„Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ“

Názov projektu :

Medzinárodné centrum excelentnosti pre výskum inteligentných a bezpečných informačno-komunikačných technológií a systémov

ITMS : 26240120039

Názov výstupu :

Pracovný balík 7: Architektúra veľkých dát , štruktúra dát a ich analýza

Obsah

Zoznam obrázkov

Zoznam tabuliek

1 Úvod

Cieľom pracovného balíka s názvom „Architektúra veľkých dát, štruktúra dát a ich analýza“ bol základný výskum v oblasti efektívneho spracovania veľkých dát. Pri analýze a získavaní relevantných informácií z veľkých dátových súborov, sme sa zamerali na všetky stránky životného cyklu veľkých dát od ich zberu, organizovania úložísk, až po inteligentnú analýzu a získavanie relevantných informácií pre podporu rozhodovania. Pozornosť sme venovali tiež návrhu referenčnej architektúry veľkých dát. V oblasti dolovania informácií z veľkých dát sme uvažovali nielen statické, ale aj dynamické dáta. Skúmali sme vhodné algoritmy dolovania informácií zo štruktúrovaných a aj neštruktúrovaných dát, pričom sme pri zhlukovaní zohľadnili charakter skúmanej dátovej množiny a pri dynamických dátach ich variabilitu. Sledovali sme techniky objavovania vzorov, asociácií, korelácií a detekcie extrémov, kde sme vhodne rozvinuli metódy spracovania prúdových dát. V rámci riešenia výziev súvisiacich s úlohami modelovania inteligentných sietí sme sa orientovali predovšetkým na analýzu, skúmanie a porovnanie viacerých metód predikcie so zohľadnením doménovo-špecifických vlastností zvolených dátových množín. Výskum sme zamerali na rozvoj existujúcich a hľadanie nových metód predikcie, pričom sme sa zamerali hlavne na vývoj nových analytických metód a algoritmov. Ťažiskom výskumu bol návrh nových analytických nástrojov so zameraním na predikčné modelovanie v rámci inteligentných sietí. Skúmali sme možnosti nových optimalizovaných algoritmov, v rámci ktorých sme analyzovali vlastnosti inkrementálnych a adaptívnych modelov a skúmali vhodnosť ich použitia na predikovanie spotreby elektrickej energie. Novovyvinuté riešenia boli testované aj na dostupných veľkých dátových množinách z oblasti genetiky a spracovania prirodzeného jazyka.

Súčasťou pracovného balíka je aj vývoj nových analytických algoritmov pre neštruktúrované dáta a jazykové technológie v pôsobnosti mnohojazyčnej Európy. Tieto nástroje sú vhodnými prostriedkami zlepšenia analýzy a porozumenia slovenských ako aj iných cudzojazyčných textov. Balík obsahuje analýzu metód automatického prekladania, ako aj návrh nových algoritmov na spracovanie metadát o používateľoch.

Témy pracovného balíka sú rozdelené do ôsmich kapitol. Prvé dve kapitoly sa zaoberajú prístupmi k riešeniu úloh veľkých dát a prinášajú prehľad vývoja nových architektúr. Tretia kapitola bola vypracovaná na základe zmluvného výskumu s ARI. Štvrtá kapitola prináša prehľad výskumu v oblasti dolovania informácií z veľkých dát. Piata kapitola sa zameriava na

problematiku inteligentných sietí. Prehľad analytických nástrojov a algoritmov používaných pre spracovanie veľkých dát v oblasti predikcie na základe časových radov a inteligentného získavania informácií, je obsahom šiestej kapitoly. Výsledky výskumu vykonaného v oblasti predikcie spotreby elektrickej energie, ako aj návrh nových analytických algoritmov a metód je obsahom siedmej kapitoly. Posledná, ôsma kapitola, sa zameriava na tému spracovania neštruktúrovaných dát najmä prirodzeného jazyka, ako aj získavania informácií z metadát o používateľoch.

2 Veľké objemy dát

Vývoj technológií na uchovávanie a spracovávanie dát za posledných päťdesiat rokov umožnil, že dnes sú voľne dostupné prostriedky pre zbieranie a spravovanie aj veľkých objemov dát. Skutočná hodnota však nie sú samotné dáta, ale informácie ukryté v dátach. Dnes sa objemy zozbieraných dát dokážu pohybovať rádovo v terabajtoch, petabajtoch a neustále rastú. Za čas, potrebný na spracovanie takéhoto množstva dát človekom, stratia informácie ukryté v dátach hodnotu. Preto sú dáta analyzované pomocou automatizovaných postupov. Informácie získané analýzou dát môžu ponúknuť výhodu na trhu v komerčných oblastiach (napr. zníženie nákladov na skladovanie tovaru na základe analýzy nákupov zákazníkov, predpovedanie správania zákazníka, dopytu po určitých položkách, personalizovaný marketing), ale môžu byť veľkým prínosom aj pre nekomerčné odvetvia (napr. monitorovanie atmosféry, kriminality, terorizmu, organizácia zdrojov v krízových situáciách, výskum v oblasti medicíny, biológie, geografie a pod.).

Veľké údajové korpusy (angl. *big data*) sú definované pomocou tzv. 3V modelu, ktorý vznikol už v roku 2001 vo výskumnej organizácii Gartner (predtým META Group) [75]. V roku 2012 táto výskumná organizácia definovala veľké údajové korpusy ako *„informačné aktívum s veľkým objemom, veľkou rýchlosťou rastu a/alebo veľkou rôznorodosťou, ktoré vyžaduje nové spôsoby spracovávania, aby bolo možné vykonávať dokonalejšie rozhodnutia, získavať prehľad o dátach a optimalizovať procesy“* [12].

Tri písmená „V“ v modeli znamenajú:

- **objem** (angl. ***V****olume*): Veľkosť údajového korpusu sa pohybuje okolo niekoľko terabajtov až petabajtov. Korpusy s obrovským objemom predstavujú napr. transakčné dáta ukladané počas niekoľkých rokov, neštruktúrované dáta, ktoré prúdia zo sociálnych médií, dáta zo senzorov. S čoraz menšími nákladmi na skladovanie obrovského množstva dát sa vynára problém ako určiť, ktoré dáta obsahujú nejakú hodnotu a oplatí sa ich ukladať.

- **rýchlosť** (angl. ***V**elocity*): Dáta rýchlo pribúdajú a potrebujú byť rýchlo spracovávané. Spracovávať veľký prúd dát v reálnom čase je v súčasnosti výzva pre mnohé organizácie.

- **rôznorodosť** (angl. ***V**ariety*): Dáta môžu pochádzať z niekoľkých zdrojov a môžu byť štruktúrované, neštruktúrované, prípadne nemusia mať textovú podobu, napr. obrázky, video. Spájanie dát z rôznych zdrojov a z rôznych formátov do spracovateľnej podoby, z ktorej dokážeme získať hodnotné informácie je ďalšia výzva, s ktorou sa potýka mnoho organizácií.

Niektoré zdroje [14, 68, 115] dodávajú ešte ďalšie „V“, napr. vierohodnosť (angl. ***V**eracity*), realizovateľnosť (angl. ***V**iability*) a hodnota (angl. ***V**alue*). Tieto charakteristiky sa však týkajú už procesu analýzy dát a necharakterizujú dáta samotné. Poukazujú na to, že pri analýze by mali byť vybraté vhodné aspekty dát (realizovateľnosť), aby sme získali analýzou čo najrozumnejší a najprínosnejší výstup (hodnota). Keďže dáta musia byť rýchlo spracovávané, nie je dostatočný časový priestor pre „očistenie“ dát. Preto pri rozhodovaní sa treba zohľadniť tento fakt a nedôverovať výstupom na sto percent (vierohodnosť).

Práca s veľkými údajovými korpusmi začína zadefinovaním problému alebo úlohy, ktorú je možné vyriešiť pomocou týchto dát, napr. monitorovať stroje vo výrobe pri produkcii a eliminovať chybné výrobky, monitorovať dopravu na frekventovaných komunikáciách a eliminovať zápchy, nehodovosť, znečistenie, optimalizovať spotrebu paliva a pod.

Na základe zadefinovaného problému vieme ako postupovať ďalej a prispôsobiť podľa toho proces spracovávania veľkých korpusov, ktorý sa skladá z nasledujúcich fáz: (a) zber dát, (b) organizovanie, (c) integrácia, (d) analýza, (e) rozhodovanie.

Jednou z hlavných vlastností veľkých údajových korpusov podľa definície je rýchlosť, t. j. dáta rýchlo pribúdajú a je potrebné ich rýchlo spracovávať. Preto ako dáta kontinuálne pribúdajú, proces sa neustále opakuje, napr. dáta sa spracovávajú raz denne a vznikajú denné reporty, ktoré pomáhajú organizácii pri rozhodovaní (pozri obrázok 1).

Obr. 1. Životný cyklus správy veľkých údajových korpusov [60].

2.1 Zber dát

V tejto fáze sa rozhodne, aké dáta a ako je potrebné ich ukladať na základe vytýčeného cieľa alebo úlohy. Dátová architektúra sa odvíja od toho, aké veľké objemy dát je potrebné skladovať, ako rýchlo je potrebné k dátam pristupovať, či je potrebné spracovávať dáta v reálnom čase, ako presné dáta musia byť, aká je povaha dát (napr. nakoľko je potrebné ich zabezpečiť).

Dáta zvyčajne pochádzajú z viacerých zdrojov a bývajú uložené vo vhodnej databáze na základe ich povahy. Štruktúrované dáta bývajú najčastejšie uložené v relačnej databáze. Ako zdroje dát treba brať do úvahy ale aj neštruktúrované dáta, napr. dáta zo sociálnych sietí, CMS systémov a pod. Pre väčšiu efektívnosť práce s veľkými objemami dát a pre ukladanie neštruktúrovaných dát sú vhodnejšie iné typy databáz ako klasické relačné, napr. dokumentovo orientovaná, grafová, stĺpcovo orientovaná, geopriestorová databáza, spolu nazývané tzv. *NoSQL* databázy.

Po určení dát, ktoré sa majú zbierať a vytvorení vhodnej dátovej architektúry, sú dáta kontinuálne ukladané a dostupné pre ďalšie nástroje pre správu dát v pokročilejších fázach procesu.

