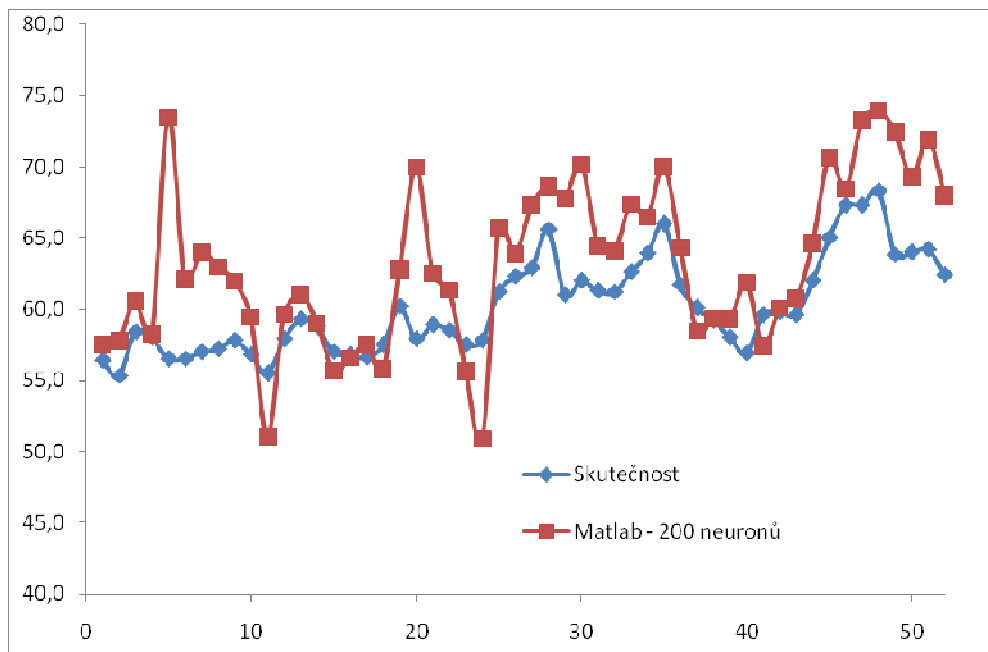
Obr. 29. 10 neuronů (MAPE=0,0115)

Zvýšení počtu neuronů na 100 nemá příliš velký vliv – chyba zůstala velmi podobná, pouze 1,67% (obr. 30). Dá se říci, že v rozsahu 10-100 neuronů se síť chová celkem rozumně a dává použitelné výsledky.



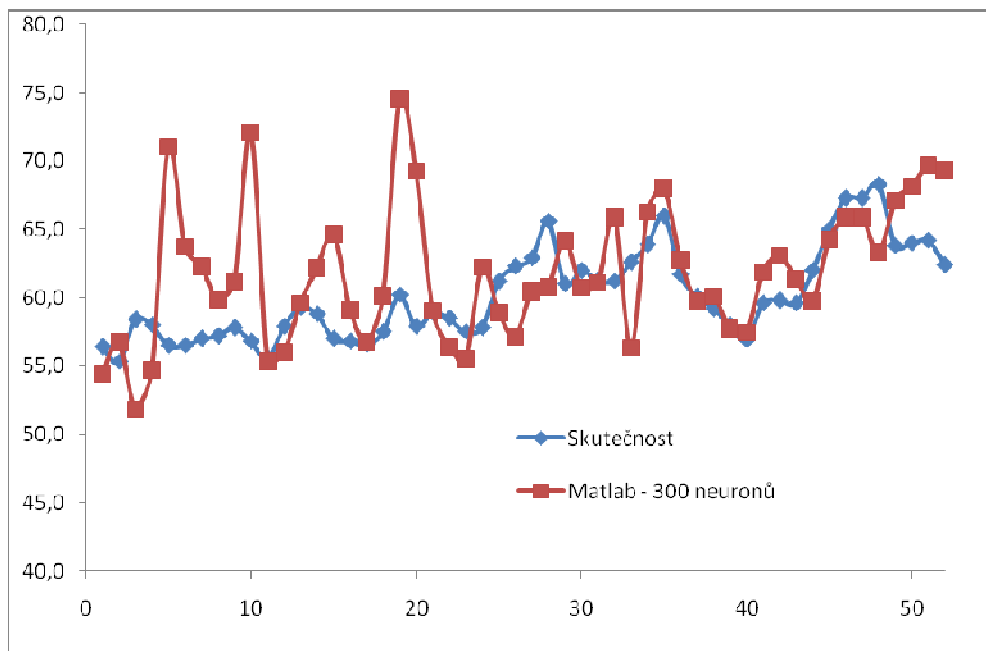
Obr. 30. 100 neuronů (MAPE=0,0167)

