

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

Evidenčné číslo: FIIT-5212-73688

Matúš Cuper

Optimalizácia konfiguračných parametrov predikčných metód

Bakalárska práca

Vedúci práce: Ing. Marek Lóderer

máj 2017

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

Evidenčné číslo: FIIT-5212-73688

Matúš Cuper

Optimalizácia konfiguračných parametrov predikčných metód

Bakalárska práca

Študijný program: Informatika

Študijný odbor: 9.2.1 Informatika

Miesto vypracovania: Ústav informatiky a softvérového inžinierstva, FIIT STU Bratislave

Vedúci práce: Ing. Marek Lóderer

máj 2017

Anotácia

Slovenská technická univerzita v Bratislave

FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLOGIÍ

Študijný program: Informatika

Autor: Matúš Cuper

Bakalárska práca: Optimalizácia konfiguračných parametrov predikčných metód

Vedúci práce: Ing. Marek Lóderer

máj 2017

V práci sme sa zamerali na problémy vznikajúce pri predikcii časových radov. V súčasnosti existuje veľké množstvo metód, ktoré nám zabezpečujú predpoveď sledovanej veličiny s prijateľne malou odchýlkou na krátke obdobie v blízkej budúcnosti. Cieľom bakalárskej práce bolo vytvoriť systém, ktorý používateľovi poskytne jednoduché rozhranie pre porovnanie jednotlivých predikčných algoritmov nad množinou dát, ktorú si sám zvolí. Hľadanie ich optimálneho nastavenia sa vykonáva pomocou optimalizačných algoritmov založených na správaní sa živočíchov v prírode.

V práci sme analyzovali a opísali množinu predikčných a optimalizačných algoritmov. Navrhli sme systém na hľadanie optimálnych parametrov predikčných metód, čím sme výrazne ovplyvnili ich presnosť. Systém bol implementovaný v programovacom jazyku R a na vytvorenie používateľského rozhrania bola použitá knižnica Shiny. Optimalizácie sme vykonávali nad dátovými množinami v doméne energetiky. Výsledný systém umožňuje používateľovi využívať silu predikčných algoritmov a nájsť ich optimálne parametre pre zabezpečenie čo najpresnejšej predikcie.

Annotation

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: Computer Science
Author: Matúš Cuper
Bachelor thesis: Optimizing configuration parameters of prediction methods
Supervisor: Ing. Marek Lóderer
May 2017

In the thesis we focused on problems, which appear in time series prediction. In present there are many methods, which predict observed value with acceptable small deviation for short time period in near future. The aim of bachelor thesis was creating system, which provides simple user interface to compare chosen prediction algorithms on dataset, which is chosen by user. Looking for their optimal setup is made by optimization algorithm based on nature-inspired behavior.

In the thesis we analyzed and described set of prediction and optimization algorithms. We designed system for searching optimal parameters of prediction methods, which influence their accuracy significantly. System was implemented in programming language R and for creating user interface was used library Shiny. Optimization was provided on dataset in energetics domain. The final system provides to user to use force of prediction algorithms and find out their optimal parameters for the most accurate prediction.

ČESTNÉ PREHLÁSENIE

Čestne prehlasujem, že bakalársku prácu som vypracoval samostatne pod vedením vedúceho bakalárskej práce a s použitím odbornej literatúry, ktorá je uvedená v zozname použitej literatúry.

.....
Matúš Cuper

POĎAKOVANIE

Ďakujem vedúcemu bakalárskej práce Ing. Marekovi Lódererovi za odborné vedenie, cenné rady a pripomienky pri spracovaní bakalárskej práce.

Obsah

1	Úvod	1
2	Analýza problému	2
2.1	Časové rady	2
2.1.1	Analýza časových radov	2
2.1.2	Zložky časových radov	3
2.2	Analýza predikčných algoritmov	5
2.2.1	Lineárna regresia	5
2.2.2	Stochastické modely	7
2.2.3	Regresia založená na podporných vektoroch	8
2.2.4	Rozhodovacie stromy	9
2.2.5	Neurónové siete	10
2.3	Analýza optimalizačných algoritmov	12
2.3.1	Genetický algoritmus	12
2.3.2	Optimalizácia svorkou divých vlkov	14
2.3.3	Umelá kolónia včiel	17
2.4	Meranie presnosti predpovede	18
2.4.1	Stredná chyba predpovede	19
2.4.2	Stredná absolútna chyba	19
2.4.3	Stredná percentuálna chyba	19
2.4.4	Stredná absolútna percentuálna chyba	19
2.4.5	Stredná štvorcová chyba	20
2.5	Zhodnotenie analýzy	20
3	Špecifikácia požiadaviek	21
4	Návrh riešenia	22
4.1	Vlastnosti vstupných dát	22
4.2	Použité predikčné metódy	23
4.2.1	Regresia založená na podporných vektoroch	23
4.2.2	Autoregresívny integrovaný model kľzavého priemeru	23
4.2.3	Náhodné lesy	24
4.3	Použité optimalizačné metódy	24
4.3.1	Optimalizácia rojom častíc	25
4.3.2	Umelá kolónia včiel	25
5	Implementácia	26
6	Zhodnotenie	27
7	Záver	28
8	Technická dokumentácia	29
	Literatúra	30

Zoznam obrázkov

1	Príklad trendovej zložky časového radu.	3
2	Príklad sezónnej zložky časového radu.	4
3	Príklad reziduálnej zložky časového radu.	4
4	Príklad multiplikatívneho modelu	5
5	Príklad aditívneho modelu.	6
6	Príklad neurónu v neurónovej sieti [26].	10
7	Príklad viacvrstvovej spätne propagovanej neurónovej siete [13].	11
8	Hierarchia vlkov vo svorke [22].	15
9	Príklad možného umiestnenia vlka v dvojrozmernom priestore na základe polohy koristi [22].	16
10	Ukážka transformácií dátumu a času pri predikcii náhodnými lesmi.	25

Zoznam tabuliek

1	Tabuľka rozdelenia vybraných prírodne inšpirovaných algoritmov [8]. . . .	12
2	Proces vytvorenia novej generácie z rodičovských chromozómov.	13
3	Rozbor formátu dátumu a času.	22

1 Úvod

V súčasnosti sme obklopení množstvom zariadení, ktoré merajú a zhromažďujú informácie z iných zariadení. Príkladom môžu byť rôzne mobilné aplikácie, meteorologické stanice alebo inteligentné merače merajúce spotrebu elektrickej energie, vody či plynu. Namerané dáta sa môžu meniť v závislosti od napr. hodiny, dňa alebo počasia. Samozrejme existujú aj merania, ktoré sú ovplyvnené ľudským správaním, ktoré sa môže líšiť v závislosti od vyššie uvedených faktorov, ale aj faktorov ako sú kultúrne tradície, zvyky či náboženstvo. Na základe nameraných dát vieme vytvoriť predpoveď, ktorá opäť môže ovplyvniť ostatné faktory a celé predpovedanie sa tak stáva opäť komplexnejším. Ak sú dáta merané v pravidelných intervaloch a konzistentne, nazývame ich časový rad.

V práci sme sa zamerali na predpovedanie veličín na základe ich historických meraní. Ostatné faktory pri tom neboli brané do úvahy. Tým sa stáva predpovedanie jednoduchšie a menej presné, čím vzniká priestor pre hlavný zámer práce, optimalizovanie predikčných metód na základe ich vstupných parametrov. Získame tým presnejšiu predpoveď ako keby sme nastavenie predikčných algoritmov nechali na náhode alebo vlastnom úsudku.

Optimalizačné algoritmy, tak ako ich názov napovedá, slúžia na nájdenie optimálneho riešenia. Nájdenie najlepšieho riešenia požaduje preskúmanie všetkých možností. Stáva sa tak pomalým a výpočtovo náročným. Optimalizačné algoritmy rýchlo nájdu riešenie, ktoré je pre potreby našej práce postačujúce. Optimalizačné algoritmy môžeme rozdeliť do viacerých skupín, v práci sa však zameriavame prednostne na prírodne inšpirované algoritmy, ktoré sú jednou z najefektívnejších podskupín.

Výsledný systém poskytuje používateľovi webové grafické rozhranie, pomocou ktorého môže predpovedať hodnoty vložených časových radov. Má možnosť zvoliť si medzi viacerými predikčnými a optimalizačnými algoritmami. Rozhranie poskytuje používateľovi základné informácie o algoritmoch, ako aj vysvetlenie efektov jednotlivých vstupných parametrov metód. Je na zvážení používateľa, aké parametre zvolí pre optimalizačné algoritmy. Vstupné parametre predikčných metód budú zvolené optimálne na základe výstupných hodnôt optimalizačných algoritmov. Používateľ má k dispozícii vyhodnotenie predpovede, ktoré porovnáva predpovedanú a skutočnú hodnotu rôznymi metrikami.

Celá práca je rozdelená do niekoľkých kapitol. V kapitole 2 sme sa zamerali na analýzu problému. Definovali sme kľúčové pojmy, použité metódy, opísali sme časové rady, ich vlastnosti, rozdelili sme predikčné a optimalizačné algoritmy do skupín. Kapitola 3 popisuje funkcionality, ktorú systém poskytuje používateľom. V kapitole 4 sa zameriavame na návrh systému a v kapitole 5 už samotnou implementáciou systému v jazyku R. Výsledky práce sú zhodnotené v kapitole 6.

2 Analýza problému

Predpovedanie spotreby elektrickej energie je kľúčovou činnosťou pre plánovanie a prevádzkovanie rôznych elektronických zariadení. Hľadaný vzor pre časové rady spotreby elektrickej energie je často komplexný a je zložitý ho nájsť aj kvôli faktorom ako sú napr. zmeny cien elektrickej energie na trhu. Preto sa stáva implementácia vhodného modelu zaručujúceho presnú predpoveď náročnou [18].

Práve preto vznikajú rôzne metódy na predpovedanie časových radov, ktoré sú bližšie popísané v nasledujúcich podkapitolách. Väčšina z nich poskytuje niekoľko vstupných parametrov ako rozhranie pre vnútorné nastavenie algoritmov. Vďaka tomu máme možnosť ovplyvňovať mieru natrénovania modelu, veľkosť chyby predikcie alebo dĺžku obdobia, na ktorom budeme model trénovať. Nastavenie týchto parametrov sa môže pri rôznych dátasetoch líšiť, a preto neexistuje univerzálne riešenie. Hľadať riešenie pomocou prírodne inšpirovaných algoritmov je efektívne a nájdené riešenie je optimálne. Ďalej sú v tejto kapitole opísané spôsoby merania chýb, ktoré slúžia na vyhodnotenie efektivity a správnosti nájdeného riešenia.

2.1 Časové rady

Časový rad je množina dátových bodov nameraná v čase postupne za sebou. Matematicky je definovaný ako množina vektorov $x(t)$, kde t reprezentuje uplynulý čas. Premenná $x(t)$ je považovaná za náhodnú premennú. Merania v časových radoch sú usporiadané v chronologickom poradí [1].

Časové rady delíme na spojité a diskrétny. Pozorovania pri spojitých časových radoch sú merané v každej jednotke času, zatiaľ čo diskrétny obsahujú iba pozorovania v diskrétnych časových bodoch. Hodnoty toku rieky, teploty či koncentrácie látok pri chemickom procese môžu byť zaznamenané ako spojitý časový rad. Naopak, populácia mesta, produkcia spoločnosti alebo kurzy mien reprezentujú diskrétny časový rad. Vtedy sú pozorovania oddelené rovnakými časovými intervalmi, napr. rokom, mesiacom či dňom [1]. V našom prípade sú namerané dáta dostupné každú celú štvrt' hodinu.

2.1.1 Analýza časových radov

V praxi je vhodný model napasovaný do daného časového radu a zodpovedajúce parametre sú predpovedané na základe známych dát. Pri predpovedaní časových radov sú dáta z predchádzajúcich meraní zhromažďované a analyzované za účelom navrhnutia vhodného matematického modelu, ktorý zaznamenáva proces generovania dát pre časové rady. Pomocou tohto modelu sú predpovedané hodnoty budúcich meraní. Takýto prístup je užitočný, keď nemáme veľa poznatkov o vzore v meraniach idúcich za sebou alebo máme model, ktorý poskytuje nedostatočne uspokojivé výsledky [1].

Cieľom predikcií časových radov je predpovedať hodnotu premennej v budúcnosti na základe doteraz nameraných dátových vzoriek. Matematicky zapísané ako

$$\hat{x}(t + \Delta_t) = f(x(t - a), x(t - b), x(t - c), \dots) \quad (1)$$

Hodnota \hat{x} vo vzorci 1 je predpovedaná hodnota jednorozmerného diskrétného časového radu x . Úlohou je nájsť takú funkciu $f(x)$, pre ktorú bude \hat{x} predstavovať predpovedanú hodnotu časového radu. Táto funkcia by mala predpovedať hodnoty v budúcnosti konzistentne a objektívne [21].