2.2 Organizovanie

Organizovanie dát znamená narábanie s dátami a vykonávanie dopytov nad dátami. Zbierané dáta pochádzajú z viacerých zdrojov, napr. senzory, rôzne stroje, vedomostné bázy. V minulosti nebolo možné poňať také obrovské množstvá dát, t. j. organizácie buď neboli schopné zachytiť a uskladniť toto množstvo alebo neboli dostupné nástroje na narábanie s týmito obrovskými objemami. Aby organizácie získavali informácie z dát v reálnom čase, boli nútené pracovať len s malými zábermi dát (angl. *snapshot*), ktoré ale nemuseli zachytiť dôležité udalosti v dátach, a preto organizácia mohla strácať informácie. Dnes sa situácia zmenila a existuje niekoľko komerčných aj voľne dostupných nástrojov vytvorených špeciálne pre organizovanie a správu veľkých údajových korpusov, napr. MapReduce a Big Table od Google alebo Hadoop od Apache.

2.3 Integrácia

V tejto fáze sú dáta z rôznych zdrojov integrované do jedného úložiska, tzv. dátového skladu (angl. *data warehouse*), určeného pre skúmanie a analýzu dát. Pre transformáciu dát z rôznych zdrojov, ktoré vlastníme alebo externých zdrojov, napr. štruktúrované dáta z našej databázy, neštruktúrované dáta zo sociálnych sietí, vedomostné bázy, externé zdroje; sú potrebné konektory a metadáta. Konektory sú použité na prepojenie rôznych zdrojov dát. Metadáta zase opisujú rôzne zdroje dát, ich zjednotenie a transformácie.

2.4 Analýza

Vo fáze analýzy sú využívané analytické nástroje pre hľadanie vzorov, predpovedí a pod. v integrovaných dátach. Získané informácie sú nápomocné pri dosahovaní zadefinovaného cieľa alebo riešení daného problému. Analytické nástroje využívajú proces objavovania znalostí (angl. *knowledge dicovery from data*, skr. *KDD*) a metódy dolovania v dátach (angl. *data mining*), ktoré sú podrobnejšie zachytené v časti 3. Čím ďalej, tým viac je populárnejšia tzv. vizuálna analytika [129], ktorá kombinuje automatizované metódy dolovania v dátach s vizualizáciou informácií. Analýza úzko súvisí aj s prognostikou, ktorú využívajú v manažmente organizácií. Slúži na predpovedanie budúceho vývoja v rôznych odvetviach. Na základe prognóz si organizácie vytvárajú plány a vykonávajú rozhodnutia do budúcnosti. Prognostika takisto využíva metódy pre dolovanie v dátach. Aj jej sa venujeme v časti 3.

2.5 Rozhodovanie

V poslednej fáze procesu sa na základe informácií získaných analýzou vykonajú príslušné rozhodnutia a akcie, napr. odporučí sa výrobok zákazníkovi na základe analýzy jeho predchádzajúcich nákupov, odkloní sa doprava, posilnia sa hliadky na nehodových úsekoch premávky a pod.

3 Vývoj nových architektúr veľkých objemov dát

Táto kapitola bude definovať referenčnú architektúru vysokej úrovne pre zobrazovanie big data, jej hlavné stavebné bloky a funkčné komponenty. Navrhneme rámec architektúry Big Data založený na existujúcej práci na medzinárodnej úrovni, sústreďujúci sa hlavne na prácu vykonanú štandardizačnými orgánmi a pri výskumných iniciatívach financovaných zo strany EÚ. Architektúra zohľadní celý hodnotový reťazec Big Data a skombinuje technické aspekty (spracovanie v reálnom čase, dávkové spracovanie, atď.) ako aj aspekty riadenia dát (orchestrácia, poskytovatelia IT, atď.).

Ako spôsob určenia referenčnej architektúry bude poskytnutých niekoľko doložení príkladov architektúry pre scenáre analytiky sociálnych sietí a SmartGrid. Tieto doloženia príkladmi sú zahrnuté z dôvodu jasnosti prístupu pri zdôvodňovaní referenčnej architektúry pre špecifické scenáre big data, v žiadnom prípade však nie sú jedinými doloženiami príkladov pre tieto dva scenáre.

3.1 Motivácia a príklady

3.1.1 Motivácia

V posledných rokoch bol v mnohých scenároch preukázaný potenciál Big Data pre transformáciu firiem. V súčasnosti už nikto nepochybuje, že takzvaná ekonómia dát sa stala trvalým javom. Príchod Big Data však nie je bez nedostatkov. Množstvo technologických možností sťažuje nezasväteným rozhodnutie o technickom rámci na podporu inštalácie Big

Data. Prostredie je plné vlastníckymi právami chránených a otvorených riešení, ktoré navzájom súťažia o kus trhu, vyhlasujúc, že prekonávajú konkurentov, a to často nepatrným spôsobom. Preto sa stalo potrebným abstrahovať od tejto urputnej konkurencie a odstúpiť o krok späť, aby sa definovali hlavné stavebné bloky referenčnej architektúry big data. Potom sa dá pri skúmaní špecifických scenárov a firemných potrieb z referencie odvodiť určitá konkrétna systémová architektúra pochopením toho, ktoré nástroje týmto potrebám lepšie vyhovujú.

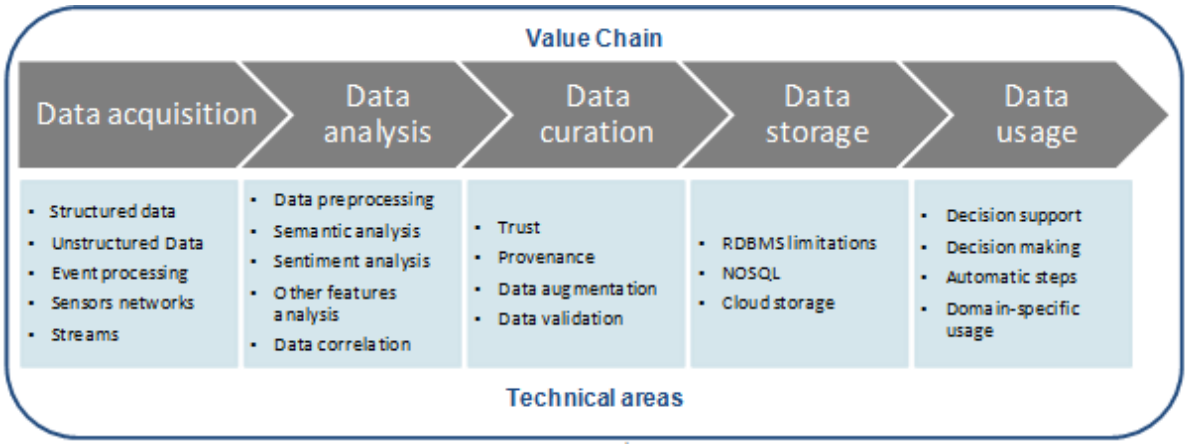
3.1.2 Hodnotový reťazec dát

Koncepcia hodnotového reťazca nie je nová. Bola zavedená Porterom [107] pochádzajúc z oblasti riadenia firiem a odvtedy sa obšírne používa ako nástroj na podporu rozhodovania pri modelovaní reťazca aktivít, ktoré organizácia vykonáva s cieľom poskytovania určitého produktu alebo služby. Hodnotový reťazec kategorizuje generické aktivity pridávania hodnoty organizáciou, umožňujúc ich pochopenie a optimalizáciu. Hodnotový reťazec sa skladá z rady podsystémov, pričom každý z nich má svoje vstupy, procesy transformácie a výstupy. Ako analytický nástroj sa dá hodnotový reťazec aplikovať na informačné systémy pre pochopenie tvorby hodnoty dátových technológií.

Hodnotový reťazec dát bol v širokej miere prevzatý na popisovanie rôznych aktivít pridávania hodnoty vykonávaných vo väčšine systémov Big Data. Preto je dobrý abstraktný model na pochopenie súboru úloh, ktoré sú obyčajne uskutočňované počas životného cyklu dát. Životný cyklus dát sa obyčajne začína **objavením** dát, kde je potrebné lokalizovať dátové zdroje, ktoré môžu pomôcť pri riešení firemných alebo výskumných problémov a vyhodnotiť pre ich použitie. Tento proces objavovania často nie je do hodnotového reťazca dát zahrnutý, pretože obyčajne k nemu dochádza mimo systému, avšak je vhodné s ním uvažovať z perspektívy životného cyklu, pretože predstavuje jednu z podstatných súčastí práce, ktorú je s dátami potrebné vykonať. Po objavení by mali dáta prejsť procesom **akvizície,** v ktorom sú dáta zo svojich pôvodných zdrojov prenesené do systému, predbežne spracované, premenené na model a formát v súlade so systémovými požiadavkami a nakoniec sú integrované. Niektorí autori považujú predbežné spracovanie v hodnotovom reťazci za odlišný krok od akvizície, zatiaľ čo iní uprednostňujú jeho presunutie do nasledovného kroku: **analýzy**. Analýza dát je hlavným bodom celého príbehu: získanie podrobného pohľadu z dát, potrebného na odvodenie podnikateľských rozhodnutí, alebo výskumných objavov. Táto fáza si obyčajne vyžaduje špecifickú funkciu –dátového vedca- ktorá je schopná navrhnúť dômyselné algoritmy s použitím škály nástrojov, ktoré dávajú zmysel pre dostupné dáta. Analýza veľmi závisí na každom z ostatných krokov v hodnotovom reťazci dát. Ďalším krokom reťazca je **ukladanie** dát. Uloženie a zachovanie dát predstavuje celý nový svet v oblasti big data, pretože obrovské objemy a vysoká rýchlosť prijímania obyčajne nie sú tradičnými systémami RDBMS dobre pokrývané. Preto boli definované špecifické modely zachovania dát a boli sprístupnené širokej škále systémov ukladania. Záverečným krokom je **používanie** dát. Výsledky analytiky a dát by mali byť sprístupnené takým spôsobom, ktorý ich robí užitočnými pre tvorbu hodnôt. Zahŕňa to rôzne techniky, so

zvláštnym dôrazom na vizualizáciu, avšak nezabúdajúc na služby umožňujúce strojné využívanie dát a detailných pohľadov, ani na rozhrania s podporou rozhodovania a iné zapojené systémy.

Nižšie môžeme vidieť prehľad typického hodnotového reťazca dát:



Obr. 2. Hlavné kroky hodnotového reťazca dát[1].