Po zvýšení neuronů na 200 síť výrazně zhoršila svůj výkon – chyba 6,32% (obr. 31).



Obr. 31. 200 neuronů (MAPE=0,0632)

Že se nejedná o náhodu, dokazuje zvýšení počtu neuronů na 300. Chyba je 6,14%. (obr. 32). Neuronová síť má chybu predikce závislou na počtu neuronů, ale nedá se předem určit, jak to má být správně, pouze empiricky.



Obr. 32. 300 neuronů (MAPE=0,0614)

Ze všech uvedených ukázek predikce vyplývá, že pro konkrétní potřeby energetického provozu je potřebné znát detaily systému jako celku a pak by bylo možné vytvořit ucelený predikční systém založený na neuronových sítích.

Možností, jak namíchat ten správný koktejl dat a nastavení je mnoho – výběr proměnných, výběr časových dat pro učení sítě, hledání optimálního počtu neuronů, několikrát opakované trénování sítě, zjištění vah jednotlivých neuronů a následné naprogramování automatické aplikace apod.

Rozsah činností potřebných pro nalezení dosažitelného optima překračuje rámec této bakalářské práce. Přesto věřím, že po důkladné statistické analýze výkonnosti jednotlivých sítí se všemi mnohokriteriálními nastaveními je možné žádoucí optimální predikční systém založený na neuronových sítích nalézt.

ZÁVĚR

V bakalářské práci jsem řešil téma data miningu v energetickém průmyslu. Cílem bylo vybrat a demonstrovat vhodnou metodu pro aplikaci data miningu pro konkrétní provoz – elektrárnu a teplárnu v Komořanech. Touto metodou jsou neuronové sítě.

V práci jsem uvedl důvody, proč je dobré využít data mining pro řízení distribuce tepla v městské aglomeraci. Popsal jsem dostupné metody a vybral jednu, které jsem se věnoval podrobně. V praktické části jsem popsal postup predikce pomocí neuronových sítí a dostupného softwaru. Vlastní výsledky predikce tvoří závěr mé práce.

Další postup v této oblasti bych viděl v systémovém přístupu k problematice predikce pomocí neuronových sítí v Matlabu. Je nutné znát provozní potřeby a možnosti elektrárny v Komořanech, až pak může tato práce přinést konkrétní výsledky. Výzkum by mohl být dobrým tématem pro diplomovou práci – podrobná statistická analýza všech možných i nemožných kombinací dat a postupů z oblasti predikce použitím neuronových sítí.

Po nalezení optima by ovšem bylo stěžejní aplikovat dokončený výzkum napojením softwarové aplikace do systému řízení energetického provozu a jeho odzkoušení a vyladění na místě. Tato aplikace by měla být plně automatizována a sloužit tak lepšímu, efektivnějšímu provozu celé společnosti.

ZÁVĚR V ANGLIČTINĚ

In the bachelor thesis I solved the theme of data mining in energetic industry. The aim was to choose and demonstrate appropriate method for application of data mining in concrete plant – power and heating plant in Komořany. This method is neural network.