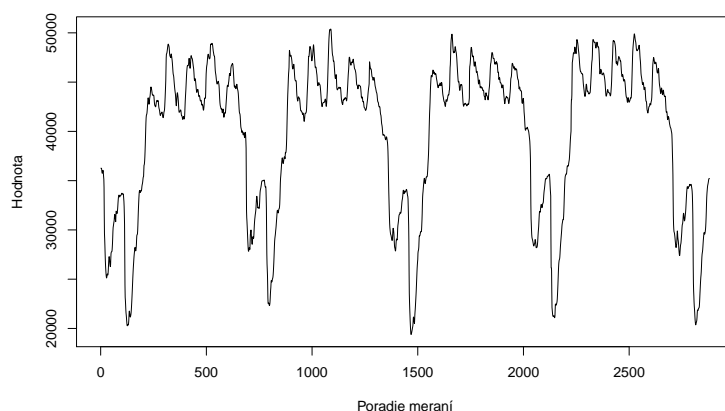
Časové rady sú najčastejšie vizualizované ako graf, kde pozorovania sú na osy y a plynúci čas na osy x . Pre lepšie vysvetlenie časových radov, budú nasledujúce odstavce obsahovať aj obrázky zobrazujúce vygenerovanú dátovú množinu.

2.1.2 Zložky časových radov

Pri predpovedaní časových radov ako napr. meraní odberu elektrickej energie vznikajú 2 typy trendov. Prvým typom je trvalá alebo dočasná zmena spôsobená ekonomickými alebo ekologickými faktormi. Druhým typom je sezónna zmena, spôsobená zmenami ročných období a množstvom denného svetla. Môžeme ju pozorovať na úrovni dní, týždňov alebo rokov. Veličina, ktorú sa snažíme predpovedať postupne mení svoje správanie a model sa tak stáva nepresným. Kvôli tomu je nutné v každom modeli rozdeliť tieto typy tendencií, aby sme vedeli model zmenám prispôbiť [9].

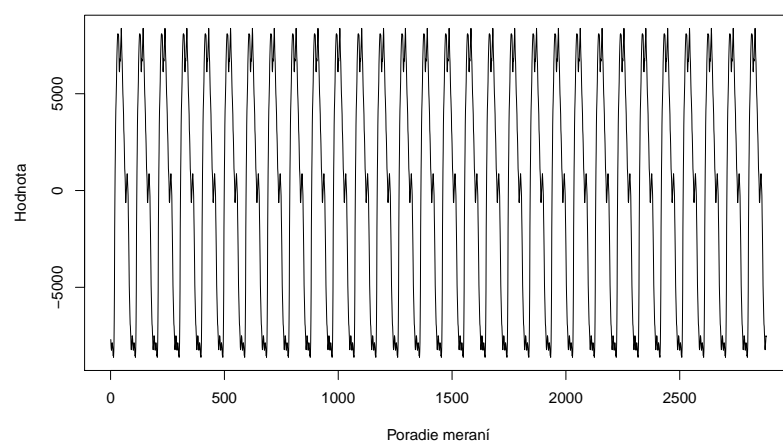
Vo všeobecnosti sú časové rady zložené zo 4 hlavných zložiek, ktoré môžeme odlíšiť od pozorovaných dát. Jedná sa o trendovú, cyklickú, sezónnu a reziduálnu zložku [1].

Trendová zložka predstavuje správanie časového radu v dlhodobom časovom horizonte. Z tohto pohľadu má časový rad tendenciu klesať, rásť alebo stagnovať. Príkladom môže byť nárast populácie či klesajúca úmrtnosť [1].

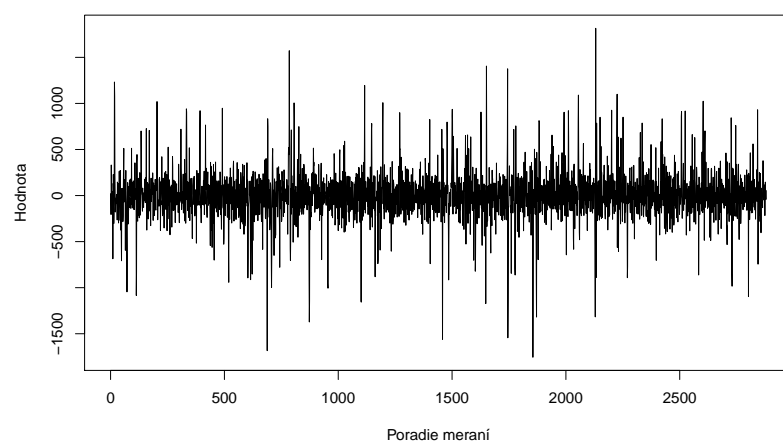


Obr. 1: Príklad trendovej zložky časového radu.

Cyklická zložka je spôsobená zmenami, ktoré sa cyklicky opakujú. Dĺžka periódy je 2 a viac rokov, čo zodpovedá strednodobému časovému horizontu. Táto zložka je zastúpená najmä pri ekonomických časových radoch napríklad podnikateľský cyklus pozostávajúci zo 4 fáz, ktoré sa stále opakujú [1].



Obr. 2: Príklad sezónnej zložky časového radu.



Obr. 3: Príklad reziduálnej zložky časového radu.

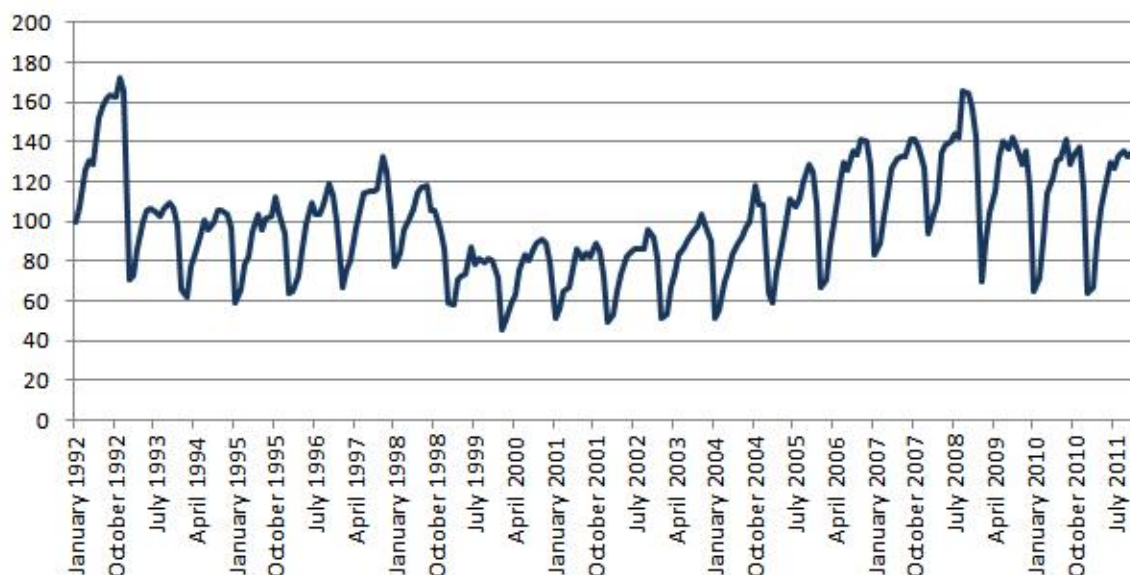
Sezónna zložka predstavuje kolísanie časových radov počas ročných období. Dôležitými faktormi pri tom sú napr. klimatické podmienky, tradície alebo počasie. Napríklad predaj zmrzliny sa v lete zvyšuje, ale počet predaných lyžiarskych súprav klesá [1].

Reziduálna zložka predstavuje veličinu, ktorá nemá žiadny opakovateľný vzor a ani dlhodobý trend. V časových radoch má nepredvídateľný vplyv na pozorovanú veličinu. V štatistike zatiaľ nie je definovaná metóda jej merania. Označuje sa aj ako náhodná zložka alebo biely šum. Je spôsobená nepredvídateľnými a nepravidelnými udalosťami [1].

Vo všeobecnosti sa pre tieto 4 zložky používajú 2 rôzne modely. Je to multiplikatívny model a aditívny model.

$$\begin{aligned} Y(t) &= T(t) \times S(t) \times C(t) \times I(t) \\ Y(t) &= T(t) + S(t) + C(t) + I(t) \end{aligned} \quad (2)$$

Vo vzorci 2 predstavuje $Y(t)$ meranie v čase t . Premenné $T(t)$, $S(t)$, $C(t)$ a $I(t)$ sú zložkami trendu, sezónnosti, cyklu a náhodnosti. Multiplikatívny model, zobrazený na obrázku 4 je založený na predpoklade, že časové rady môžu byť na sebe závislé a môžu byť ovplyvňované medzi sebou, zatiaľ čo aditívny model, zobrazený na obrázku 5 predpokladá nezávislosť zložiek [1].



Obr. 4: Príklad multiplikatívneho modelu
Index stavebnej produkcie Slovenska

2.2 Analýza predikčných algoritmov

Na základe množstva výskumov, ktoré analyzovali predikčné algoritmy, sme si zvolili reprezentatívnu vzorku algoritmov. Našou snahou bude rôzne predikčné metódy optimalizovať pomocou optimalizačných algoritmov. Pred tým je potrebné analyzovať a pochopiť použité predikčné metódy. Taktiež je potrebné identifikovať ich vstupné parametre, keďže ich neskôr budeme optimalizovať.

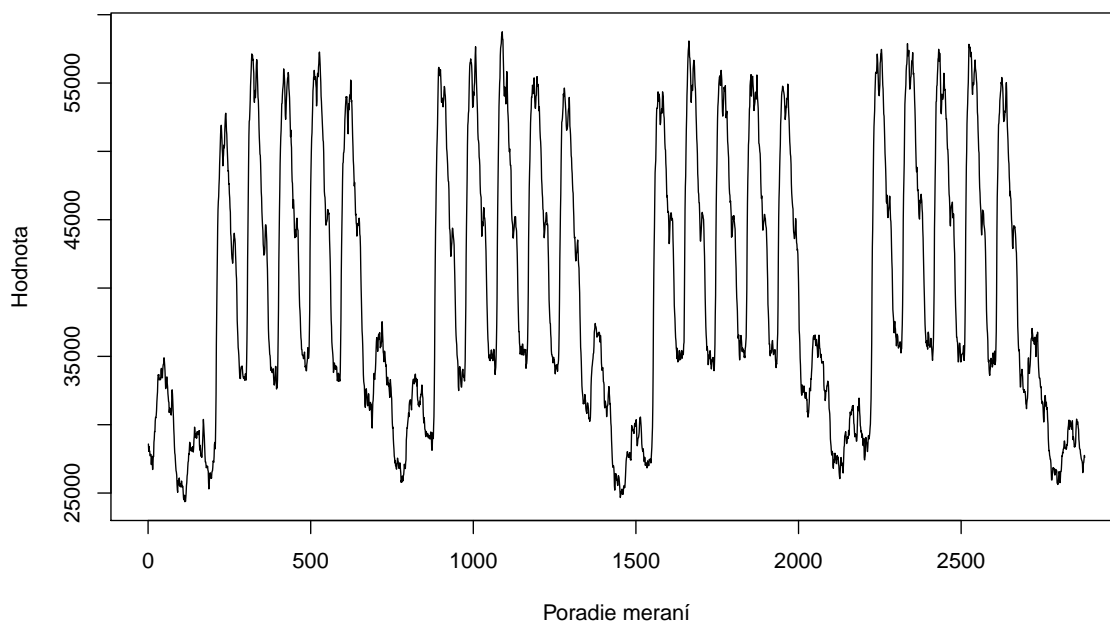
2.2.1 Lineárna regresia

Najpoužívanejšia štatistická metóda, ktorá modeluje vzťah závislej premennej a vysvetľujúcej premennej. Závislú premennú predstavuje veličina, ktorú sa snažíme predpovedať, čo je v našom prípade spotreba elektrickej energie. Vysvetľujúca premenná v sebe zahŕňa rôzne faktory, ktoré ovplyvňujú závislú premennú. Môžeme si pod tým predstaviť deň v týždni, počasie, tradície alebo rôzne udalosti, ktoré majú vplyv na predpoveď. [14, 18].

Predpokladajme typický regresný problém. Dáta pozostávajúce z množiny n meraní majú formát $\{(x_1, f(x_1)), \dots, (x_n, f(x_n))\}$. Úlohou regresie je odvodiť funkciu \hat{f} z dát, kde

$$\hat{f} : X \rightarrow \mathbb{R}, \text{ kde, } \hat{f}(x) = f(x), \forall x \in X, \quad (3)$$

Funkcia f vo vzorci 3 reprezentuje reálnu neznámu funkciu. Algoritmus použitý na odvodenie funkcie \hat{f} sa nazýva indukčný algoritmus. Funkcia \hat{f} sa nazýva model alebo prediktor.



Obr. 5: Príklad aditívneho modelu.

Obvykle je úlohou regresie minimalizovať odchýlku funkcie pre štvorcovú chybu, konkrétne strednú štvorcovú chybu MSE [19].

Keďže časový rad pozostáva z viacerých zložiek, môžeme ho zapísať ako funkciu $L(t)$ definovanú ako

$$L(t) = L_n(t) + \sum a_i x_i(t) + e(t) \quad (4)$$

Vo vzorci 4 funkcia $L_n(t)$ predstavuje odber elektrickej energie v čase t . Hodnota a_i je odhadovaný pomaly meniaci sa koeficient. Faktory $x_i(t)$ nezávisle vplývajú na spotrebu elektrickej energie. Môže sa jednať napr. o počasie alebo zvyky ľudí. Komponent $e(t)$ je biely šum, ktorý má nulovú strednú hodnotu a pevnú varianciu. Číslo n je počet meraní, obvykle 24 alebo 168, v našom prípade 96 meraní počas jedného dňa [14].