Stojí za to spomenúť, že hodnotový reťazec dát nie je funkčným sledom krokov, ale logickým reťazcom činností, hoci jeho predstavenie môže navodzovať niečo iné. Napríklad uloženie dát sa môže uskutočniť kedykoľvek v procese, alebo vôbec nie, pretože existujú výkonné technológie typu in-memory, ktoré si vôbec nemusia vyžadovať ukladanie, hoci vo väčšine prípadov sa surové dáta alebo spracované dáta uchovávajú pre potenciálnu ďalšiu analýzu. Aj usporadúvanie dát sa uskutočňuje v ľubovoľnom danom momente od skutočného objavenia dát, pričom ovplyvňuje všetky krok od akvizície až po použitie.

Podobne ako v hodnotovom reťazci dát, niektorí autori definujú nový hodnotový reťazec pre big data: **Hodnotový reťazec IT**. Tento hodnotový reťazec IT sa z pohľadu architektúry považuje za ortogonálny hodnotový reťazec. Hodnotový reťazec IT pokrýva technologické aspekty, ktoré sa dajú znovu použiť v mnohých krokoch hodnotového reťazca dát a sú určitým spôsobom technologickými základmi architektúry. Neexistuje žiadna požiadavka pre všetky prípady referenčnej architektúry, aby boli založené na tej istej technologickej štruktúre, alebo aby boli zabezpečené jediným predávajúcim. V skutočnosti väčšina implementácií spája rôzne technologické prístupy, čím sa poskytuje architektúre väčšia pružnosť, avšak za cenu rozšírenia technologického záberu, čo komplikuje celkový obraz.

Hlavné vrstvy hodnotového reťazca IT súvisia s **infraštruktúrou** (fyzickou a virtualizovanou), **uložením** a **výpočtovými rámcami** (pre dávkové a streamové spracovanie, strojné výučbové nástroje, atď.). Riadenie a použitie týchto IT zdrojov v hodnotovom reťazci dát umožňuje referenčnú architektúru, ktorá sa dá zavádzať s použitím hardvéru a softvéru a rôznym pôvodom.

[1] Zdroj: Projekt BIG (www.big-project.eu/)

3.1.3 Príklady existujúcich architektúr

Definovali sme hodnotový reťazec IT na popísanie toho, že nie je žiadna požiadavka toho, aby mali všetky prípady architektúry tú istú technológiu, alebo aby pochádzali od toho istého predávajúceho. Táto časť presne popisuje niektoré príklady architektúry, ktoré pochádzajú od konkrétnych predávajúcich, aby sa ukázalo ako riešia hodnotový reťazec dát z rôznych perspektív. Nepredstavuje to vyčerpávajúci zoznam predávajúcich, ale zvolený výber niektorých z nich, ktorý dostatočne poukazuje na rôzne architektonické prístupy, ktoré odvetvie poskytovateľov big data využíva.

3.1.3.1 IBM

Dolu je ukázaná referenčná architektúra Big Data od IBM[2]:

Obr. 3. Big Data architektúra IBM.

Prístup IBM k architektúre sa skladá z nasledovných stavebných blokov:

1. Vizualizácia a objavenie: IBM uprednostňuje začať od posledného kroku hodnotového reťazca dát zabezpečením stavebného bloku, ktorý poskytuje zjednotený mechanizmus vyhľadávania, navigácie a vizualizácie pre prístup k dátam v ich systéme.
2. Systém Hadoop: Toto je jadro spracovania dát pre uložené dáta. IBM navrhuje konkrétnu implementáciu založenú na Hadoop.

[2] So zvolením IBM: http://www.ibmbigdatahub.com/blog/part-iv-flexible-platform-based-approach-big-data

3. Streamová výpočtová činnosť: Rovnako ako v predchádzajúcom kroku, avšak pre časť streamového spracovania, IBM ponúka rámec na uskutočňovanie spracovania dát v reálnom čase.

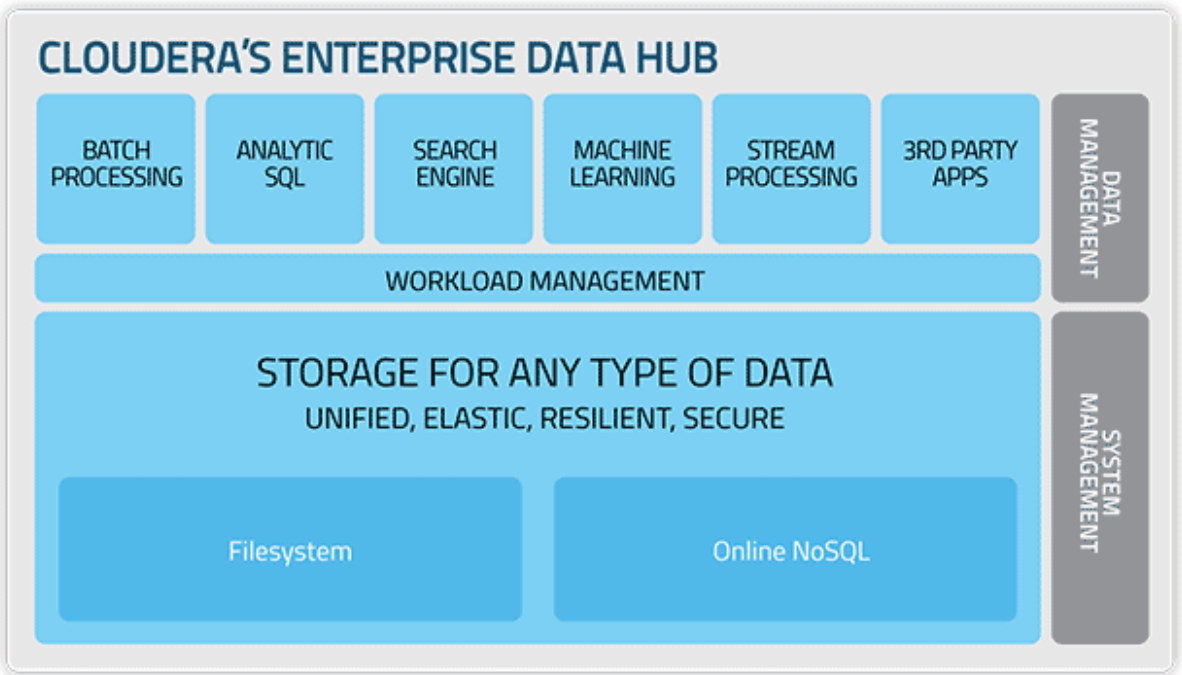
4. Nástroj Data Warehouse (úložiska dát): Pre krok ukladania dát IBM navrhuje špecifický komponent pre komplexné analytické pracovné záťaže na uložených dátach.

5. Analytické akcelerátory: Pre krok analýzy dát IBM poskytuje súbor zabudovaných komponentov vo forme matematických, štatistických a strojných výučbových algoritmov na urýchlenie procesu tvorby analytických aplikácií. Tvrdia, že poskytujú komponenty na spracovanie textu, obrázkov, videa, akustiky, časových radov, geopriestorových a sociálnych dát.

6. Integrácia a riadenie informácií: IBM pokrýva kroky akvizície dát poskytnutím konektorov pre prijímanie dát z mnohých typov bežných formátov, databáz a streamov. Poskytuje tiež nástroje predbežného spracovania, ktoré presúvajú prijaté dáta do systému analytiky a ukladania. V rámci tohto bloku IBM rieši niektoré problémy súvisiace s usporiadaním a bezpečnosťou dát.

3.1.3.2 Cloudera

Cloudera je jedným z najznámejších predajcov Big Data, ponúkajúcim produkty a služby v tejto oblasti. Verzia Enterprise hlavného produktu Big Data má nasledovnú architektúru vysokej úrovne:



Obr. 4. Big Data architektúra Cloudera.

Vo vyššie uvedenej schéme chýbajú niektoré kroky ako je akvizícia dát alebo usporiadanie dát. Cloudera sa sústreďuje špeciálne na spracovanie, analýzu, uloženie a používanie dát. Prístup Cloudera k architektúre pokrýva nasledovné stavebné bloky:

1. Uloženie ľubovoľného typu dát: Cloudera ponúka duálnu vrstvu ukladania založenú na systéme súborov (Hadoop File System alebo HDFS) a databázu NoSQL (HBase).

2. Riadenie pracovnej záťaže: Cloudera ponúka nástroje na zaistenie toho, že systém bude správne škálovaný v závislosti na pracovnej záťaži.

3. Spracovanie po dávkach: Pre analytické kroky je rámec dávkového spracovania Cloudera založený na Hadoop.

4. Streamové spracovanie: Pre analytické kroky v reálnom čase, streamové spracovanie Cloudera využíva Apache Spark a Apache Storm.

5. Analytický SQL: firma Cloudera navrhla nástroj (Cloudera Impala[3]) implementáciu otvoreného zdroja stroja podobného ako SQL na uskutočňovanie analytických dopytovaní na dátach uložených na vrstve úložiska. Je to súčasť kroku využívania dát v hodnotovom reťazci dát.

6. Vyhľadávací stroj: Cloudera tiež vykonáva smart indexing obsahu uloženého v Hadoop, čím ho robí dostupným pre typ vyhľadávania podobný Googlu. Je to súčasť kroku využívania dát v hodnotovom reťazci dát.

7. Strojná výučba: Cloudera ponúka knižnice strojnej výučby založené na Apache Spark.

8. Aplikácie 3.strán: Cloudera ponúka podporu pri jednoduchom doplnení aplikácií tretích strán do svojho súboru.

9. Nástroje pre riadenie dát a systému: Cloudera ponúka nástroje na uľahčenie procesu riadenia viacerých dátových procesov (objavenie, navigáciu a usporiadanie) a monitorovania systému.

Cloudera poskytuje súbor referenčných architektúr pre rôzne umiestnenia u známych poskytovateľov cloudov ich produktov[4].

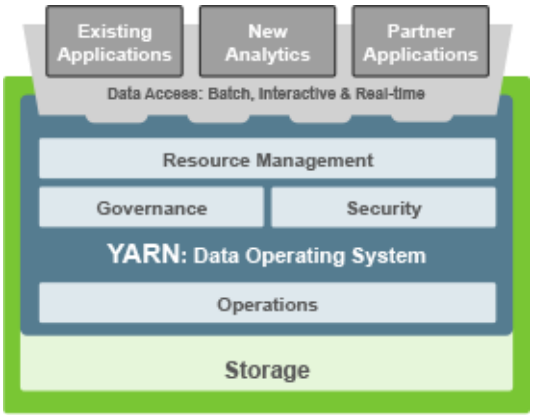
3.1.3.3 Hortonworks

Dátová platforma Hortonworks (Hortonworks data platform -HDP) je otvorená architektúra, ktorá ponúka to najnovšie v komunitou riadenej inovácii Hadoop. Hadoop Data Platform je dátová platforma pre spracovanie dát s rôznou pracovnou záťažou v celej škále metód

[3] https://github.com/cloudera/impala
[4] http://www.cloudera.com/content/cloudera/en/documentation/reference-architecture/latest.html

spracovania – od dávkového, cez interaktívne, až po spracovanie v reálnom čase – pričom jadrom platformy je YARN[5].