In the thesis I introduced reasons, why it is a good idea to use data mining for controlling of distribution of heat in a city agglomeration. I described available methods and chose one, which I dedicated in detail. In the practical part I described instructions for prediction using neural networks and available software. Actual results of prediction form conclusion of my thesis.

The next progress in this field could be seen in systematic attitude to questions of prediction using neural networks in Matlab. It is necessary to know operational needs and possibilities of power plant in Komořany and then can this work bring concrete results. The research could be a good theme for master thesis – detailed statistic analysis of all possible and impossible combinations of data and procedures from the field of prediction using neural networks.

After the optimum will be found, it is crucial to apply finished research by connecting a software application to the system of controlling energetic plant and its testing and tuning on the spot. This application should be fully automated and to serve for better, more effective running of the whole company.

SEZNAM POUŽITÉ LITERATURY

- [1] KOUT J. - HEJNÝ T. - KLÉMA J. *Inteligentní řízení a rozhodování v distribučních sítích*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <http://www.certicon.cz/fileadmin/Certicon/downloads/researchArticles/1_4.pdf>.
- [2] *Data mining*. Wikimedia Foundation. [online]. [cit. 2009-04-15]. Dostupný z WWW: <http://cs.wikipedia.org/wiki/Data_mining>.
- [3] *Híbková analýza dat*. Wikimedia Foundation. [online]. [cit. 2009-04-15]. Dostupný z WWW: <http://sk.wikipedia.org/wiki/Data_mining>.
- [4] KLÍMEK, Petr. *Získávání znalostí z podnikových dat (data mining)*. 1. vyd. Zlín: Univerzita Tomáše Bati ve Zlíně, 2005. 35 s. ISBN 8073182416 (brož.)
- [5] BERKA, Petr. *Aplikace systémů dobývání znalostí pro analýzu medicínských dat*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://euromise.vse.cz/kdd/index.php?page=metody#soused>>.
- [6] SERRANO J. I. – TOMEČKOVÁ M. – ZVÁROVÁ J. *Metody strojového učení pro vyhledávání znalostí v lékařských datech o ateroskleróze*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://www.ejbi.org/articles/200608/25/2.html>>.
- [7] KVASNIČKA M., VAŠÍČEK O. *Úvod do analýzy časových řad*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://www.econ.muni.cz/~qasar/vyuka/emm2/skriptaemmii.pdf>>.
- [8] *Sequence analysis*. Wikimedia Foundation. [online]. [cit. 2009-04-15]. Dostupný z WWW: <http://en.wikipedia.org/wiki/Sequence_analysis>.
- [9] *Analysis of kovariance*. Wikimedia Foundation. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://en.wikipedia.org/wiki/ANCOVA>>.
- [10] *Analýza rozptylu*. Wikimedia Foundation. [online]. [cit. 2009-04-15]. Dostupný z WWW: <http://cs.wikipedia.org/wiki/Analýza_rozptylu>.
- [11] HANZELKA, David. *Bayesovská umělá inteligence*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <hilbert.chtf.stuba.sk/KUZV/download/kuzv-hanzelka.pdf>.
- [12] MELOUN, Milan. *Diskriminační analýza*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://meloun.upce.cz/kapitoly/4fmetody.pdf>>.