Lineárna regresná analýza je štatistická metóda používaná na modelovanie vzťahov, ktoré môžu existovať medzi veličinami. Nachádza súvislosti medzi závislou premennou a potenciálnymi vysvetľujúcimi premennými. Používame pri tom vysvetľujúce premenné, ktoré môžu byť namerané súčasne so závislými premennými alebo aj premenné z úplne iných zdrojov. Regresná analýza môže byť tiež použitá na zlúčenie trendu a sezónnych zložiek do modelu. Keď je raz model vytvorený, môže byť použitý na zásah do spomínaných vzťahov alebo, v prípade dostupnosti vysvetľujúcich premenných, na vytvorenie predikcie [17].

Viacnásobná lineárna regresia sa snaží modelovať vzťah medzi dvoma alebo viacerými vysvetľujúcimi premennými a závislou premennou vhodnou lineárnou rovnicou pre pozorované dáta. Výsledný model je vyjadrený ako funkcia viacerých vysvetľujúcich premenných [9].

Túto funkciu môžeme zapísať ako

$$Y(t) = V_t a_t + e_t \quad (5)$$

Vo vzorci 5 t označuje čas, kedy bolo meranie uskutočnené. $Y(t)$ predstavuje celkový nameraný odber elektrickej energie. Vektor V_t reprezentuje hodnoty vysvetľujúcich premenných v čase merania. Vysvetľujúce premenné môžu predstavovať meteorologické vplyvy, ekonomický nárast, ceny elektrickej energie či kurzy mien. Chybu modelu v čase t zapíšeme ako e_t [14, 18].

2.2.2 Stochastické modely

Tieto metódy časových radov sú založené na predpoklade, že dáta majú vnútornú štruktúru, ako napr. autokoreláciu, trend či sezónnu variáciu. Najprv sa precízne zostaví vzor zodpovedajúci dostupným dátam a potom sa na jeho základe predpovie budúca hodnota veličiny [14].

Autoregresný model predpovedá budúcu hodnotu premennej ako súčet lineárnej kombinácie p predchádzajúcich meraní, náhodnej chyby a konštanty. V literatúre sa označuje ako AR (autoregressive model). Matematicky môžeme autoregresný model zapísať ako

$$y_t = c + \sum_{i=1}^p \varphi_i y_{t-i} + \varepsilon_t \quad (6)$$

Vo vzorci 6 hodnota y_t predstavuje predpovedanú hodnotu v čase t . Náhodnú chybu v čase t zapíšeme ako ε_t . Hodnoty φ_i sú parametre modelu a c je konštanta. Konštantou p označujeme rad modelu [1].

Model kľavého priemeru na rozdiel od autoregresného modelu používa ako vysvetľujúce premenné chyby predchádzajúcich meraní a nie priamo hodnoty. V literatúre sa označuje ako MA (moving average). Matematicky môžeme tento vzťah zapísať ako

$$y_t = \mu + \sum_{j=1}^q \Theta_j \varepsilon_{t-j} + \varepsilon_t \quad (7)$$

Vo vzorci 7 hodnota y_t predstavuje strednú hodnotu postupnosti meraní v čase t . Hodnoty Θ_j sú parametre modelu a konštantou q označujeme rad modelu. Vychádzame z predpokladu, že náhodná zložka ε_t je biely šum, čo je rovnomerne distribuovaná náhodná premenná, ktorá má nulovú strednú hodnotu a konštantnú variáciu σ^2 [1].

Autoregresívny model kľavého priemeru reprezentuje súčasnú hodnotu časového rádu lineárne na základe jeho hodnôt a hodnôt bieleho šumu v predchádzajúcich periódach. V literatúre sa označuje ako ARMA (autoregressive moving average) [14].

Ide o kombináciu autoregresie (AR) a kľavého priemeru (MA), vhodnú pre modelovanie jednorozmerných časových radov. Matematicky môžeme reprezentovať tento model ako súčet predchádzajúcich modelov

$$y_t = c + \varepsilon_t + \sum_{i=1}^p \varphi_i y_{t-i} + \sum_{j=1}^q \Theta_j \varepsilon_{t-j} \quad (8)$$

Rad modelu určuje p a q [1].

Autoregresívny integrovaný model kľzavého priemeru je generalizáciou modelu ARMA. V literatúre sa označuje ako ARIMA (autoregressive integrated moving average). Modely typu ARMA môžu byť použité iba na statické časové rady. Mnoho časových radov v praxi vykazuje nestatické správanie a napr. tie, ktoré obsahujú komponenty trendu a sezónnosti. Kvôli tomu bol navrhnutý model ARIMA, ktorý zahŕňa v sebe aj prípady nestatických časových radov. Z nestatických časových radov sa vytvárajú statické pomocou konečného počtu derivovaní dátových bodov. Vzniká tak matematický model, ktorý môžeme zapísať ako

$$\left(1 - \sum_{i=1}^p \varphi_i L^i\right)(1 - L)^d y_t = \left(1 + \sum_{j=1}^q \Theta_j L^j\right) + \varepsilon_t \quad (9)$$

Vzorec 9 môžeme zapísať aj jednoduchšie a to

$$\varphi(L)(1 - L)^d y_t = \Theta(L)\varepsilon_t \quad (10)$$

Vo vzorci 9 predstavujú premenné p , d a q rad autoregresného modelu, modelu kľzavého priemeru a integrovaného modelu. Hodnota d zodpovedá stupňu derivovania, zvyčajne je rovná 1. V prípade, že $d = 0$ dostaneme klasický ARMA model. Rovnakým spôsobom vieme dostať modely AR a MA [1].

2.2.3 Regresia založená na podporných vektoroch

Regresia založená na podporných vektoroch a metóda podporných vektorov je založená na štatistickej teórii učenia, nazývanej aj VC teória, podľa svojich autorov, Vapnik a Chervonenkisa. Táto predikčná metóda poskytuje malý počet voľných parametrov. Garantuje konvergenciu k ideálnemu riešeniu a môže byť výpočtovo efektívna [21].

Metóda podporných vektorov je použitá na množstvo úloh strojového učenia ako je rozoznávanie vzorov, klasifikácia objektov a v prípade predikcií časových radov to je regresná analýza. Regresia založená na podporných vektoroch je postup, ktorého funkcia je predpovedaná pomocou nameraných dát, ktorými je metóda podporných vektorov natrénovaná. Toto je odklon od tradičných predpovedí časových radov v zmysle, že metóda podporných vektorov nepoužíva žiadny model, ale predikciu riadia samotné dáta [21].

Uvažujme množinu tréovacích dát $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \subset \chi \times \mathbb{R}$, kde v našom prípade hodnota x_i predstavuje odber elektrickej energie v čase i a množina χ označuje vstupnú množinu dát. Úlohou algoritmu je nájsť funkciu $f(x_i)$, pre ktorú hodnota ε nadobúda pre všetky tréovacie dáta čo najväčšie hodnoty oproti y_i a súčasne nadobúda najmenšie možné hodnoty menšie ako ε [25].

Matematicky môžeme túto lineárnu zapísať ako

$$f(x) = \langle w, x \rangle + b \quad (11)$$

Vo vzorci 11 platí $w \in \chi$ a $b \in \mathbb{R}$. Výsledok vektorového súčinu $\langle w, x \rangle$ sa nachádza v množine χ . Našou snahou je dosiahnuť čo najmenšiu hodnotu w . Zabezpečiť to môžeme minimalizovaním hodnoty $\|w\|^2 = \langle w, w \rangle$. Pri celi minimalizovať hodnotu $\frac{1}{2}\|w\|^2$, môžeme opisovaný problém zapísať ako

$$\begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases} \quad (12)$$

Predpokladom pre vzorec 12 je, že funkcia f existuje pre všetky páry (x_i, y_i) s presnosťou ε [25].

2.2.4 Rozhodovacie stromy

Rozhodovacie stromy sú jednou z najrozšírenejších učiacich metód. Používajú sa najmä na klasifikáciu, ale v súčasnosti sa využívajú aj na regresiu. Rozhodovací strom je reprezentovaný ako množina uzlov a im prislúchajúcich hrán. Uzly reprezentujú atribúty a výstupné hrany sú vždy označené konkrétnou hodnotou pre atribút, z ktorého vychádzajú. Rozhodovanie začína v koreni stromu a končí po dosiahnutí listového uzla. Pre riešenie jedného problému je možné vytvoriť stromy s rôznym počtom a usporiadaním uzlov. Najlepším riešením je strom s najmenším počtom rozhodovacích uzlov [20].

Hlavnou výhodou rozhodovacích stromov oproti ostatným modelom je, že strom produkuje model, ktorý je reprezentovateľný ako pravidlá alebo logické výroky. Taktiež klasifikácia sa môže vykonávať bez komplikovaných výpočtov. Táto technika môže byť použitá ako na kategorické premenné tak aj na spojité. Rozhodovacie stromy neposkytujú také dobré výsledky pre nelineárne dáta ako neurónové siete. Vo všeobecnosti je táto metóda vhodnejšia pre predpovedanie kategorických dát alebo na dáta, ktoré v sebe obsahujú viditeľný trend [26].

Regresný rozhodovací strom obsahuje odozvoový vektor Y reprezentujúci odozvu hodnotu ku každému meraniu v matici X . Vetvy sú rozdelené na základe štvorcového zostatkového minimalizačného algoritmu. Ten zabezpečuje, že očakávaný súčet variancií dvoch uzlov bude minimalizovaný. Algoritmus tak nájde optimálnu podmienku rozdelenia uzlu na jeho potomkov [3].

Priradením hodnoty 1 pre triedu k a hodnoty 0 pre ostatné triedy, získame varianciu rovnú $p(k|t)[1 - p(k|t)]$. Sčítaním K tried získame funkciu

$$i(t) = 1 - \sum_{k=1}^K p^2(k|t) \quad (13)$$

Tým sme vytvorili maximálny strom, čo znamená, že uzly sa rozdeľovali až do posledného merania, ktoré sa nachádzalo v trénovacej množine. Maximálny strom sa tak môže stať veľmi veľkým [3].

Jedným z najpoužívanejších algoritmov na rozdelenie uzlov sa nazýva index Gini alebo rozhodovacie pravidlo Gini. Tento algoritmus zapíšeme ako

$$i(t) = \sum_{k \neq l} p(k|t)p(l|t) \quad (14)$$

Vo vzorci 14 predstavujú k a l indexy tried o 1 po K a $p(k|t)$ označuje podmienenú pravdepodobnosť triedy k za predpokladu, že aktuálny uzol je uzol t . Aplikovaním Gini algoritmu nájdeme v trénovacej množine najväčšiu triedu, ktorú odizolujeme od ostatných dát [3].

Náhodné lesy sú kombináciou predpovedí stromov. Každý strom závisí od hodnoty náhodného vektora s rovnakým rozdelením. Chyba lesu závisí od sily jednotlivých stromov a koreláciou medzi nimi. Náhodný les môžeme definovať ako klasifikátor pozostávajúci z množiny stromov $\{h(x, \Theta_k), k = 1, 2, \dots\}$, kde $\{\Theta_k\}$ sú nezávislé rovnomerne distribuované náhodné vektory a každý strom sa podieľa hlasom na voľbe triedy vstupu x . S nárastom počtu stromov hodnota $\{\Theta_k\}$ konverguje k určitému bodu. Tým je zabezpečené, že náhodné lesy sa s pridávaním počtu stromov nepretrénujú, ale veľkosť chyby sa postupne ustáli.

Pri výbere náhodného vektora sa snažíme pri zachovaní jeho sily minimalizovať koreláciu, čím zvyšujeme presnosť celého výpočtu [4].

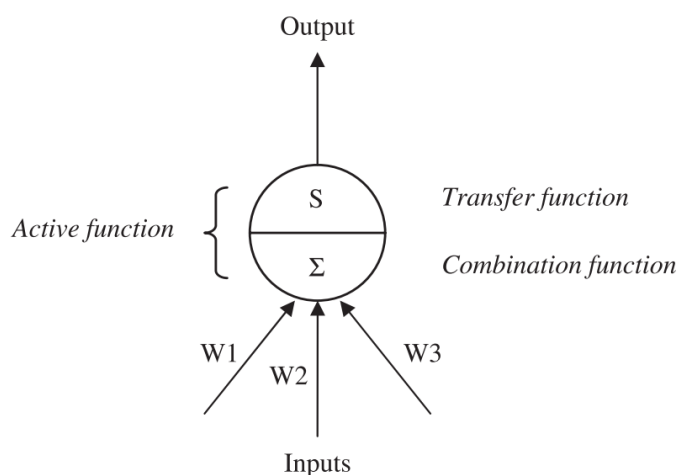
Väčšinou sú náhodné lesy používané na klasifikačné problémy, avšak je možné ich aplikovať aj na regresiu. Regresné náhodné lesy sú tvorené rastom stromov závislých na náhodnom vektore $\{\Theta\}$. Prediktor stromu $h(x, \Theta)$ nadobúda číselné hodnoty na rozdiel od štítkov tried ako je to pri klasifikačných problémoch. Predpokladáme trénovaciu množinu, ktorá je nezávislou distribúciou náhodného vektora Y, X . Potom môžeme strednú štvorcovú generalizačnú chybu pre číselný prediktor h matematicky zapísať ako

$$E_{X,Y}(Y - h(X))^2 \quad (15)$$

Prediktor náhodného lesu je tvorený priemerom k stromov, čo zapíšeme ako $h(x, \Theta_k)$ [4].