Typická architektúra s najreprezentatívnejšími vrstvami zahŕňa nasledovné:



Obr. 5. Big Data architektúra Hortonworks.

Hlavné črty Open Enterprise Hadoop sa opierajú o komponenty otvorených zdrojov a o otvorenú komunitu. Ako výsledok tejto stratégie poskytujú riešenia Open Enterprise Hadoop nasledovné:

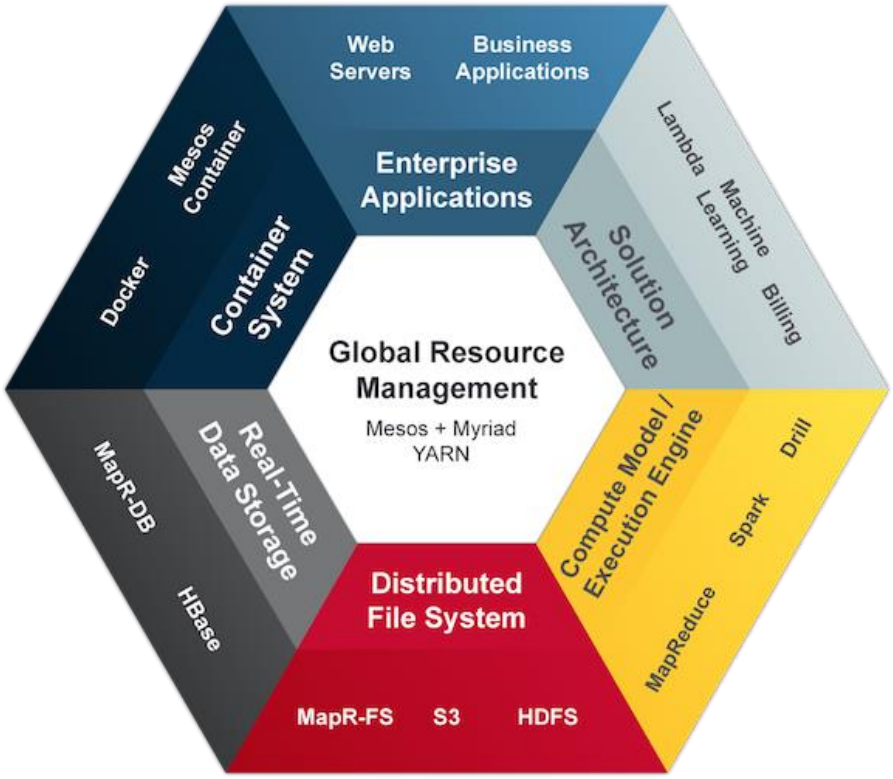
- **Účinne kombinujú silu vývoja otvorených zdrojov s poslednými inováciami v komunite**
- Open Enterprise Hadoop tiež zabezpečuje plnú interoperabilitu nad rámec jadra Hadoop prostredníctvom podpory otvorených štandardov pre široký ekosystém technológie.
- **Poskytujú robustné operácie, bezpečnosť a schopnosti riadenia s platformou, ktorá je už na úrovni podniku**

3.1.3.4 MapR

MapR Distribution zahŕňajúca Apache Hadoop, poskytuje dátovú platformu pre spoľahlivé ukladanie a spracovanie rozsiahlych a rýchlych dát. MapR Distribution, založená na technológii Hadoop, vám poskytuje ten najlepší základ pre prácu dávkových, interaktívnych aplikácií a aplikácií reálneho času. So svojím prístupom otvoreného výberu otvoreného zdroja vám MapR dáva k dispozícii široký rozsah technológií, vrátane množstva projektov pre SQL-on-Hadoop, databázy NoSQL, vykonávacie stroje ako je Spark, atď.

Zeta Architecture[6] je firemná architektúrna konštrukcia podobajúca sa na architektúru Lambda, ktorá umožňuje zjednodušené procesy firemnej činnosti a definuje škálovateľný spôsob na zvýšenie rýchlosti integrácie dát do firemnej činnosti:

[5] http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html



Obr. 6. Big Data architektúra MapR.

Existuje sedem pripojiteľných komponentov Zeta Architecture, ktoré navzájom spolupracujú:

- Distribuovaný systém súborov – čítanie a písanie všetkých aplikácií v bežnom, škálovateľnom riešení, čo zjednodušuje systémovú architektúru.
- Ukladanie dát v reálnom čase – podporuje potrebu business aplikácií s vysokou rýchlosťou prostredníctvom používania databáz v reálnom čase.
- Vykonávajúca jednotka –poskytuje rôzne jednotky spracovania a modely, s cieľom splnenia potrieb rôznych business aplikácií a užívateľov v určitej organizácii.
- Nasadenie / Kontajnerový systém riadenia – poskytuje štandardizovaný prístup pre nasadenie softvéru. Všetci spotrebitelia zdrojov sú izolovaní a nasadení štandardným spôsobom.

[6] https://www.mapr.com/solutions/zeta-enterprise-architecture

- Architektúra riešenia –sústreďuje sa na riešenie konkrétnych business problémov a spája jednu alebo viac aplikácií vybudovaných na poskytnutie kompletného riešenia. Tieto architektúry riešenia zahŕňajú interakciu vyššej úrovne medzi spoločnými algoritmami alebo knižnicami, softvérovými komponentmi a firemnými postupmi prác.
- Podnikové aplikácie – prináša jednoduchosť a opakovanú použiteľnosť poskytovaním komponentov potrebných na pochopenie všetkých business cieľov definovaných pre určitú aplikáciu.
- Dynamické a globálne riadenie zdrojov – umožňuje dynamickú alokáciu zdrojov, takže môžete vyhovieť požiadavkám ľubovoľnej úlohy, ktorá je v ten deň najdôležitejšia.

3.2 Snaha o štandardizáciu a projekty Big Data

3.2.1 Snahy EÚ

3.2.1.1 Mechanizmy financovania v Európe

Európska komisia (EK) má dlhú históriu záznamov o financovaní projektov súvisiacich s big data, už od doby keď sa big data takto ešte ani nenazývali. V posledných výzvach Siedmeho rámcového programu (**FP7**)[7] sa medzi inými nachádzali aj vyhradené výzvy ICT týkajúce sa Big Data (Výzva 4 – Ciele 4.1 a 4.2)[8]. Iné výzvy ICT financovali Big Data pre stredné a malé podniky a financovanie bolo udelené dokonca aj vo výzvach mimo ICT, ktoré trochu zaváňali témou big data. Výsledkom je, že celá skupina RTD projektov, z ktorých sa mnohé ešte realizujú, riešila rôzne aspekty Big Data.

K tomuto úsiliu prispeli aj iné programy financovania, ako je **EIT ICT Labs**[9]. Príkladom je vývoj Apache Flint, ktorý využíval financovanie z rôznych agentúr EÚ a národných agentúr a z Laboratórií ICT.

Od roku 2014 sa väčšina financovania z EÚ realizuje cez program Horizon 2020 (**H2020**)[10]. Tento nový pracovný program, nasledovník FP7, sa silno sústreďuje na aplikáciu technológie. V oblasti big data to znamená, že EK očakáva návrhy projektov na zvládnutie problémov dát v reálnom čase, s realistickými a obrovskými datasetmi a sústreďujúce sa na zvýšenie úrovne technológie. Doteraz boli vydané dve výzvy ICT na inovácie big data (ICT16-2014) a

[7] FP7: http://cordis.europa.eu/fp7/ict/home_en.html
[8] 2013 ICT Workprogramme: http://cordis.europa.eu/fp7/ict/docs/ict-wp2013-10-7-2013-with-cover-issn.pdf
[9] EIT ICT Labs project Stratosphere (Apache Flink): http://stratosphere.eu/disclaimer.html
[10] H2020: http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/index.html

výskum(ICT15-2015). Pozrite tiež pracovný program H2020 ICT, kde sú uvedené ďalšie podrobnosti[11].

Nedávno sa v rámci spoločného úsilia spojila EK a viaceré zainteresované subjekty z európskeho priemyslu spojené pod dáždnikom **Big Data Value Association** –BDVA- (medzi nimi aj ATOS)[12], do **Partnerstva verejného a súkromného sektora (Public-Private Partnership**-PPP)[13] s cieľom spolupráce vo výskume a inováciách súvisiacich s dátami, podpory vytvorenia komunity v oblasti dát a na položenie základov prekvitajúcej ekonomiky v Európe, hnanej vpred s pomocou dát. V nasledujúcich rokoch bude toto partnerstvo poskytovať väčšinu financovania Big Data v Európe. PPP sa snaží o posilnenie hodnotového reťazca dát, so zámerom toho, aby sa Európe umožnilo zohrávať relevantnú rolu v oblasti Big Data na globálnom trhu. BDVA vydalo Agendu strategického výskumu a inovácií (Strategic Research and Innovation Agenda-SRIA)[14] k Big Data ako kľúčový vstup do Pracovného programu PPP, ktorý sa má spustiť v poslednom štvrťroku 2015.