- [13] *Metodiky mnohorozměrné statistiky*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://tns.factum.cz/mva>>.
- [14] TEDA, Jaroslav. *Genetické algoritmy a jejich aplikace v praxi*. [online]. [cit. 2009-04-20]. Dostupný z WWW: <<http://programujte.com/index.php?akce=clanek&cl=2005072601-geneticke-algoritmy-a-jejich-aplikace-v-praxi>>.
- [15] *Genetický algoritmus*. Wikimedia Foundation. [online]. [cit. 2009-04-15]. Dostupný z WWW: <http://cs.wikipedia.org/wiki/Genetický_algoritmus>.
- [16] MELOUN, Milan. *Faktorová analýza*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://meloun.upce.cz/kapitoly/4dmetody.pdf>>.
- [17] RYDVAL, Slávek. *Základy fuzzy logiky*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://www.rydval.cz/phprs/view.php?cislocianku=2005061701>>.
- [18] VOJÁČEK, Antonín. *Samoučící se neuronová síť - SOM, Kohonenovy mapy*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://automatizace.hw.cz/clanek/2006051401>>.
- [19] *Kontingenční tabulka*. Wikimedia Foundation. [online]. [cit. 2009-04-15]. Dostupný z WWW: <http://cs.wikipedia.org/wiki/Kontingenční_tabulka>.
- [20] *Korelace*. Wikimedia Foundation. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://cs.wikipedia.org/wiki/Korelace>>.
- [21] KOTOUČEK M. - SKOPALOVÁ J. - ADAMOVSÝ P. *Příklady z analytické chemie*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://ach.upol.cz/ucebnice/hodnoceni7.htm>>.
- [22] *Logistická regrese*. Wikimedia Foundation. [online]. [cit. 2009-04-15]. Dostupný z WWW: <http://cs.wikipedia.org/wiki/Logistická_regrese>.
- [23] BERÁNEK, L., HAVRÁNKOVÁ R. *Biostatika*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <http://www.zsf.jcu.cz/struktura/katedry/radio/informace-pro-studenty/ucebni_texty/ochrana-obyvatelstva-se-zamerenim-na-cbrne-aplikovana-radiobiologie-a-toxikologie-krizova-radiobiologie-a-toxikologie/biostatistika.doc/>.

- [24] *Odhad parametrů metodou maximální věrohodnosti*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://cmp.felk.cvut.cz/cmp/courses/recognition/Labs/mlodhad/index.html>>.
- [25] KELBEL J., ŠILHÁN D. *Shluková analýza*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://gerstner.felk.cvut.cz/biolab/X33BMI/slides/KMeans.pdf>>.
- [26] ŽÁK, Libor. *Shluková analýza I*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <http://www.volny.cz/elzet/Libor/Aut_cl_1.pdf>.
- [27] *Support vector machines (SVM)*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <http://is.muni.cz/el/1433/podzim2006/PA034/09_SVM.pdf?fakulta=1433;obdobi=3523;kod=PA034>.
- [28] OBITKO, Marek. *Prediction using neural networks*. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://obitko.com/tutorials/neural-network-prediction/>>.
- [29] ZELINKA, Ivan. *Umělá inteligence*. 2. vyd. Zlín: Univerzita Tomáše Bati ve Zlíně. 2005. 127 s. ISBN 80-7318-277-7.
- [30] VEELANTURF, L.P.J. *Analysis and Applications of Artificial Neural Networks*. 1. vyd. New Jersey: Prentice Hall Inc., 1995. 259 s. ISBN 0-13-489832-X.
- [31] HAGAN M. T. - DEMUTH H. - B. BEALE M. H. *Neural Network Design*. 1. vyd. Boston: PWS Publishing, 1996. 734 s. ISBN 0971732108.
- [32] DOSTÁL, Petr. *Advanced economic analyses*. 1. vyd. Brno: Akademické nakladatelství CERM, 2008. 80 s. ISBN 978-80-214-3564-3.
- [33] DOSTÁL, Petr. *Moderní metody ekonomických analýz*. 1. vyd. Zlín: Univerzita Tomáše Bati ve Zlíně, 2002. 110 s. ISBN 80-7318-075-8.
- [34] McNELIS, Paul D. *Neural networks in finance*. 1. vyd. San Diego: Elsevier Inc., 2005. 261 s. ISBN 0-12-485967-4.
- [35] *Time Series Analysis*. StatSoft, Inc. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://www.statsoft.com/textbook/stathome.html?sttimser.html&1>>.
- [36] ŠNOREK, Milan. *Neuronové sítě a neuropočítače*. 1. vyd. Praha: Vydavatelství ČVUT - výroba, 2004. 406 s. ISBN 80-01-02549-7.

[37] *Mean absolute percentage error*. Wikimedia Foundation. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://en.wikipedia.org/wiki/MAPE>>.