2.2.5 Neurónové siete

Návrh neurónových sietí je inšpirovaný neurofyziológiou ľudskeho mozgu. Model je analytická technika modelujúca procesy učenia v kognitívnych systémoch a neurologických funkciách mozgu. Má schopnosť predpovedať budúcu hodnotu merania konkrétnej premennej na základe hodnôt z predchádzajúcich meraní. Tento proces sa inak nazýva aj učenie z existujúcich dát. Tok v neurónovej sieti preteká cez jednotlivé neuróny. Na obrázku 6 môžeme vidieť príklad takéhoto neurónu [26].

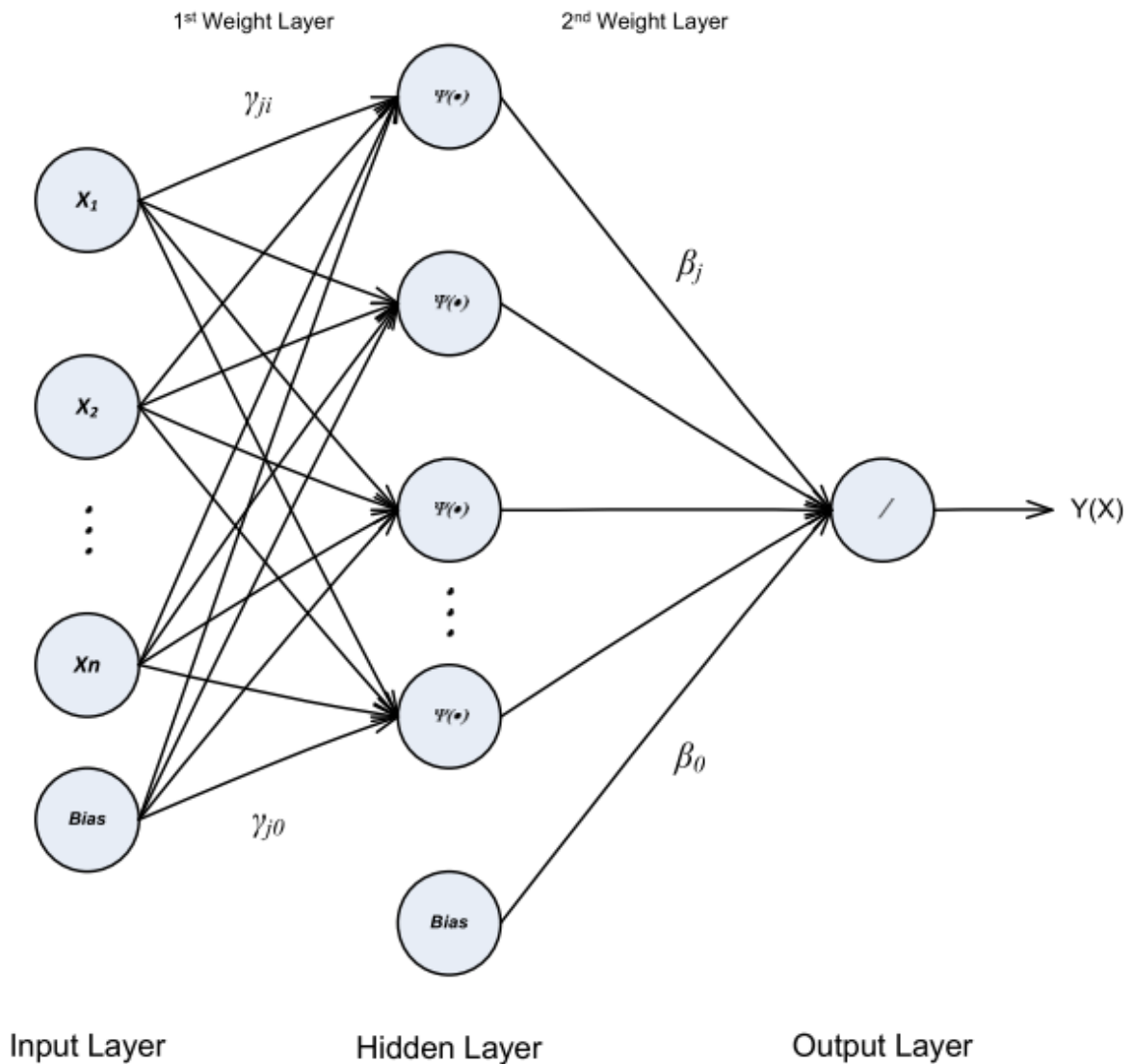


Obr. 6: Príklad neurónu v neurónovej sieti [26].

Neurónová sieť predstavuje orientovaný graf uzlov. Uzol neurónovej siete sa nazýva neurón. Každý uzol počíta svoj výstup na základe vstupov od susedných uzlov. Výpočet prebieha aplikovaním funkcie, ktorá sa nazýva sigmoid, na vážený súčet vstupov [10].

Trénovanie siete je proces nastavovania, čo najlepších váh na vstupy jednotlivých neurónov. Chyba neurónovej siete sa najčastejšie počíta pomocou spätnej propagácie (backpropagation), čím dostaneme rast chyby pre danú neurónovú sieť [26].

Je veľa typov neurónových sietí napríklad viacvrstvové perceptrónové siete, samoriadiace siete, siete s viacerými skrytými vrstvami, alebo aj viacvrstvové spätne propagované neurónové siete, pričom príklad takejto siete môžeme vidieť aj na obrázku 7. V každej skrytej vrstve je množstvo neurónov. Hlavnou výhodou je, že väčšina sietí nepotrebuje model. Na druhej strane, trénovanie obvykle zaberá veľa času. Výstupom siete je lineárna rovnica váh prepojených so vstupom [14].



Obr. 7: Príklad viacvrstvovej späťne propagovanej neurónovej siete [13].

Viacvrstvový perceptrón je jednou z najpoužívanejších neurónových sietí. Pozostáva z uzlov a im prislúchajúcim hranám. Uzly sú zoskupované do rôznych vrstiev. Prvá vrstva je vstupná vrstva, kde počet d označuje počet vstupných parametrov vstupujúcich do siete. Táto vrstva je následne prepojená hranami so skrytou vrstvou pozostávajúcou z h uzlov. Tá je potom prepojená s výstupnou vrstvou s c uzlami. Kvôli tomu sa tieto siete zvyknú označovať aj ako dopredné siete [20].

Elementy skrytých a výstupných vrstiev sú umelé neuróny pozostávajúce z uzlov, viacerých vstupujúcich a jednou výstupnou hranou. Funkciou neurónu je transformovať lineárnu kombináciu vstupov pomocou nelineárnej aktivačnej funkcie, čiže každú vstupnú hranu prenásobiť jej váhou a výsledok týchto súčinov sčítať. Tak dostaneme pre neurón j vzorec 16 opisujúci lineárnu kombináciu vstupov a_j

$$a_j = \sum_{i=1}^d w_{ji} x_i \quad (16)$$

Tabuľka 1: Tabuľka rozdelenia vybraných prírodne inšpirovaných algoritmov [8].

Model ľudskej mysle	Umelé imunitné systémy	Rojová inteligencia	Ostatné napr. prírodné úkazy
Teória fuzzy množín	Genetický algoritmus	Optimalizácia kolóniu mravcov	Tektonické dosky
Teória približných množín	Umelé neurónové siete	Optimalizácia rojom častíc	Veľký kolaps
Granulárne výpočty	Celulárny automat	Inteligentná kvapka vody	Oceánske prúdy
Predbežné výpočty	Membránové výpočty	Včelí zhukovací algoritmus	Prílivové vlny
		Vyhľadávanie kukučkou	Sopečné erupcie
			Zemetrasenie

Pričom váha w_{ji} označuje váhu medzi neurónom i na vstupnej vrstve a neurónom j na skrytej vrstve [20].

Aktivovanie neurónu j závisí od jeho aktivačnej funkcie $g(a_j)$. Jednu z najpoužívanějších aktivačných funkcií, logistickú sigmoidnú funkciu, môžeme matematicky zapísať ako

$$g(a) \equiv \frac{1}{1 + \exp(-a)} \quad (17)$$

Je zrejmé, že funkcia zo vzorca 17 vracia hodnoty v rozmedzí $(0, 1)$ [20].

2.3 Analýza optimalizačných algoritmov

Optimalizačné algoritmy, ktoré majú potenciál nájsť globálne alebo lokálne riešenie problému. Lokálne optimalizácie, nazývané aj vyhľadávacie algoritmy, sa pokúšajú nájsť lokálne minimum v okolí štartovacieho riešenia. Väčšina týchto algoritmov je deterministická. Pri hľadaní minima používajú vyhodnocovaciu funkciu, na základe ktorej aktualizujú doterajšie riešenie. Z nových možných riešení je najlepšie to s najnižšou hodnotou, čiže to, ktoré je najlacnejšie. Kvôli tejto vlastnosti sa zvyknú tieto algoritmy označovať aj ako nenásytne algoritmy (greedy algorithms). Spravidla tieto algoritmy nenájdu globálne minimum v prípade, že sa nachádza ďalej od štartovacieho riešenia ako nejaké lokálne minimum [23].

Na druhej strane, optimalizačné algoritmy s potenciálom nájdenia globálneho minima nachádzajú riešenie, ktoré postupne konverguje k optimálnemu riešeniu. To ale nie je algoritmami úplne garantované. Algoritmy s globálnym potenciálom majú väčší prehľad o svojom okolí, a preto uviaznutie v lokálnom minime je zriedkavé [23].

Nasledujúca tabuľka 1 zobrazuje rozdelenie algoritmov s potenciálom nájdenia globálneho minima. Algoritmy sú prírodne inšpirované a ich ďalšie delenie vyplýva zo spoločných znakov, na ktorých sú založené.

2.3.1 Genetický algoritmus

Genetický algoritmus patrí medzi prírodne inšpirované algoritmy patriace do triedy umelých imunitných systémov. Genetický algoritmus je stochastický optimalizačný algoritmus, ktorého úlohou je nájdenie globálneho riešenia pre zadaný problém, čiže sa nestane, že riešenie

spadne do lokálne minima a nenájde sa tak optimálne riešenie. Od tradičných algoritmov sa líšia hlavne v počte riešení, ktoré sú kandidátmi na najlepšie riešenie. Tradičné vyhľadávacie algoritmy prehľadávajú dôkladne iba jedno riešenie, zatiaľ čo genetické algoritmy hlbšie spracujú viacero kandidátov naraz. Každý kandidát na optimálne riešenie problému je reprezentovaný dátovou štruktúrou, ktorú označujeme pojmom jedinec. Súbor jedincov tvorí populáciu. Začiatok procesu začína náhodnými riešeniami populácie, ktorý sa postupne zlepšuje [5].

Vytvorenie novej generácie sa vykonáva pomocou genetických operátorov: a to selekciou, krížením a mutáciou. Proces selekcie vyberie kvalitnejšie chromozómy, ktoré prežijú a vyskytnú sa tak aj v ďalšej generácii [24].

Pri genetických algoritmoch sa zavádzajú pojmy ako chromozóm, fitness funkcia, kríženie, elitárstvo, operátor reprodukcie či mutácie [5].

Chromozóm je pomenovanie pre jedinca. V literatúre sa tiež zvykne používať označenie kandidát [2].

Inicializácia je proces, ktorý po vygenerovaní generácie priradí kandidátom náhodné hodnoty. Pri prehľadávaní dvojrozmerného priestoru to budú náhodné hodnoty oboch súradníc [16].

Fitness funkcia je funkcia určujúca efektívnosť chromozómu. Každý jedinec v množine je ohodnotený pomocou tejto funkcie, ktorá vracia číselnú reprezentáciu riešenia, čiže ako dobre daný kandidát vyriešil zadaný problém. Pri hľadaní optimálneho riešenia sa porovnáva hodnota fitness funkcie aktuálneho riešenia s hodnotou funkcie cieľového riešenia, ale vo všeobecnosti, čím je hodnota väčšia, tým je kandidátovo riešenie lepšie [5, 16, 24].

Operátor reprodukcie je obvykle prvý operátor, ktorý sa uplatní na populáciu. Operátor náhodne vyberie reťazce z dvoch chromozómov na párenie [5].

Kríženie je operátor kombinácie. Kríženie vykonáva výmenu blokov chromozómov. Z druhého chromozómu je vybraný reťazec náhodnej veľkosti, ktorý sa vymení s rovnako dlhým reťazcom z prvého chromozómu. V tabuľke 2 je tento proces znázornený graficky [5].