3.2.1.2 Vzorové projekty

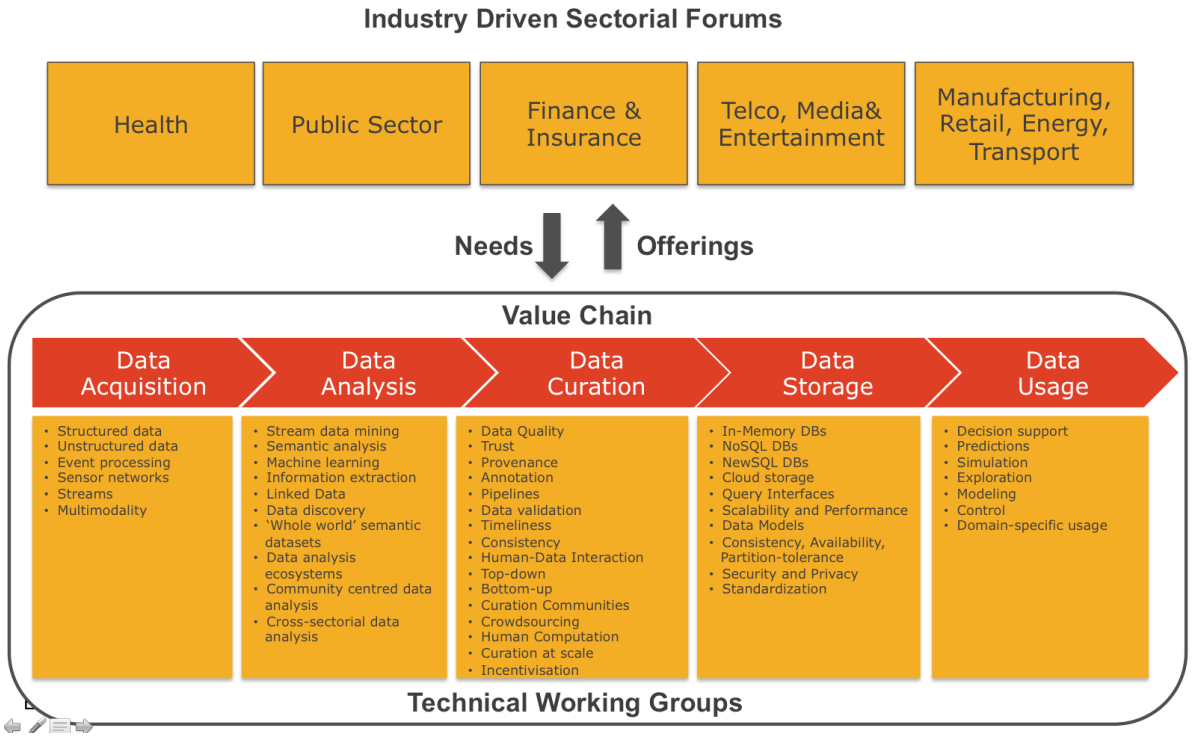
V tejto časti bude predstavených niekoľko projektov financovaných EÚ, zaujímavých z pohľadu architektúry, súvisiacich s big data.

Projekt BIG

BIG[15] je opatrením koordinácie a podpory FP7, ktorého hlavným cieľom bolo poskytnúť „cestovný poriadok“ pre Big Data v rámci Európy. Program BIG prebiehal do konca roka 2014 a poskytol vstup pre EK smerom k definícii PPP venovaného big data.

BIG bol štruktúrovaný do viacerých priemyslom riadených odvetvových fór na zabezpečenie podnikateľského pohľadu na Big Data z rôznych oblastí. Súčasne sa niekoľko Technických pracovných skupín sústredilo na technológie Big Data pre každý krok v hodnotovom reťazci dát, aby sa preskúmali schopnosti a aktuálna zrelosť technológií v týchto oblastiach.

[11] Pracovný program ICT H2020 ICT: https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-leit-ict_en.pdf
[12] BDVA: http://bdva.eu/
[13] vPPP: http://ec.europa.eu/digital-agenda/en/big-data-value-public-private-partnership
[14] http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?action=display&doc_id=7151
[15] www.big-project.eu/



Obr. 7. Hodnotový reťazec dát a priemyslom riadené odvetvové fóra sledované projektom BIG.

BIG má preto zvláštnu dôležitosť pre pochopenie úrovne zrelosti technológií súvisiacich s hodnotovým reťazcom dát. „Biela kniha technických pracovných skupín Big Data“ [17] poskytuje prehľad postupov a nástrojov, ktoré sa dajú používať v rôznych krokoch hodnotového reťazca dát.

BIG však neposkytuje referenčnú architektúru pre Big Data.

Projekt MLi

MLi je projektom Podporného opatrenia zameraným na poskytnutie strategickej vízie a operačných špecifikácií potrebných pre vybudovanie komplexnej Európskej viacjazyčnej infraštruktúry dát a služieb, súbežne s viacročným plánom pre jej vývoj a nasadenie a na upevnenie aliancií viacerých zainteresovaných subjektov, zabezpečiac jej dlhodobú udržateľnosť.

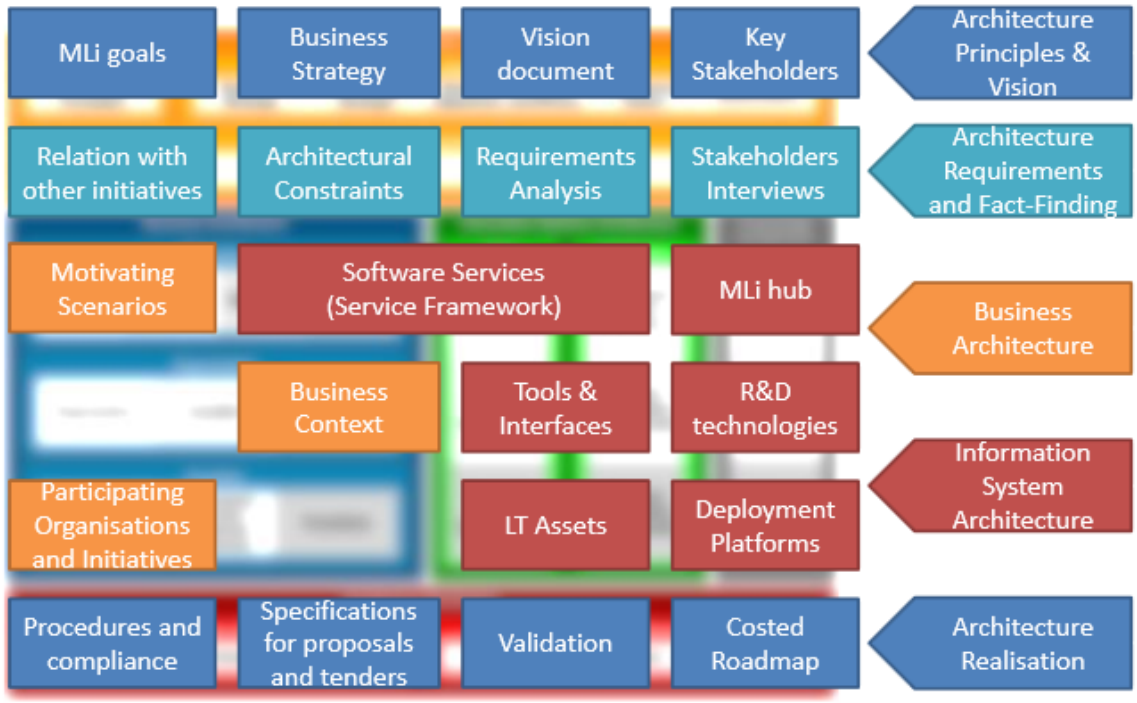
Nasledujúc Podnikovú architektúru (Enterprise Architecture), sa ako referenčný model architektúry zvolil prístup TOGAF[16]. Referenčný meta model poskytuje návod na organizáciu rôznych aspektov celého Uzla MLi do viacerých logických blokov:

- Princípy, vízia a požiadavky architektúry
- Business architektúra

[16] https://www.opengroup.org/togaf/

- Architektúra informačných systémov
- Architektúra technológií
- Realizácia architektúry

Vytvorená bola architektúra projektu MLi na pokrytie hlavných artefaktov vrcholovej úrovne projektu:



Obr. 8. Meta model referenčnej architektúry, detailnejšie rozdelený pre projekt MLi.

Analytika priemyslových dát (Projekt IDA): Analytika priemyslových dát pre sieťové odvetvia a Smart Grid

Architektúra Analytiky priemyslových dát pre Smart Grid je spoločnou platformou pre rôzne prípady použitia v oblasti sieťových odvetví. Je to projekt vykonaný v rámci spolupráce medzi firmami Atos a Siemens a je základom scenára pre sieťové odvetvia.

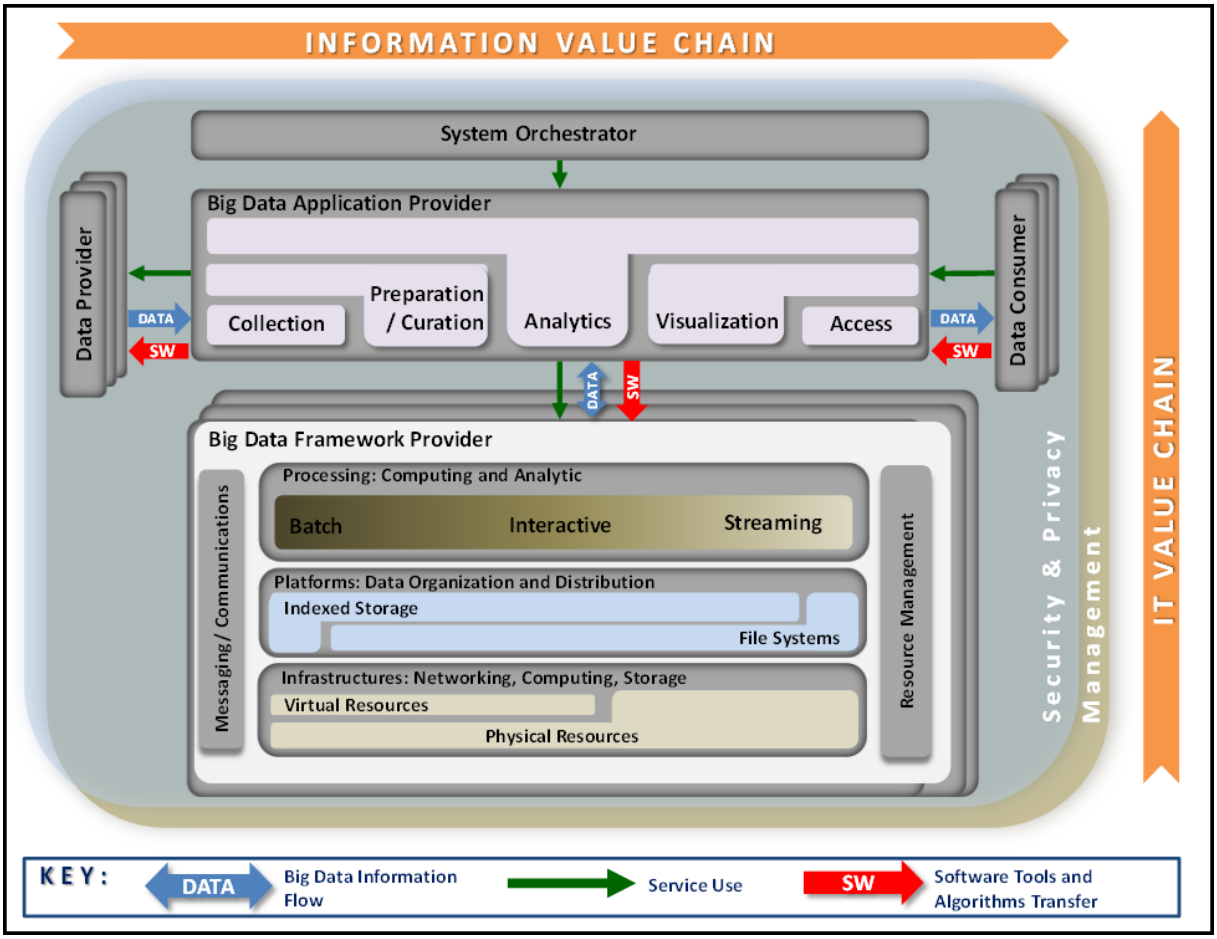
3.2.2 Štandardy týkajúce sa architektúr pre Big Data

V posledných rokoch sme boli svedkami snáh smerom ku štandardizácii určitých nástrojov big data, API, atď. Veľmi málo sa však dosiahlo v zmysle štandardizácia architektúr big data a práce tu stále ešte napredujú. Sú tu však zaujímavé pokroky, ktoré má zmysel zohľadniť na ceste smerom k definícii referenčnej architektúry.