[38] United Energy, a. s. [online]. [cit. 2009-04-15]. Dostupný z WWW: <<http://www.ue.cz>>.

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

ANN Artificial neural networks.

MAPE Mean absolute percentage error.

MWe Megawatt electrical.

MWt Megawatt thermal.

nftool Neural fitting tool.

nntool Neural network tool.

SEZNAM OBRÁZKŮ

<i>Obr. 1. Klasifikace dle nejbližšího souseda [5]</i>	13
<i>Obr. 2. Rozdělení pravděpodobnosti v uzlech Bayesovské sítě [11]</i>	16
<i>Obr. 3. Objekty ve dvojrozměrném prostoru: jedná se o 2 nebo 3 shluky? [26]</i>	23
<i>Obr. 4. Princip vzniku možnosti lineárního oddělení dvou tříd s nelineárními hranicemi pomocí přidané dimenze [27]</i>	24
<i>Obr. 5. Zjednodušený biologický neuron [30]</i>	27
<i>Obr. 6. Umělý model neuronu [30]</i>	27
<i>Obr. 7. Funkce log sigmoid [34]</i>	28
<i>Obr. 8. Struktura vícevrstvé sítě se vstupní, skrytou a výstupní vrstvou [30]</i>	29
<i>Obr. 9. Vícenásobná predikce kurzů měn [29]</i>	31
<i>Obr. 10. Start – Toolboxes – Neural Network</i>	39
<i>Obr. 11. Nástroj nftool po spuštění</i>	40
<i>Obr. 12. Výběr dat pro import</i>	40
<i>Obr. 13. Rozdělení dat na trénovací, validační a testovací</i>	41
<i>Obr. 14. Nastavení počtu neuronů v síti</i>	41
<i>Obr. 15. Trénink sítě a výsledky tréninku</i>	42
<i>Obr. 16. Opakování tréninku</i>	42
<i>Obr. 17. Vizualizace regrese</i>	43
<i>Obr. 18. Uložení sítě a výsledků tréninku</i>	43
<i>Obr. 19. Prostředí pro práci se sítěma</i>	44
<i>Obr. 20. Vytvoření nového vstupu do sítě</i>	44
<i>Obr. 21. Predikce v síti pomocí simulace</i>	45
<i>Obr. 22. Výslední vektor predikce</i>	45
<i>Obr. 23. Predikce teploty topné vody v 1h (MAPE=0,0281)</i>	46
<i>Obr. 24. Predikce teploty topné vody v 11h (MAPE=0,0259)</i>	47
<i>Obr. 25. Predikce teploty topné vody v 1h – 6 parametrů (MAPE=0,0128)</i>	47
<i>Obr. 26. Specialista na 11h (MAPE=0,0119)</i>	48
<i>Obr. 27. Specialista na 19h (MAPE = 0,0097)</i>	49
<i>Obr. 28. 20 neuronů (MAPE=0,0138)</i>	50
<i>Obr. 29. 10 neuronů (MAPE=0,0115)</i>	50
<i>Obr. 30. 100 neuronů (MAPE=0,0167)</i>	51

<i>Obr. 31. 200 neuronů (MAPE=0,0632)</i>	51
<i>Obr. 32. 300 neuronů (MAPE=0,0614)</i>	52

SEZNAM TABULEK

<i>Tab. 1. Rozhodovací pravidla [5]</i>	22
<i>Tab. 2. Rozdíly mezi PC a neuronovou sítí [29]</i>	26
<i>Tab. 3. Popis použitých dat – rok 2005</i>	37
<i>Tab. 4. Popis použitých dat – rok 2006</i>	37
<i>Tab. 5. Popis použitých dat – 1. část roku 2007</i>	37
<i>Tab. 6. Popis použitých dat – 2. část roku 2007</i>	38

SEZNAM PŘÍLOH

P I CD-ROM

Obsahuje pracovní soubory data miningu

PŘÍLOHA P I: CD-ROM