Tabuľka 2: Proces vytvorenia novej generácie z rodičovských chromozómov.

rodič č. 1	0110	0101 0010	0011
rodič č. 2	1100	1101 1110	1111
potomok č. 1	0110	1101 1110	0011
potomok č. 2	1100	0101 0010	1111

Vzniká problém ako čo najvhodnejšie určiť bod kríženia a vybrať tak najkvalitnejšie bloky chromozómu. Kvôli tomu je potrebné definovať nový element nazývaný väzba, označovaný ako d_i

$$d_i = \frac{a_i + a_{i+1}}{2} \quad (18)$$

Za predpokladu, že chromozóm reprezentujeme ako maticu veľkosti $1 \times N$, potom väzbu definovanú vzorcom 18 môžeme reprezentovať ako maticu $1 \times (N - 1)$. Vypočítaním priemeru susedných hodnôt aj pre cieľovú maticu a nájdením priemeru matice, určíme body kríženia. Chromozóm rozdelíme na miestach, kde je väzba, čo najmenšia [24].

Elitárstvo je proces pridávania chromozómov s najlepšou hodnotou funkcie fitness priamo do ďalšej populácie. Zaisťuje to, že najlepšie riešenie budúcej generácie bude vždy lepšie alebo pri najhoršom rovnaké, ako najlepšie riešenie predchádzajúcej generácie [6].

Operátor mutácie sa vykoná po vykonaní operátora reprodukcie. Mutácia chromozómu väčšinou predstavuje operáciu, ktorá s nízkou pravdepodobnosťou invertuje jeden bit v chromozóme [5].

Princíp genetického algoritmu môžeme znázorniť v nasledujúcom pseudokóde [5]

Algorithm 1 Pseudokód genetického algoritmu

- 1: Náhodne inicializovanie jedincov
 - 2: Vyhodnotenie fitness funkcie pre každého jedinca
 - 3: Výber jedincov pre ďalšiu populáciu na základe fitness funkcie
 - 4: Kríženie jedincov
 - 5: Mutovanie jedincov
 - 6: Ak bolo nájdené žiadané optimálne riešenie pokračuj, inak návrat na krok 2
 - 7: Vráť optimálne riešenie
-

Veľkou výhodou genetického algoritmu je, že mutácia predchádza sklznutiu do lokálnych miním a kombinácia chromozómov vedie k rýchlemu približovaniu sa k optimálnemu riešeniu. Napriek týmto výhodám, majú genetické algoritmy aj niekoľko nevýhod [6].

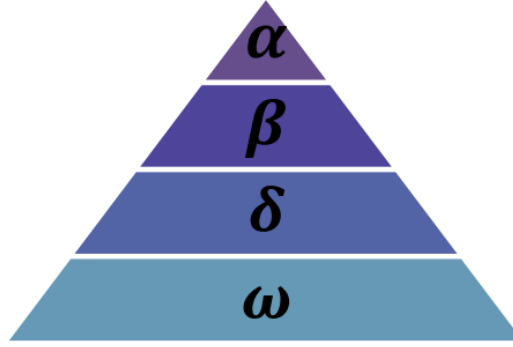
- Reprezentovanie kandidátov je príliš obmedzujúce
- Mutácia a kríženie sú v súčasnosti aplikovateľné iba na chromozómy reprezentované bitovým reťazcom alebo číslami
- Definovanie fitness funkcie je často netriviálnou záležitosťou a jej generalizácia je náročná

2.3.2 Optimalizácia svorkou divých vlkov

Algoritmus je založený na správaní vlka sivého, ktorý je na vrchole potravinového reťazca a preferuje život vo svorke. Lov pozostáva zo stopovania, prenasledovania, približovania sa, obklúčenia koristi a útoku na korisť. Vlkov vo svorke možno rozdeliť do niekoľkých skupín. V obrázku 8 je znázornená hierarchia týchto skupín [22].

Alfa vlky sú vodcami svorky. Ich úlohou je robiť rozhodnutia ohľadom lovu, miesta na spanie či času zobudenia. Tieto príkazy diktujú svorke, avšak bolo pozorované aj demokratické správanie, kedy alfa vlky nasledovali ostatných členov svorky. Všetky vlky uznávajú postavenie alfy vo svorke. Zaujímavosťou je, že vodcom nemusí byť najsilnejší jedinec, ale môže to byť aj jedinec najlepší v organizovaní svorky [22].

Beta vlky sú druhým stupňom v hierarchii, podriadené alfa vlkom, ktorým pomáhajú v rozhodovaní a organizácii. Sú najlepšími kandidátmi na alfu v prípade úmrtia alebo zosarnutia alfa jedincov. Hrajú rolu radcu a ďalej distribuujú príkazy a vracajú sa s odozvou na ne [22].



Obr. 8: Hierarchia vlkov vo svorke [22].

Delta vlky inak nazývané aj podriadené, zastupujú vo svorke úlohy prieskumníkov, strážcov, starejších, lovcov či opatrovateľov. Prieskumníci hliadkujú hranice a varujú svorku pred nebezpečím. Strážcovia sa starajú o bezpečie svorky. Starejší sú skúsenými vlkmi, ktorí boli na pozícii alfa alebo beta. Lovci pomáhajú alfa a beta vlkom pri love. Opatrovatelia sa starajú o slabé, choré alebo zranené jedince [22].

Omega vlky majú najnižšiu hodnotu vo svorke. Hrajú rolu obetných baránkov, ktoré sa podrobia ostatným. K potrave sa dostanú ako úplne posledné. Aj keď sa možno javí ich postavenie zbytočné, boli pozorované prípady, kedy ich strata spôsobila vo svorke nedorozumenia [22].

Spoločenská hierarchia reprezentovaná matematickým modelom, označuje najvhodnejšie riešenie ako alfa, druhé a tretie najvhodnejšie ako beta resp. delta. Riešenia ostatných kandidátov označujeme ako omega. Algoritmus optimalizácie (v prírode lovu) je vedený alfa, beta a delta kandidátmi, ktorí sú nasledovaní kandidátmi omega [22].

Obkľúčenie koristi jednotlivými vlkmi α , β a δ môžeme matematicky vyjadriť nasledujúcimi rovnicami

$$\vec{D}_\alpha = |\vec{C}_1 \cdot \vec{X}_\alpha - \vec{X}| \quad (19)$$

$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1 \cdot \vec{D}_\alpha \quad (20)$$

$$\vec{D}_\beta = |\vec{C}_2 \cdot \vec{X}_\beta - \vec{X}| \quad (21)$$

$$\vec{X}_2 = \vec{X}_\beta - \vec{A}_2 \cdot \vec{D}_\beta \quad (22)$$

$$\vec{D}_\delta = |\vec{C}_3 \cdot \vec{X}_\delta - \vec{X}| \quad (23)$$

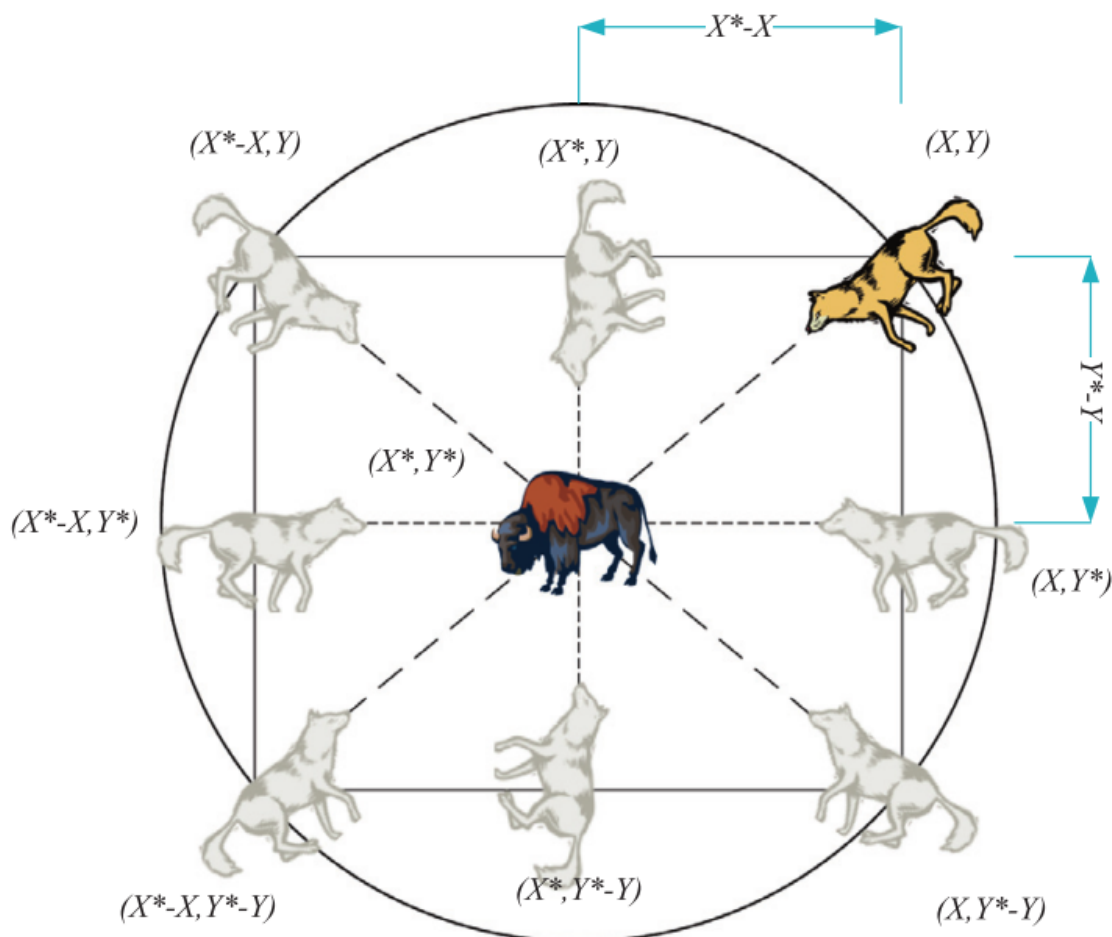
$$\vec{X}_3 = \vec{X}_\delta - \vec{A}_3 \cdot \vec{D}_\delta \quad (24)$$

Vo vzorcoch 20, 22 a 24 predstavujú vektory \vec{X}_1 , \vec{X}_2 a \vec{X}_3 polohu koristi a vektory \vec{X}_α , \vec{X}_β a \vec{X}_δ polohu vlkov α , β a δ . Vektory \vec{A} a \vec{C} sú koeficienty, ktoré zapíšeme ako

$$\vec{A} = 2\vec{a} \cdot \vec{r}_1 - \vec{a} \quad (25)$$

$$\vec{C} = 2 \cdot \vec{r}_2 \quad (26)$$

Vo vzorci 25 sa premenná \vec{a} počas výpočtu lineárne znižuje od 2 po 0. Vektory \vec{r}_1 a \vec{r}_2 vo vzorci 26 sú náhodnými vektormi v rozsahu $[0, 1]$. Vlk môže svoju pozíciu (X, Y) aktualizovať v závislosti od pozície koristi (X^*, Y^*) . Obrázok 9 ilustruje možné aktualizované pozície, ktoré môže najlepší agent dosiahnuť. Tieto pozície získame aplikovaním vzorcov 19 až 24 [22].



Obr. 9: Príklad možného umiestnenia vlka v dvojrozmernom priestore na základe polohy koristi [22].

Hľadánie koristi je zabezpečené vektorom \vec{A} , ktorý mimo intervalu $[-1, 1]$ núti vlky divergovať od seba, čím je zdôraznená potreba prehl'adávať okolie a neskĺznuť do lokálneho minima. Náhodný vektor \vec{C} simuluje prekážky v prírode, s ktorými sa vlk stretne pri hľadaní koristi. V závislosti od vygenerovanej hodnoty, môže simulovať aj opačný prípad, ktorý je pre vlka priaznivejší [22].

Lov je vedený alfa vlkami, beta a delta vlky sa tiež môžu na ňom príležitostne podieľať. V prírode disponujú schopnosťou rozpoznať umiestnenie koristi a obkľúčiť ju, avšak pri simulácii tohto správania, nemáme vedomosť o presnej polohe koristi. Preto na základe prvých

troch najlepších riešení vypočítame predpokladanú polohu koristi, čím vznikne vzorec 27. Kandidáti, vrátane omega vlkov, potom na základe rovníc 20, 22 a 24 aktualizujú svoju polohu [22].

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad (27)$$

Útok na koristi v matematickom modeli dosiahneme približovaním sa ku koristi, čiže znížením hodnoty \vec{a} . Tak dosiahneme aj zníženie hodnoty \vec{A} až bude jeho hodnota v intervale $[-1, 1]$. Potom ďalšia vypočítaná hodnota pozície vlka sa bude nachádzať medzi jeho pôvodnou polohou a polohou koristi. V prípade, že hodnota \vec{A} sa nachádza mimo spomínaného intervalu, vlk diverguje od koristi, čo nastáva vo fáze jej hľadania. Tým je zabezpečené prehľadávanie priestoru agentami [22].

Princíp optimalizácie svorkou divých vlkov môžeme znázorniť v nasledujúcom pseudokóde [22]

Algorithm 2 Pseudokód optimalizácie svorkou divých vlkov

- 1: Náhodná inicializácia vlkov
 - 2: Inicializácia a , A a C
 - 3: Vyhodnotenie fitness funkcie pre každého jedinca
 - 4: Priradenie do 3 najlepších riešení do X_α , X_β resp. D_δ
 - 5: Opakuj kroky 5 až 10 pokiaľ nebude prekročený maximálny počet iterácií
 - 6: Aktualizovanie polohy na základe rovníc 20, 22 a 24
 - 7: Aktualizovanie a , A a C
 - 8: Vyhodnotenie fitness funkcie pre každého jedinca
 - 9: Aktualizovanie X_α , X_β a D_δ
 - 10: Aktualizovanie počtu iterácií
 - 11: Vráť optimálne riešenie X_α
-

2.3.3 Umelá kolónia včiel

V literatúre označovaný ako ABC algoritmus (Artificial bee colony) je pomerne nový medzi rojovými algoritmami. Princíp je založený na biologickom procese správaní medonosných včiel pri hľadaní potravy. Každá včela pracujúca v roji sa spolupodieľa na tvorbe celého systému na globálnej úrovni. Správanie systému je určené lokálnym správaním, kde spolupráca a zladenie jedincov vedie k štruktúrovanému kolaboračnému systému [5].