V prvom rade existuje Študijná skupina ISO/IEC JTC 1 k Big Data (SGBD)[17]. Táto študijná skupina zahájila svoju prácu v roku 2014 a uskutočnila viacero osobných rokovaní ako v USA, tak aj v Európe. Táto skupina ISO úzko spolupracuje so skupinou NBD-PWD popísanou nižšie.

Národný inštitút pre štandardy a technológiu(NIST)[18] je agentúrou amerického Ministerstva obchodu, ktorá sa zaoberá normami. NIST spustil v roku 2014 Pracovnú skupinu NIST Big Data (NBD-PWD)[19] aby sa pokúsila priniesť štandardy na poli big data. NIST poskytuje definície, klasifikácie, požiadavky, bezpečnosť a dôvernosť a, čo je dôležité pre tento dokument, referenčnú architektúru pre big data a technický „cestovný poriadok“. Hoci práca ešte v čase písania tohto dokumentu pokračuje, odozva na NBD-PWD je dostatočne zrelá na to, aby bola predmetom záujmu pri definovaní referenčnej architektúry pre big data.

Aktuálny pohľad NBD-PWD na referenčnú architektúru big je vidno dolu:



Obr. 9. Štandardná referenčná architektúra NBD-PWD.

Zo schémy architektúry NBD-PWD sa dá odvodiť viacero dôležitých kľúčových faktov:

[17] http://jtc1bigdatasg.nist.gov/home.php
[18] http://www.nist.gov/
[19] http://bigdatawg.nist.gov/home.php

- Architektúra je zobrazená ako 2-osý diagram, kde vodorovná os znázorňuje hodnotový reťazec dát, zatiaľ čo zvislá os detailnejšie znázorňuje hodnotový reťazec IT a preto je tu nezávislosť predajcov a technológií.

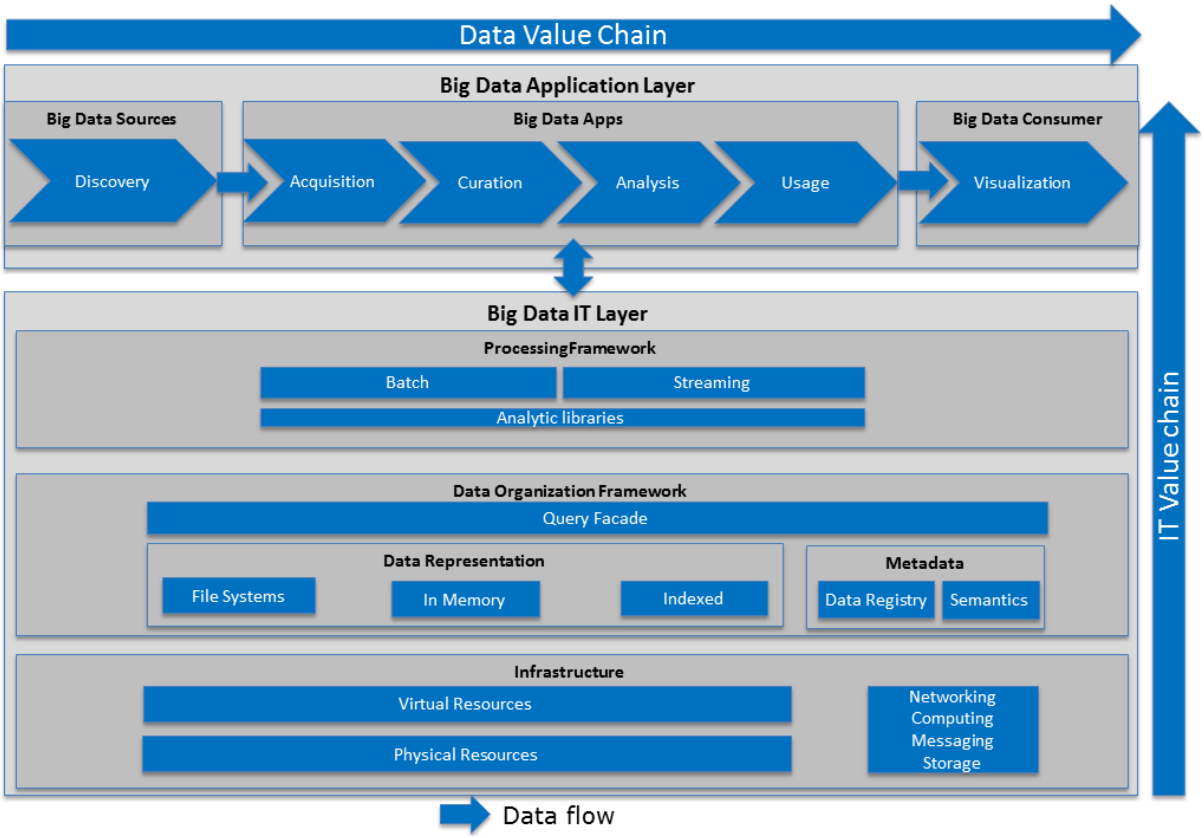
- Hodnotový reťazec dát sa podobá tomu, ktorý je v našom dokumente zobrazený skôr, avšak nie je ten istý. Táto schéma predstavuje dátovú analýzu v strede, týkajúcu sa zvyšku krokov reťazca, zatiaľ čo z obrázku chýba uloženie dát. Ukladanie dát a fyzické a virtuálne výpočtové zdroje sú zobrazené na nižšej vrstve, viac súvisiace s technologickými aspektmi. Toto je alternatívny spôsob nazerania na hodnotový reťazec dát, avšak, nakoniec, je to perfektne platné predstavenie a veľmi sa podobá na hodnotový reťazec dát, ktorý sledujeme v tomto projekte.

- Oddelenie technológie od hodnotového reťazca dát poskytuje spôsob ako oddeliť funkčné aspekty od nástrojov a postupov, čo je vhodné pri štruktúrovaní architektúry.

- Architektúra zavádza radu funkcií (Organizátor systému, Poskytovateľ dát, Spotrebiteľ dát, Poskytovateľ aplikácie Big Data a Rámcový poskytovateľ Big Data). Niektoré z týchto funkcií presahujú rámec predmetu tohto dokumentu, avšak sú jasne súčasťou funkcií, ktoré je potrebné pri definovaní architektúry big data zohľadniť.

3.3 Smerujúc k referenčnej architektúre Big Data

Zohľadniac predchádzajúce projekty a snahy o štandardizáciu, schéma predstavená na nasledujúcom obrázku zobrazuje referenčnú architektúru, ktorú bude projekt sledovať v nasledujúcich častiach:



Obr. 10. Projektová referenčná architektúra vysokej úrovne.

Obrázok hore má štruktúru dvoch hlavných osí (Hodnotový reťazec dát a Hodnotový reťazec IT), ktoré pokrývajú hlavné stavebné bloky (IT vrstva Big Data a Aplikačná vrstva Big Data). V tomto zmysle preniká IT do horných vrstiev, ako aj do dolných a je na základoch osi Hodnotového reťazca dát s viacerými funkciami. Táto architektúra je založená hlavne na snahách o štandardizáciu, vysvetlených v časti 3.2.2 a pozostáva z nasledovných vrstiev a funkčných prvkov:

Začínajúc zospodu obrázka 10, **IT vrstva Big Data** sa väčšinou vzťahuje na hodnotový reťazec IT. Potenciálne doloženie IT vrstvy Big Data príkladmi sa nemusí riadiť technologickou zostavou od konkrétneho predajcu, alebo riešením Open Source, t.j. z otvoreného zdroja. Ako sme už spomenuli, predajcovia pokrývajú niektoré alebo všetky stavebné bloky tejto vrstvy. V praxi môže príklad architektúry zvoliť komponenty a nástroje od mnohých predajcov/riešení otvorených zdrojov na pokrytie funkčnosti stavebných blokov. Toto je podstata referenčnej architektúry: architektúra vysokej úrovne, ktorá abstrahuje od skutočného zavedenia konkrétneho riešenia.

- Stavebný blok **Infraštruktúra (Infrastructure)** je zodpovedný za poskytnutie potrebných zdrojov na hostiteľstvo/zbiehanie činností ostatných komponentov Big Data systému. Infraštruktúra pozostáva z **kombinácie fyzických zdrojov, ktoré môžu byť hostiteľmi/podporovať podobné virtuálne zdroje**. Tieto zdroje sa týkajú elementov ako je **networking** a konektivita, **computing** a vizualizácia, **messaging**,

a distribuované uloženie (**storage),** starajúce sa o veľké objemy dát s vysokou rýchlosťou.