V algoritme, kolónia umelých včiel pozostáva z 3 skupín. Sú to včely robotnice, diváčky a prieskumníčky. Informácie o zdrojoch potravy si vymieňajú tancovaním (waggle dance) na tzv. tanečnej ploche. Diváčky čakajúce na tanečnej ploche sa rozhodujú, ktorý zdroj potravy si vyberú. Robotnice najskôr tieto zdroje navštívia a prieskumníčky uskutočňujú náhodné prehľadávanie priestoru. Pri aplikovaní ABC algoritmu, polovica včiel predstavuje robotnice, druhá polovica diváčky. Počet robotníc zodpovedá počtu zdrojov potravy v okolí úľu. Robotnica, ktorej zdroj je vyčerpaný inými včelami sa stáva prieskumníčkou [12].

Zväčšovaním množstva zdrojov potravy sa zväčšuje aj pravdepodobnosť vybratia zdroja diváčkou. Vďaka tomu je zabezpečené, že tanec robotníc, ktoré navštívili zdroje s najväčším množstvom nektáru, presvedčí najviac diváčok. Tie na základe informácie o polohe zdroja nájdu v jeho okolí nový zdroj, z ktorého budú zberať nektár. Po vyčerpaní nektáru je tento zdroj opustený a včela sa presunie na zdroj, ktorý našli prieskumníčky [12].

Pozícia zdrojov jedla je reprezentácia možného riešenia daného problému. Množstvo nektáru je úmerné kvalite riešenia, ktoré je určené fitness funkciou. Počet robotníčok a diváčok je rovný počtu riešení v danej populácii, pričom každé je reprezentovateľné ako D -dimenzionálny vektor. Hodnota D môže predstavovať počet optimalizačných parametrov [12].

Včely majú riešenie uložené vo vlastnej pamäti, z ktorého modifikáciou vznikajú nové, ktoré následne vyskúšajú. Ak je nové riešenie lepšie ako predchádzajúce, včely si ho zapamätajú a predchádzajúce zabudnú. Inak si pamätajú staré. Keď všetky robotnice skončia proces hľadania, zdieľajú informácie o zdrojoch s diváčkami. Tie vyhodnotia zhromaždené informácie a vyberú zdroj s najlepším riešením a jeho modifikovaním vytvoria nové, ktoré následne skontrolujú [12].

Diváčky vyberajú zdroj v závislosti od pravdepodobnosti zdroja p_i , ktorú vypočítame ako

$$p_i = \frac{fit_i}{\sum_{j=1}^n fit_j} \quad (28)$$

Vo vzorci 28 predstavuje fit_i fitness funkciu, ktorá ohodnocuje riešenie i . Týmto spôsobom si vymieňajú medzi sebou informácie robotnice a diváčky [12].

Algoritmus počíta nové riešenie na základe starého, čo matematicky zapíšeme ako

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \quad (29)$$

Ak si počet včiel definujeme ako B potom na základe vzorca 29 platí $k \in 1, 2, \dots, B$ a $j \in 1, 2, \dots, D$. Indexy k a j sú vybrané náhodne, ale k musí byť rôzne od i . Hodnota ϕ_{ij} je z intervalu $[-1, 1]$ vybraná náhodne. Je zrejmé, že čím bude menší rozdiel medzi $(x_{ij}$ a $x_{kj})$, tým bude menší rozdiel aj medzi starým a novým riešením [12].

Algoritmus má 3 konfigurovateľné parametre a to: počet robotníčok alebo diváčok, ktorý je rovný počtu zdrojov jedla, limit, ktorý ohraničuje prehľadávaný priestor a maximálny počet cyklov, ktorý sa vykoná ak sa skôr nenájde optimálne riešenie [12].

Princíp umelej kolónie včiel znázorňuje nasledujúci pseudokód [12]

Algorithm 3 Pseudokód umelej kolónie včiel

- 1: Inicializácia včiel na náhodné zdroje potravy
 - 2: Opakuj kroky 2 až 7 pokiaľ nebude dosiahnutý cieľ
 - 3: Umiestnenie robotníčok na zdroje potravy, ktoré boli nájdené
 - 4: Vyhodnotenie zdroja na základe množstva nektáru
 - 5: Umiestnenie diváčok na zdroje potravy na základe získaných informácií
 - 6: Vyhodnotenie, ktoré včely sa stanú prieskumníčkami
 - 7: Vyslanie prieskumníčok do prehľadávanie priestoru za účelom objavenia nových zdrojov potravy
 - 8: Vráť optimálne riešenie
-

2.4 Meranie presnosti predpovede

Pre vyhodnotenie efektívnosti a presnosti modelov je potrebné merať ich vlastnosti tak, aby sme ich vedeli medzi sebou porovnávať. V nasledujúcich spôsoboch merania sú použité pojmy ako aktuálna hodnota y_t , predpovedaná hodnota f_t alebo chyba predpovede e_t definovaná ako $e_t = y_t - f_t$. Veľkosť testovacej množiny budeme označovať ako n [1].

2.4.1 Stredná chyba predpovede

V literatúre označovaná ako MFE (mean forecast error). Matematickú funkciu môžeme zapísať ako

$$MFE = \frac{1}{n} \sum_{t=1}^n e_t \quad (30)$$

Týmto spôsobom meriame priemernú odchýlku predpovedanej hodnoty od aktuálnej. Zistíme tak smer chyby. Nevýhodou je, že kladné a záporné chyby sa vynulujú a potom nie je možné zistiť presnú hodnotu chyby. Pri nameraní extrémnych chýb nedochádza k žiadnej špeciálnej penalizácii. Taktiež hodnota chyby závisí od škály meraní a môže byť ovplyvnená aj transformáciami dát. Dobré predpovede majú hodnotu blízku 0 [1].

2.4.2 Stredná absolútna chyba

V literatúre označovaná ako MAE (mean absolute error). Patrí k jedným z najpoužívanějších. Funkciu môžeme zapísať ako

$$MAE = \frac{1}{n} \sum_{t=1}^n |e_t| \quad (31)$$

Týmto spôsobom meriame priemernú absolútnu odchýlku predpovedanej hodnoty od aktuálnej. Zistíme tak celkový rozsah chyby, ktorá nastala počas predpovede. Na rozdiel od merania chyby pomocou vzorca 30 sa kladné a záporné chyby nevynulujú, no ani napriek tomu nevieme určiť celkový smer chyby. Na druhej strane tiež nenastáva žiadna penalizácia pri extrémnych chybách. Hodnota chyby závisí od škály meraní a transformácií dát. Dobré predpovede majú hodnotu čo najbližšiu 0 [1, 11].

2.4.3 Stredná percentuálna chyba

V literatúre označovaná ako MPE (mean percentage error). Matematicky môžeme túto funkciu zapísať ako

$$MPE = \frac{1}{n} \sum_{t=1}^n \frac{e_t}{y_t} \times 100 \quad (32)$$

Vlastnosti sú veľmi podobné ako pri MAPE v časti 2.4.4. Chyba nám poskytuje prehľad o priemernej chybe, ktorá sa vyskytla počas predpovede. Navyše oproti MAPE získame prehľad o smere chyby, čo má však za následok, že opačné znamienka sa vynulujú. O modely, ktorého chyba MPE sa blíži k 0, nemôžeme s určitosťou tvrdiť, že funguje správne [1].

2.4.4 Stredná absolútna percentuálna chyba

V literatúre označovaná ako MAPE (mean absolute percentage error). Vzorec, ktorým ju zapíšeme bude veľmi podobný vzorcu 32

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{e_t}{y_t} \right| \times 100 \quad (33)$$

Pomocou tohto merania chyby získavame percentuálny prehľad o priemernej absolútnej chybe, ktorá sa vyskytla počas predpovedi. Veľkosť chyby nezávisí od škály merania, ale je závislá od transformácií dát. Tiež nie je možné zistiť smer chyby a ani nenastáva žiadna penalizácia pri extrémnych chybách [1].

2.4.5 Stredná štvorcová chyba

V literatúre označovaná ako MSE (mean squared error). Vzorcom ju zapíšeme ako

$$MSE = \frac{1}{n} \sum_{t=1}^n e_t^2 \quad (34)$$

Chyba meria priemernú štvorcovú odchýlku predpovedanej hodnoty. Opačné znamienka sa neovplyvňujú. Neposkytuje nám pohľad na smer chyby. Zabezpečuje penalizáciu extrémnych chýb. Zdôrazňuje fakt, že celková chyba je viac ovplyvnená jednotlivými veľkými chybami ako viacerými malými. Nevýhodou je, že chyba je veľmi citlivá na zmenu škály alebo transformáciu dát [1].

2.5 Zhodnotenie analýzy

S rastúcim množstvom dát z rôznych zdrojov vzniká potreba predpovedať správanie týchto veličín v budúcnosti. Existuje množstvo článkov, ktoré sa zaoberajú týmto problémom. Neustále prinášajú a vylepšujú existujúce matematické modely, ktoré sú schopné predpovedať budúce hodnoty meraných veličín. Výsledky nie sú presné, ale chyba výpočtu je dostatočne nízka na to, aby sme takýto výsledok mohli považovať za relevantný. Zvýšiť presnosť výsledkov je možné viacerými metódami, či už kombináciou viacerých matematických modelov alebo ich optimálnym nastavením. My sme sa v práci zamerali na hľadanie optimálneho nastavenia používaných modelov.

Preskúmané predikčné modely potrebujú pre svoje správne fungovanie vstupné parametre. Pomocou nich vieme ovplyvniť chybu predikcie. Problémom je nájdenie takých parametrov, pre ktoré by bola chyba predikcie čo najmenšia. Hľadanie najlepšieho riešenia by bolo časovo neprípustné, a preto budeme hľadať iba optimálne riešenie, ktoré nám poskytne prijateľnú chybu predikcie. Je dôležité poznamenať, že vstupné parametre, ktoré nájdeme optimálnu predikciu závisia aj od samotných dát. Ako bolo spomenuté v sekcii 2.1.1, dáta pozostávajú z viacerých zložiek a ich pomer sa môže líšiť pri rôznych veličinách.

Na hľadanie optimálneho nastavenie predikčných algoritmov sme použili prírodne inšpirované optimalizačné algoritmy. Ako ich názov napovedá, hľadanie optimálneho riešenia sa vykonáva na základe správania sa nejakého živočíšneho druhu alebo prírodného javu. Tieto algoritmy sa osvedčili ako efektívne a rýchle riešenie problémov, ktorých prehládavaný priestor riešení nie je možné prehládať celý. Algoritmy sa vyhýbajú spadnutiu do lokálnych miním a tak je nájdené riešenie obvykle optimálne v globálnom rozsahu. A tak tiež tieto algoritmy poskytujú niekoľko konfiguračných parametrov ovplyvňujúce rýchlosť a presnosť nájdeného riešenia. Avšak cieľom tejto práce je optimalizovať predikčné a nie optimalizačné algoritmy a poskytnúť tak používateľovi univerzálne a jednoduché riešenie predikčných problémov.

Dôležitou súčasťou predpovedí je vyhodnotenie ich úspešnosti porovnávaním predpovedanej hodnoty s ich skutočnými nameranými hodnotami. Existuje množstvo metód, ktorými vieme zmerať chybu predpovede. Niektoré nám poskytujú informáciu o smere chyby, iné zohľadňujú extrémne chyby. Výsledky niektorých sa viažu na škálu, v ktorej sa nachádzajú naše merania, iné sú nezávislé, merané v percentách. Naším cieľom bolo poskytnúť používateľovi čo najviac informácií o presnosti predpovede, a preto sú použité viaceré metódy merania chýb predikcií.

3 Špecifikácia požiadaviek

Cieľom práce je navrhnuť systém, ktorý umožní používateľovi nájsť optimálne nastavenie vstupných parametrov predikčných metód. Používateľ bude mať k dispozícii niekoľko predikčných a optimalizačných algoritmov, bude môcť meniť ich konfiguračné parametre a vykonávať výpočty, vďaka ktorým zistí okrem veľkosti predikčnej chyby aj optimálne nastavenie parametrov. Tieto výsledky môže použiť vo vlastnej aplikácii alebo výskume. Predmetom optimalizácie budú parametre predikčných metód, konfigurácia optimalizačných algoritmov je ponechaná na používateľovi. Množinu dát si sám zvolí, aplikácia je nezávislá na doméne, v ktorej boli dáta namerané. Ku každému z uvedených parametrov bude poskytnuté vysvetlenie a opis zmeny správania na základe ich zmeny. Presnosť predpovede bude určená pomocou používaných metrík na meranie chyby predikcie. Keďže metrík je viacero, používateľ bude mať možnosť si vybrať jednu z nich a tá bude následne použitá optimalizačným algoritmom.