- Stavebný blok **Organizačného rámca dát** poskytuje vrstvu na organizáciu a distribuovanie dát súbežne s poskytovaním metód prístupu.
    - **Query Façade**: „Fasáda dopytovania“ poskytuje súbor API, služby a jazyky dopytovania podobné SQL potrebné na prístup k dátam.
    - **Prezentácia dát**: Ako bolo diskutované vyššie, rámec big data má obyčajne vrstvu „Data Representation“ založenú na distribúcii dát. Táto prezentácia môže využívať distribuované súborové systémy, indexované (SQL, NoSQL, newSQL, index. jednotky, atď.) alebo v pamäti uložené prístupy.
    - **Metadáta**: V niektorých prípadoch si môže organizácia dát vyžadovať metadáta, obyčajne poskytnúc dátový register a súvisiace sémantické popisy dát.
- Stavebný blok **Framework spracovania** (**Processing Framework)** poskytuje výpočtovú infraštruktúru potrebnú na zavedenie aplikácií big data. Rôzne kroky hodnotového reťazca dát obyčajne vyžadujú podporu nástrojov a komponentov spracovania dát, ktoré definujú, ako sú organizované výpočtová činnosť a spracovanie dát. Spracovanie dát obyčajne spadá do dvoch kategórií: dávkové a streamové. Veľmi to závisí na špecifických firemných potrebách a na latentnosti potrebnej na prijímanie, spracovanie a poskytovanie analytických výsledkov. Celkovo budú mnohé architektúry Big Data zahŕňať viaceré frameworky na podporu širokej škály požiadaviek.
    - **Dávkové spracovanie**: Frameworky dávkového spracovania rozsiahlych dát sú jadrom toho, čo väčšina chápe ako Big Data. Obyčajne tieto frameworky poskytujú model programovania, kde sa dajú v určitej mierke aplikovať rôzne algoritmy. Tieto modely spracovania sú normálne distribuované v priereze viacerých uzlov určitého klastra. Hlavným modelom dávkového spracovania je Map-Reduce, hoci niektoré iné nástroje a frameworky využívajú odlišné prístupy.
    - **Streamové spracovanie**: Frameworky spracovania streamových dát sú vhodné na rýchle spracovanie dát. Niektorí autori nazývajú tieto frameworky Fast Data. Schopnosť prijímať a spracovať dáta s požadovanou rýchlosťou je v týchto frameworkoch kľúčovým aspektom. Frameworky ako sú Spark Streaming alebo Apache Storm, často doplnené o nástroje fast messaging, publish-subscription alebo Control Event Processing (CEP) poskytujú tento typ frameworkov.

- **Analytické knižnice**: Hlavným cieľom aplikácie Big Data je často poskytnúť detailné pohľady založené na dátach, na odvodenie podnikateľských rozhodnutí. Analytické knižnice sú kľúčovým prvkom fáz spracovania a analýzy. Analytické knižnice sú v mnohých prípadoch napojené na špecifické frameworky spracovania dát (t.j. Spark MLlib pre knižnicu Machine Learning), zatiaľ čo iné knižnice a analytické nástroje sa môžu pripojiť k architektúre.

Na druhej strane, vrstva **Data Application Layer** zobrazená na obrázku 10 úzko dodržiava koncepciu dátového hodnotového reťazca - Data Value Chain. V tejto architektúre sme sa rozhodli zahrnúť na ľavom a pravom okraji tejto vrstvy Objavovanie dát (Data Discovery), resp. Vizualizáciu dát (Data Visualization). Tieto pokrývajú kroky pred zberom dát a konečné postupy vizualizácie poskytnuté koncovým užívateľom. Stavebné bloky zvažované vo vrstve Aplikácie dát sú preto nasledovné:

- **Big Data Sources – Zdroje Big Data**: Tento stavebný blok reprezentuje rôzne zdroje dát, ktoré zásobujú systém. Dáta môžu byť interné z organizácie konečného užívateľa, alebo externé, z rôznych zdrojov. Zdroje externých dát môžu byť veľmi rozličné, napríklad ako Internet of Things (senzory a zariadenia IoT), sociálne siete (Twitter, Facebook, atď.), Open Data, zápisy štruktúrované dáta, atď..

    - **Objavenie dát** - **Discovery**: Tento krok dátového hodnotového reťazca pokrýva aspekty predchádzajúce zberu dát. Je o identifikácii správnych dátových zdrojov pre riešený problém, ešte pred samotnou akvizíciou. Ľudia, ktorí majú toto na starosti, budú často potrebovať zohľadniť aspekty účtu ako je dostupnosť dát, cena, kvalita dát zdroja, otvorené API, použitie ETL, atď.

- **Big Data Apps**: Toto je hlavný stavebný blok zobrazujúci dátový hodnotový reťazec od zberu dát až po ich použitie. Hovoriac všeobecne, táto vrstva buduje aplikácie zhora na nástrojoch vrstvy Big Data IT, na zber dát, ich usporiadanie, analýzu a poskytnutie výsledkov analýzy k dispozícii. Hlavné funkčné bloky sa zhodujú s typickým dátovým hodnotovým reťazcom (**Akvizícia** dát, **Usporiadanie** dát, **Analýza** dát a **Používanie** dát).

- **Spotrebitelia Big Data**: Tento stavebný blok predstavuje pohľad spotrebiteľa dát a najviac sa týka vizualizácie výsledkov analýzy. Preberá výsledky z posledného kroku (použitie) predchádzajúceho stavebného bloku a poskytuje prostriedky na to, aby sa stali zrozumiteľnými pre človeka. Hlavným funkčným krokom tohto stavebného bloku je **Vizualizácia**.

3.4 Konkretizácia referenčnej architektúry

Existuje veľká výzva súvisiaca s konkretizáciou referenčnej architektúry Big Data: **rýchle zabezpečenie a ustanovenie** určitej architektúry Big Data, podporujúcej využívanie zdrojov dát na všetkých vrstvách sústavy Big Data a nákladovo efektívne objavovanie a výber

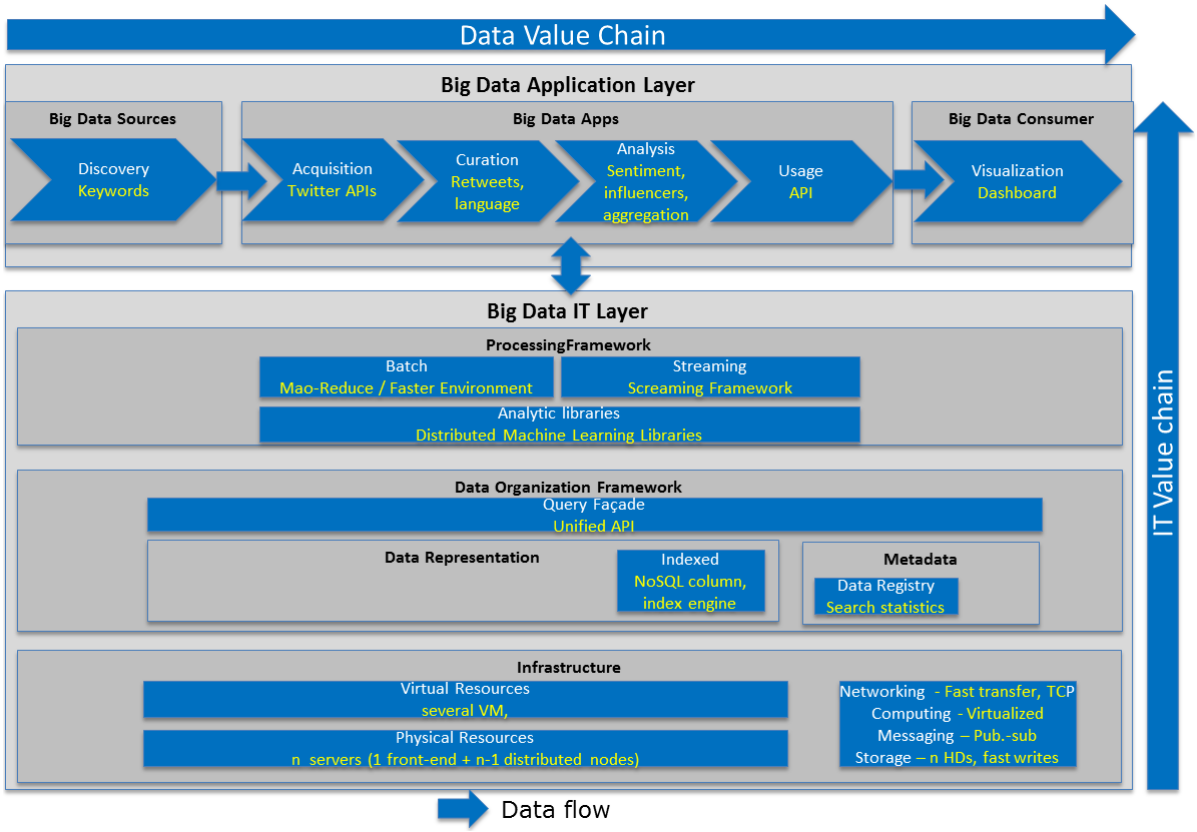
výpočtových zdrojov a zdrojov dát. V tejto časti predstavíme pár príkladov konkretizácie výberu architektúry pre dva scenáre: spracovanie textových dát a SmartGrid.

3.4.1 Architektúra Big Data pre textové dáta

Táto časť zahŕňa:

- Referenčnú architektúru pre narábanie s veľkými objemami neštruktúrovaných dát
- Návrh rámca pre zber a analýzu neštruktúrovaných dát, obzvlášť zo sociálnych sietí a webu

Aby sme zdôvodnili technickú architektúru pre špecifický scenár spracovania textov, zvolili sme potenciálnu aplikáciu, ktoré zbiera dáta zo sociálnych sietí za účelom poskytnutia určitej analytiky. Referenčná architektúra pre tento typ aplikácie je zobrazená na obrázku.



Obr. 11. Úvodná charakterizácia referenčnej architektúry pre spracovanie textov.

Obrázok ukazuje počiatočné kroky na adaptáciu navrhovanej referenčnej architektúry na analýzu textových informácií pochádzajúcich z Twitteru. Hlavné rozhodnutia a požiadavky, ktoré majú byť pokryté a ktoré súvisia s funkčnými stavebnými blokmi, sú vysvetlené nižšie:

IT vrstva Big Data

- **Infraštruktúra:** Tento scenár by mal poskytnúť minimálnu infraštruktúru schopnú škálovania v závislosti na potrebách (t.j. počet streamov Twitteru – potrebné 1 IP na každý stream, alebo počet tweetov za sekundu na spracovanie). Minimálnou

konfiguráciou pre testovanie by mohla byť lokálna inštalácia, avšak sú potrebné minimálne 3 hlavné virtualizované uzly na zabezpečenie minimálnej distribúcie spracovania a ukladanie. Postačuje cloudová infraštruktúra od poskytovateľa cloudov.