Pri návrhu aplikácie boli použité dáta z inteligentných elektromeračov používaných na Slovensku. Dostupné sú každých 15 minút z rôznych obcí. Keďže budeme vždy predpovedať práve jednu veličinu práve jedného merača, každé meranie potom bude jednoznačne identifikovateľné pomocou dátumu a času, kedy bolo uskutočnené. Nameraná hodnota bude reprezentovaná reálnym číslom. Našou úlohou nie je optimalizovať predpovede na základe faktorov, ktoré ich ovplyvňujú a teda merania, ktoré majú k dispozícii aj hodnotu nejakej vysvetľujúcej premennej budú spracovávané rovnakým spôsobom ako tie bez nej. Fungovanie aplikácie nie je závislé na doméne, z ktorej dáta pochádzajú. Používateľ tak môže optimalizovať predikčné parametre aj na základe iných dát ako z elektromeračov.

Používateľ si bude môcť vyberať z ponúknutých predikčných a optimalizačných algoritmov. K dispozícii bude mať stochastické modely, regresiu založenú na podporných vektoroch, rozhodovacie stromy a prírodné inšpirované optimalizačné algoritmy ako umelú kolóniu včiel alebo optimalizáciu založenú na správaní roja častíc. Systém bude navrhnutý tak, aby pridanie ďalších algoritmov nespôsobovalo zmeny v pôvodnej aplikácii a pozostávalo tak iba z pridania implementácie algoritmov a prípadne zmeny konfigurácie. Prípadá do úvahy, že samotný systém bude poskytovať rozhranie, ktoré používateľovi umožní použiť vlastnú implementáciu iných algoritmov.

Optimalizačné algoritmy potrebujú mať určený definičný obor. Hodnoty z neho používajú na nájdenie optimálneho riešenia, čo je v našom prípade nájdenie kombinácie vstupných parametrov predikčných metód, ktorých výsledná chyba je čo najmenšia. Z tohto dôvodu je dôležité myslieť pri návrhu aplikácie na to, aby výstupom zvolených metód na meranie chyby boli nezáporné čísla. V aplikácii sa budú nachádzať prednastavené definičné obory, ktoré bude môcť používateľ v definovanom rozmedzí upravovať. Predídeme tak dlhým odozvám systému pri zvolení veľkého definičného oboru a zároveň nevzniknú situácie, kedy používateľ zvolí hodnotu, ktorá nebude patriť do definičného oboru predikčnej metódy.

Systém bude implementovaný v jazyku R a webové používateľské rozhranie bude vytvorené pomocou knižnice Shiny. Web aplikácie bude navrhnutý tak, aby používateľovi poskytoval interaktívne rozhranie s predikčnými metódami. Bude dostatočne robustný a intuitívny, aby s jeho používaním nemali problémy ani tí menej skúsení. Zároveň bude skúsenému používateľovi poskytovať modulárnosť, konfigurovateľnosť, ale hlavne rozhranie pre jednoduché pridávanie ďalších algoritmov, ktoré si môže sám implementovať.

4 Návrh riešenia

Výsledná webová aplikácia bude používateľovi poskytovať možnosť zvoliť si optimalizačný algoritmus, ktorým bude optimalizovať nastavenie predikčného algoritmu. Keďže každý predikčný algoritmus potrebuje pre správnu funkcionálnosť iné parametre s rôznymi hodnotami, je najskôr potrebné implementovať rozhranie, ktoré zabezpečí správnu interakciu s optimalizačným algoritmom. Používateľ preto bude vyberať iba z algoritmov, ktoré budú predom implementované alebo tak sám urobí.

V práci som sa rozhodol zvoliť ako predikčné algoritmy metódu podporných vektorov, autoregresívny integrovaný model kľzavého priemeru a náhodné lesy. Aplikácia je však navrhnutá genericky a pridanie ďalšieho algoritmu vyžaduje minimálne úsilie. Kvôli malému počtu implementovaných optimalizačných algoritmov v jazyku R som zvolil optimalizáciu rojom častíc a ABC optimalizáciu. Používateľ bude môcť pomocou rozhrania určiť parametre pre tieto algoritmy, ako napr. počet častí alebo cyklov.

4.1 Vlastnosti vstupných dát

Vstupné dáta, ktoré vkladá do aplikácie používateľ by mali byť vo formáte CSV, teda súbor s tabuľkovými dátami, kde hodnoty v stĺpcoch sú oddelené čiarkami. Súbor by mal pri tom obsahovať aspoň dva pomenované stĺpce, kde stĺpec “timestamp” obsahuje dátum a čas získania hodnoty, ktorá sa nachádza v stĺpci “value”. Hodnoty v ostatných stĺpcoch budú ignorované. Zámerom bolo, aby zápis dátumu a času zodpovedal medzinárodnému štandardu ISO 8601. Kvôli zložitosti čítania času v takomto formáte v použítom implementačnom prostredí som zvolil vlastný formát dátumu, ktorý mu je veľmi podobný. Formát dátumu je “%Y-%m-%d %H:%M:%S”, podrobný rozbor sa nachádza v tabuľke 3. Z pohľadu návrhu aplikácie je potrebné, aby vstupné dáta mali rovnaký počet meraní nameraných v jeden deň v celej vstupnej množine. To znamená, že veľkosť periódy dát by mala byť rovnaká a zároveň sa rovnala veľkosti periódy, ktorú zvolil používateľ vo webovej aplikácii.

Tabuľka 3: Rozbor formátu dátumu a času.

Názov	Formátový reťazec	Hodnota
Rok	%Y	2013
Mesiac	%m	07
Deň	%d	10
Hodina	%H	20
Minúta	%M	15
Sekunda	%S	15

* Dátum z tabuľky je “2013-07-10 20:15:00”

Stĺpec s hodnotou obsahuje reálne číslo, ktoré reprezentuje veľkosť nameranej veličiny, ktorú sa bude aplikácia snažiť predpovedať. Jednotky, v ktorých budú hodnoty uvedené nie sú podstatné, výpočet je nezávislý na použitej škále. Obmedzenia na tento stĺpec sú, aby skutočne obsahoval iba reálne číslo zodpovedajúce nameranej hodnote a aby v rámci jedného súboru bola konzistentne použitá rovnaká škála hodnôt. V stĺpci sa tak nesmie nachádzať názov či skratka jednotky, v ktorej je hodnota nameraná.

V prípade, že pri vkladaní súboru aplikácia nenájde spomínané stĺpce, nedovolí používateľovi spustiť výpočet. Pri návrhu aplikácie sa predpokladalo, že vstupné dáta budú vždy pochádzať z jedného zariadenia a budú teda mať rovnakú časovú zónu. Z tohto dôvodu

stĺpec “timestamp” neobsahuje údaj o časovej zóne, v ktorej bola hodnota nameraná. Dáta, ktoré som používal pri návrhu aplikácie preto museli byť orezané tak, aby všetky dni mali rovnaký počet meraní alebo čas musel byť prekonvertovaný do takej časovej zóny, ktorá by nespôsobovala rôznu veľkosť periódy. Keďže množina dát, s ktorými som pracoval bola dostatočne veľká zvolil som prvú možnosť.

Dáta by samozrejme nemali byť náhodné, ale malo by sa jednať o časový rad, ktorý je možné dekomponovať na cyklickú, trendovú a náhodnú zložku. Niektoré predikčné algoritmy priamo používajú dekompozíciu časových radov na predpovedanie hodnôt. Jednotlivé záznamy o meraniach by mali za sebou chronologicky nasledovať. V rámci jedného vstupného súboru musia mať dáta všetky záznamy rovnakú periódu časového radu. Ak používateľ vo webovom rozhraní zvolí, že vložené dáta majú periódu napr. 96 meraní za deň, potom aj aplikácia bude očakávať zvolenú veľkosť periódy. Ak pri overovaní dojde k zlyhaniu aplikácie ohlási chybu.

4.2 Použité predikčné metódy

Z pohľadu návrhu aplikácie nás budú pre každú predikčnú metódu zaujímať najmä vstupné parametre, ktoré budeme optimalizovať a transformácie, ktoré je potrebné nad vstupnými dátami vykonať. Takéto delenie je dôležité aj kvôli organizácii programu pri implementácii, keďže úprava dát sa vykoná iba raz, zatiaľ čo vstupné parametre sa budú meniť neustále. Výstupom výpočtu predpovede bude samotná predpoveď alebo veľkosť chyby dosiahnutá pri nej dosiahnutá.

4.2.1 Regresia založená na podporných vektoroch

Parametre, ktoré budeme optimalizovať budú C a ε , ktoré môžu nadobúdať iba kladné hodnoty. Hodnota parametru C reprezentuje cenu za prekročenie hranice, ktorú sme definovali. Hranicu si označíme ako ε . To znamená, že pri veľmi nízkej hodnote sa výsledná predikcia podobá na konštantnú funkciu. Pri veľmi vysokej hodnote môže dôjsť k pretrénovaniu modelu a zvýšeniu výslednej chyby na množine testovacích dát. Hodnota parametru ε je hodnota použitá v nesenzitívnej stratovej funkcii, ktorá sa používa pre regresiu. Zvyšovaním tejto hodnoty sa funkcia reprezentujúca predpoveď postupne vyhladzuje, až sa bude podobáť na konštantnú funkciu.

Vstupné dáta je potrebné transformovať na 2 matice, jednu trénovaciu maticu a jednu testovaciu maticu. Trénovacia matica obsahuje stĺpec s hodnotami, ktoré boli v daných časoch namerané, testovacia má tento stĺpec prázdny, predpovedané hodnoty doplní algoritmus. Počet ďalších stĺpcov sa môže líšiť v závislosti od veľkosti zvolenej periódy. Ak si označíme počet meraní za deň ako n , potom stĺpce C_1 až C_n budú obsahovať hodnotu 1 práva vtedy, keď poradie merania v dni bude totožné s poradím stĺpca. Rovnakým postupom pridáme ešte ďalších 7 stĺpcov, ktoré budú predstavovať poradie dňa v týždni. Pre iné stĺpce C_1 až C_7 bude potom platiť, že napr. v prvom stĺpci bude priradená hodnota 1 tým záznamom, ktoré boli namerané v pondelok. Zvyšok matíc bude doplnený nulami.

4.2.2 Autoregresívny integrovaný model kĺzavého priemeru

Parametre stochastického modelu ARIMA sú v literatúre označované ako p , d a q . Z kapitoly 2.2.2 vieme, že model ARIMA je generalizáciou modelu ARMA, ktorý vznikol kombináciou autoregresívneho modelu a modelu kĺzavých priemerov. Vstupné parametre sú ne-

záporné celé čísla, kde p a q predstavuje rad modelu a d zodpovedá stupňu derivovania. Pri veľkom počte derivovaní sa funkcia predpovednej hodnoty podobá lineárnej funkcii.

Príprava vstupných dát je veľmi jednoduchá. Keďže model pracuje priamo s časovými radmi a vstupné dáta sú časový rad, jedinou úpravou je rozdelenie dát na tréningovú a testovaciu množinu a určenie veľkosti periódy, ktorú zadáva používateľ.

4.2.3 Náhodné lesy

Vstupné parametre, ktoré sa budeme snažiť v aplikácii optimalizovať bude počet stromov a minimálna veľkosť uzlu v stromoch. Je zrejmé, že číslo definujúce počet stromov bude musieť byť kladné a celé, zatiaľ čo veľkosť uzlu môže byť definovaná ako ľubovoľné nezáporné číslo. Pri zvyšovaní počtu stromov, ktoré budú vytvorené sa funkcia predpovedanej hodnoty javí, akoby každé pridanie stromu poskytovalo funkcii ďalšiu možnosť zmeny smeru. Funkcia tak pri malom počte stromov vyzerá kostrbato a náhodne, zväčšovaním počtu sa extrémny vyhladzujú a predpovedané hodnoty postupne konvergujú do funkcie, ktorá je obrysmi podobná funkcii reálnych dát. Zväčšovaním uzlov sa funkcia predikovaných hodnôt postupne vyhladzuje, až nakoniec opisuje konštantnú funkciu.

Rovnako ako model ARIMA tak aj náhodné lesy pracujú priamo s časovými radmi. Je však potrebné pridať k dátam informáciu o tom, kedy boli namerané. Zároveň by táto funkcia mala byť spojitá. To najjednoduchšie dosiahneme tak, že z dátumu extrahujeme poradie dňa v týždni a z časovej zložky zasa poradie merania v dni. Funkcie zatiaľ nie sú spojité, a preto na ne aplikujeme transformáciu popísanú vo vzorci 35 [15]

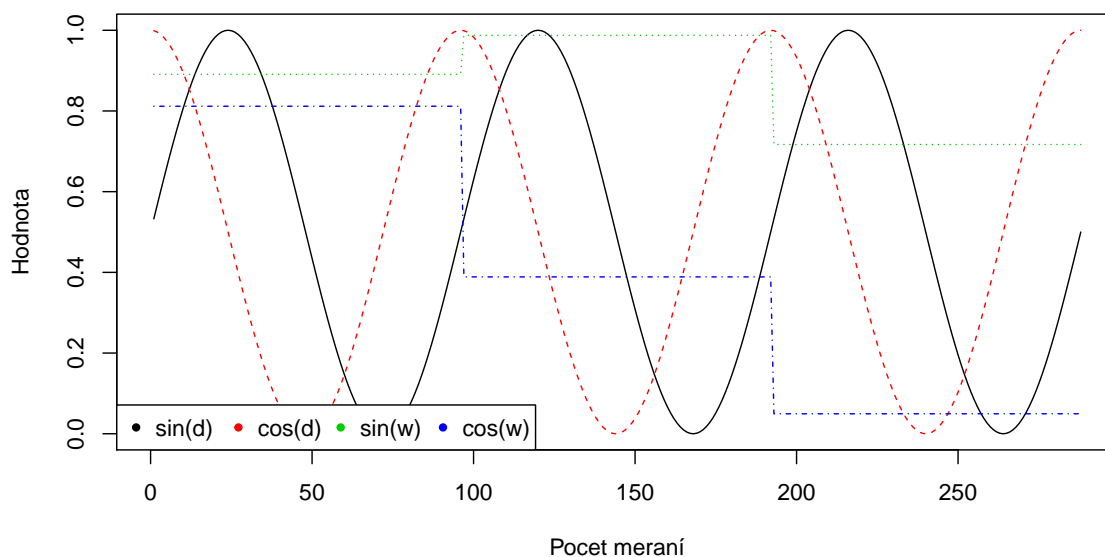
$$\frac{\sin(2\pi \frac{\text{poradie merania}}{\text{veľkosť periódy}}) + 1}{2} \text{ resp. } \frac{\cos(2\pi \frac{\text{poradie merania}}{\text{veľkosť periódy}}) + 1}{2} \quad (35)$$

a goniometrické funkcie sínus a kosínus. Výsledné funkcie sú spojité a svojou harmonickosťou opisujú vlastnosti dát skryté v dátume a čase. Pribeh 4 pomocných funkcií je znázornený na obrázku 10.

4.3 Použité optimalizačné metódy

Aplikácia bude používateľovi poskytovať rozhranie na nastavenie vstupných parametrov predikčných metód a taktiež nastavovanie rozsahu hodnôt, ktoré majú tieto metódy použiť pri optimalizovaní. Výsledkom výpočtu bude porovnanie reálnych a predpovedaných dát v podobe grafu a veľkosti výslednej chyby. Taktiež budú používateľovi zobrazené optimálne parametre pre zvolenú predikčnú metódu. Každá kombinácia parametrov vystupuje v optimalizačnom algoritme ako agent, ktorého úlohou je dosiahnuť, čo najnižšiu hodnotu fitness funkcie. V závislosti od použitého algoritmu a získaných vedomostí sa v nasledujúcich cykloch menia hodnoty reprezentujúce agenta. Hodnoty parametrov môžu byť ovplyvňované aj voľbou funkcie na meranie chyby, ktorá plní úlohu fitness funkcie pre optimalizačné algoritmy.

Pri návrhu aplikácie je dôležité najmä správne definovať rozhranie medzi predikčnými a optimalizačnými metódami. Je nemysliteľné, aby pri pridávaní ďalších algoritmov bolo potrebné upravovať už používané algoritmy, najmä kvôli množstvu kombinácií, ktoré môže vzniknúť. Našťastie každého agenta definujeme práve hodnotami, ktoré vstupujú do predikčného algoritmu. Na výstupe očakáva optimalizačná metóda hodnotu fitness funkcie, ktorá určuje ďalšie smerovanie agentov. Tento proces sa opakuje až pokiaľ nebude dosiahnutá ukončovacia podmienka. V použitých algoritmoch to je maximálny počet iterácií algoritmu. Túto hodnotu môže používateľ nastaviť v aplikácii v definovanom rozsahu.



Obr. 10: Ukážka transformácií dátumu a času pri predikcii náhodnými lesmi.

4.3.1 Optimalizácia rojom častíc

Používateľ bude mať možnosť v aplikácii zvoliť počet častíc, ktoré budú použité na optimalizovanie problému. Pod každou časticou si môžeme predstaviť agenta, ktorý bude nositeľom informácie o okolitom prostredí a o hodnotách, ktoré ho viedli k aktuálnemu riešeniu. Pre každý optimalizovaný parameter je potrebné definovať rozsah hodnôt, ktoré môže používateľ v aplikácii nastaviť. Je dôležité poznamenať, že okrem hraníc rozsahu nás zaujímajú aj samotné hodnoty, ktoré bude algoritmus používať na optimalizáciu. Je zrejmé, že používať desatinné čísla pre parameter označujúci napr. počet stromov nie je logické a aplikácia by mala byť schopná sa s tým vysporiadať.

V prostredí jazyka R sú k dispozícii 2 rôzne knižnice implementujúce optimalizáciu rojom častíc. V aplikácii sa budú pre porovnanie použité obidve implementácie.

4.3.2 Umelá kolónia včiel

Rovnako ako pri optimalizácii rojom častíc tak aj pri tomto algoritme bude mať používateľ možnosť zvoliť rozsah hodnôt, ktoré budú použité na optimalizovanie. Opäť bude dôležité vysporiadať sa s desatinými číslami. Používateľ bude môcť navyše nastavovať parametre ABC algoritmu ako je počet zdrojov potravy alebo limit pre daný zdroj. Pod limitom môžeme rozumieť počet iterácií cyklu, počas ktorého sa daný zdroj potravy nezlepšil. Týmto parametrom môžeme výrazne urýchliť nájdenie optimálneho riešenia, na druhej strane môže nastať situácia, kedy by sme rýchlym opustením zdroja potravy mohli stratiť lepšie riešenie. Zvyšovaním počtu zdrojov potravy sa zvyšuje aj čas výpočtu, ale zároveň bude nájdené optimálnejšie riešenie.

5 Implementácia

Aplikácia je kvôli prehľadnosti organizovaná do viacerých adresárov. Pri vytváraní webovej aplikácie pomocou knižnice Shiny, sa program rozdeľuje na časť webového rozhrania a serverovú časť. Tie sa pri jednoduchých aplikáciách môžu nachádzať v jednom súbore, ja som ich podľa dokumentácie rozdelil do súborov “ui.R” resp. “server.R” pre webové rozhranie resp. serverovú logiku. V adresári “src” sa nachádzajú zdrojové súbory, ktoré v sebe zahŕňajú všetky výpočty aplikácie. Od prípravy vstupných dát, cez predikčné a optimalizačné algoritmy až po fitness funkcie. Adresár “conf” obsahuje konfiguračné súbory, ktoré poskytujú celej aplikácii flexibilitu pri jej ďalšom rozširovaní a udržiavaní.

6 Zhodnotenie

7 Záver

8 Technická dokumentácia

Literatúra

- [1] Adhikari, R.: *An Introductory Study on Time Series Modeling and Forecasting*. Saarbrücken: LAP LAMBERT Academic Publishing, 2013, ISBN 9783659335082.
- [2] Arun, K.; Rejimoan, R.: A survey on network path identification using bio inspired algorithms. In *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Feb 2016, s. 387–389, doi:10.1109/AEEICB.2016.7538314.
- [3] Bel, L.; Allard, D.; Laurent, J. M.; aj.: CART Algorithm for Spatial Data: Application to Environmental and Ecological Data. *Comput. Stat. Data Anal.*, ročník 53, č. 8, Jún 2009: s. 3082–3093, ISSN 0167-9473, doi:10.1016/j.csda.2008.09.012.
URL <http://dx.doi.org/10.1016/j.csda.2008.09.012>
- [4] Breiman, L.: Random Forests. *Machine Learning*, ročník 45, č. 1, 2001: s. 5–32, ISSN 1573-0565, doi:10.1023/A:1010933404324.
URL <http://dx.doi.org/10.1023/A:1010933404324>
- [5] Chavan, S. D.; Kulkarni, A. V.; Khot, T.; aj.: Bio inspired algorithm for disaster management. In *2015 International Conference on Energy Systems and Applications*, Oct 2015, s. 776–781, doi:10.1109/ICESA.2015.7503455.
- [6] Deolekar, R. V.: Defining parameters for examining effectiveness of genetic algorithm for optimization problems. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, March 2016, s. 1061–1064.
- [7] Eurostat: Multiplicative decomposition. <https://ec.europa.eu/eurostat/sa-elearning/multiplicative-decomposition>, 2012, dostupné 2017-04-28.
- [8] Goel, L.; Gupta, D.; Panchal, V. K.; aj.: Taxonomy of nature inspired computational intelligence: A remote sensing perspective. In *2012 Fourth World Congress on Nature and Biologically Inspired Computing (NaBIC)*, Nov 2012, s. 200–206, doi:10.1109/NaBIC.2012.6402262.
- [9] Grmanová, G.; Laurinec, P.; Rozinajová, V.; aj.: Incremental Ensemble Learning for Electricity Load Forecasting. *Acta Polytechnica Hungarica*, ročník 13, č. 2, 2016.
- [10] Gruau, F.; Lyon I, L. C. B.; Doctorat, O. A. D. D.; aj.: Neural Network Synthesis Using Cellular Encoding And The Genetic Algorithm. 1994.
- [11] Gutiérrez, P. A.; Pérez-Ortiz, M.; Sánchez-Monedero, J.; aj.: Ordinal Regression Methods: Survey and Experimental Study. *IEEE Transactions on Knowledge and Data Engineering*, ročník 28, č. 1, Jan 2016: s. 127–146, ISSN 1041-4347, doi:10.1109/TKDE.2015.2457911.
- [12] Karaboga, D.; Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *Journal of Global Optimization*, ročník 39, č. 3, 2007: s. 459–471, ISSN 1573-2916, doi:10.1007/s10898-007-9149-x.
URL <http://dx.doi.org/10.1007/s10898-007-9149-x>

- [13] Kennedy, J.; Eberhart, R.: Particle swarm optimization. In *Neural Networks, 1995. Proceedings., IEEE International Conference on*, ročník 4, Nov 1995, s. 1942–1948 vol.4, doi:10.1109/ICNN.1995.488968.
- [14] Kumar Singh, A.; Khatoon, S.; Muazzam, M.; aj.: An Overview of Electricity Demand Forecasting Techniques. *Network and Complex Systems*, ročník 3, č. 3, 2013, ISSN 2225-0603.
URL www.iiste.org
- [15] Laurinec, P.: R <- Slovakia meetup started to build community in Bratislava. <https://petolau.github.io/First-R-Slovakia-meetup/>, 2017, dostupné 2017-04-26.
- [16] Lazinica, A.: *Particle swarm optimization*. Rijek, Croatia: InTech, 2009, ISBN 978-953-7619-48-0.
- [17] Liu, L.; Hudak, G.: *Forecasting and Time Series Analysis Using the SCA Statistical System*, ročník zv. 1. Scientific computing Associates Corporation, 1992.
URL <https://books.google.sk/books?id=8-HrAAAACAAJ>
- [18] Mahalakshmi, G.; Sridevi, S.; Rajaram, S.: A survey on forecasting of time series data. In *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, Jan 2016, s. 1–8, doi:10.1109/ICCTIDE.2016.7725358.
- [19] Mendes-Moreira, J. a.; Soares, C.; Jorge, A. M.; aj.: Ensemble Approaches for Regression: A Survey. *ACM Comput. Surv.*, ročník 45, č. 1, December 2012: s. 10:1–10:40, ISSN 0360-0300, doi:10.1145/2379776.2379786.
URL <http://doi.acm.org/10.1145/2379776.2379786>
- [20] Merz, C. J.: *Classification and Regression by Combining Models*. Dizertačná práca, University of California, 1998, aAI9821450.
- [21] Sapankevych, N. I.; Sankar, R.: Time Series Prediction Using Support Vector Machines: A Survey. *IEEE Computational Intelligence Magazine*, ročník 4, č. 2, May 2009: s. 24–38, ISSN 1556-603X, doi:10.1109/MCI.2009.932254.
- [22] Seeley, T. D.; Camazine, S.; Sneyd, J.: Collective decision-making in honey bees: how colonies choose among nectar sources. *Behavioral Ecology and Sociobiology*, ročník 28, č. 4, 1991: s. 277–290, ISSN 1432-0762, doi:10.1007/BF00175101.
URL <http://dx.doi.org/10.1007/BF00175101>
- [23] Sen, M.; Stoffa, P.: *Global Optimization Methods in Geophysical Inversion*. Advances in Exploration Geophysics, Elsevier Science, 1995, ISBN 9780080532561.
URL <https://books.google.sk/books?id=PT5MqSivREMC>
- [24] Simonova, S.; Panus, J.: Genetic algorithms for optimization of thematic regional clusters. In *EUROCON 2007 - The International Conference on Computer as a Tool*, Sept 2007, s. 2061–2066, doi:10.1109/EURCON.2007.4400359.
- [25] Smola, A. J.; Schölkopf, B.: A tutorial on support vector regression. *Statistics and Computing*, ročník 14, č. 3, 2004: s. 199–222, ISSN 1573-1375, doi:10.1023/B:STCO.0000035301.49549.88.
URL <http://dx.doi.org/10.1023/B:STCO.0000035301.49549.88>

- [26] Tso, G. K.; Yau, K. K.: Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, ročník 32, č. 9, 2007: s. 1761 – 1768, ISSN 0360-5442, doi:<http://dx.doi.org/10.1016/j.energy.2006.11.010>.
URL <http://www.sciencedirect.com/science/article/pii/S0360544206003288>