- **Framework pre organizáciu dát**:
    - **Reprezentácia dát**: Dáta z Twittra a výsledky analýzy by mali byť ukladané podľa indexovaného modelu, ak má byť systém schopný umožniť plnotextové vyhľadávanie a ponúknuť škálu dopytovaní na dátach. Odporúčame kombináciu databázy NoSQL (HBase alebo MongoDB) pre surové dáta a hlavné dopytovania, plus Apache Solr pre vyspelejšie mechanizmy dopytovania.
    - **Metadáta**: Na kategorizáciu dát z Twittra je potrebný minimálny súbor metadát. Tweety sa dajú zoskupiť pod ovládanými slovníkmi (kategóriami), objektmi záujmu (t.j. zoskupenie tweetov hovoriacich o tom istom podujatí) atď.
    - **Query Façade**: Dobre by bolo mať vlastnosti rozhraní RESTful na prístup ku zvoleným úložiskám, prinajmenšom pre najbežnejšie dopytovania. Tiež budú zaujímavé mechanizmy na kešovanie informácií, ktoré sa tak často nemenia.
- **Framework spracovania:**
    - **Dávkové spracovanie**: Použitie Hadoop v kombinácii s Apache Flink (alebo alternatívne Apache Spark).
    - **Streamové spracovanie**: Odporúča sa použitie Apache Flink alebo Spark Streaming. Pre určité účely je tiež možnosťou Apache Storm.
    - **Analytické knižnice**: Odporúča sa použitie R knižníc pre Machine Learning v kombinácii s Flink alebo Spark. Flink a Spark poskytujú zabudované knižnice ML, ktoré sa dajú použiť, kombinované s R a Apache Mahout pre vyspelejšie vlastnosti.

Vrstva dátovej aplikácie

- **Zdroje Big Data**:
    - **Objavovanie** dát: Tento krok je o rozhodnutí o tom, ktoré dáta sa majú zbierať a o zdrojoch dát. V tomto prípade sa dajú twittrové dáta zbierať s použitím otvorených API Twitter (ako sú API Twitter Search alebo Twitter Streaming), alebo sa dá zaplatiť tretím poskytovateľom dát aby dáta zbierali pre vás. Obe možnosti majú svoje za a proti. V podstate, ak potrebné dáta neprekračujú rámec limitov otvorených API Twitter, tak sa odporúča použitie týchto API a je

to lacnejšie. V opačnom prípade je nutné kontaktovať poskytovateľa dát a vykonať posúdenie nákladov ešte predtým, než sa budú vôbec realizovať zvyšné kroky. V tomto prípade sme sa rozhodli pre získavanie dát s použitím otvorených API Twitter. V tomto kroku sa musí vykonať posúdenie dostupných dopytovaní na API, kľúčových slov, hashtagov, atď., aby sa preverila dostupnosť dát a ich vhodnosť pre ďalšiu akvizíciu.

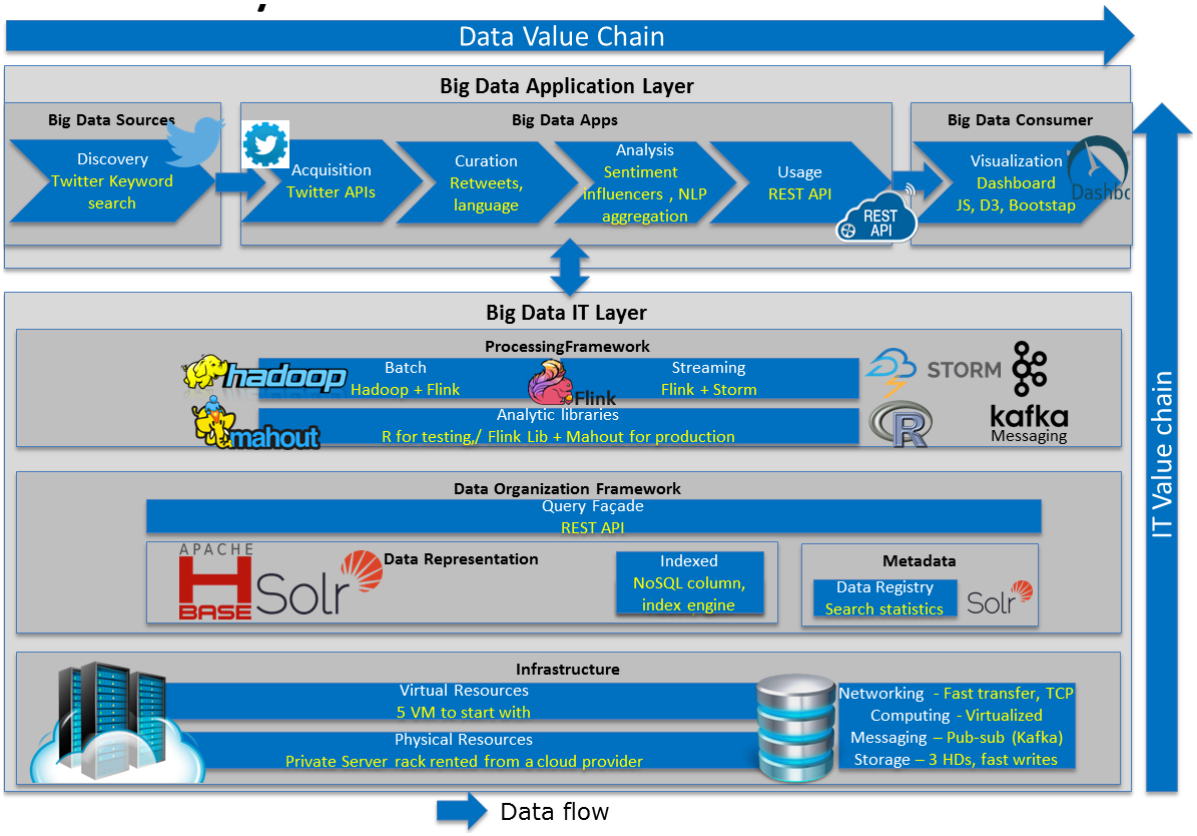
- **Aplikácie Big Data**:
    - **Akvizícia** dát**:** Zber dát zo SN (a konkrétne pre Twitter) sa dá vykonávať jednoducho s použitím uvedených otvorených API. Je potrebné prijať rozhodnutia o pretrvávaní zozbieraných dát a dátovom modeli. V závislosti na budúcich potrebách môže systém získavať dáta a ukladať ich pre ďalšie použitie alebo predbežne spracovať v takmer reálnom čase. Malo by sa posúdiť používanie ako vyhľadávacích, tak aj streamových API. V tomto konkrétnom prípade budeme mať scenáre použitia ako pre dávkové spracovanie (založené na kombinácii Twitter Search a Streaming API) a streamové spracovanie (založené výlučne na Twitter Streaming API). Nutné je používanie engines ako dávkového, tak aj streamového spracovania z vrstvy IT. Pre dávkové spracovanie je základom Hadoop hoci Apache Flink (alebo Apache Spark) sa dá upraviť tak, aby poskytoval odporu pre spracovanie dávkové a v reálnom čase. Pre spracovanie v reálnom čase by mali byť alternatívami Storm alebo Spark. Získané dáta musia byť pre ďalšie použitie uložené v HBase.
    - **Usporiadanie** dát**:** Dáta z Twittra musia byť usporiadané a spracované. Niektoré tweety sa vyfiltrujú po predbežnom spracovaní (t.j. retweets, tweety z nepoužívaných jazykov, tweety obsahujúce iba linky a žiadny text, atď.). Ako pre dávkové usporiadanie, tak aj pre usporiadanie v reálnom čase sa môžu použiť niektoré dokonalejšie mechanizmy usporiadania.
    - **Analýza** dát: Systém musí byť pripravený na použitie analytických knižníc. Ako Flink, tak aj Spark poskytujú zabudované knižnice pre strojové učenie, ktoré môžu pracovať v dávkach alebo v reálnom čase. Tieto dva systémy neposkytujú rozsiahly súbor algoritmov ML, takže alternatívou môže byť používanie Apache Mahout (alebo Weka) na doplnenie analytických knižníc. Systém musí zaviesť aplikáciu, ktorá vykonáva dve hlavné analýzy: analýzu sentimentu a vplyvu. Čo sa týka sentimentu, systém priraďuje sentiment tweetom a potom agreguje údaje sentimentu pre rôzne kritériá. Sentiment by sa dal vypočítať s použitím špecializovaných knižníc a zdrojov (t.j. Stanford alebo iní predajcovia), alebo implementovať. V tomto prípade sme sa rozhodli pre implementáciu engine strohej analýzy sentimentu a využitím čistého prístupu ML plus niektorých slov vyjadrujúcich náladu. Čo sa týka

ovplyvňujúcich subjektov, systém by mal skúmať profily niektorých kľúčových užívateľov, aby sa určilo, kto ovplyvňuje v škále konkrétnej značky. Existuje obšírna literatúra o výpočte vplyvu. Výsledky analýzy musia byť vhodne formátované a indexované, najpravdepodobnejšie s použitím ako HBase, tak aj Solr, v závislosti na povahe výsledkov.

- **Použitie** dát**:** Systém musí poskytovať na používanie jednoduché servisné rozhranie pre prístup k výsledkom analýzy. Musia byť poskytované agregované dáta, dáta sentimentu, hlavné ovplyvňujúce subjekty a niektoré funkcie vyhľadávania. Pretože dáta sa ukladajú najmenej v dvoch rôznych distribuovaných úložiskách (HBase a Solr), vrstva použitia dát musí poskytovať metódy na uniformný prístup k dátam. Toto sa vykoná zavedením vyhradeného REST API popri metódach poskytnutých databázami a Query Façade.

- **Spotrebiteľ Big Data**:
  - **Vizualizácia** dát**:** Cieľom je poskytnúť Dashboard – prehľadnú stránku - ukazujúcu výsledky monitorovania. Zvolili sme jednoduché dashboardy založené na Javascript a HTML5 pre ich relatívnu jednoduchosť a potenciál opätovného použitia existujúcich vizualizácií zavedených týmito technológiami.

Preto môže technická architektúra nižšej úrovne odvodená z predchádzajúcej diskusie vyzerať tak, ako je to zobrazené na obrázku.



Obr. 12. Architektúra pre spracovanie textov.

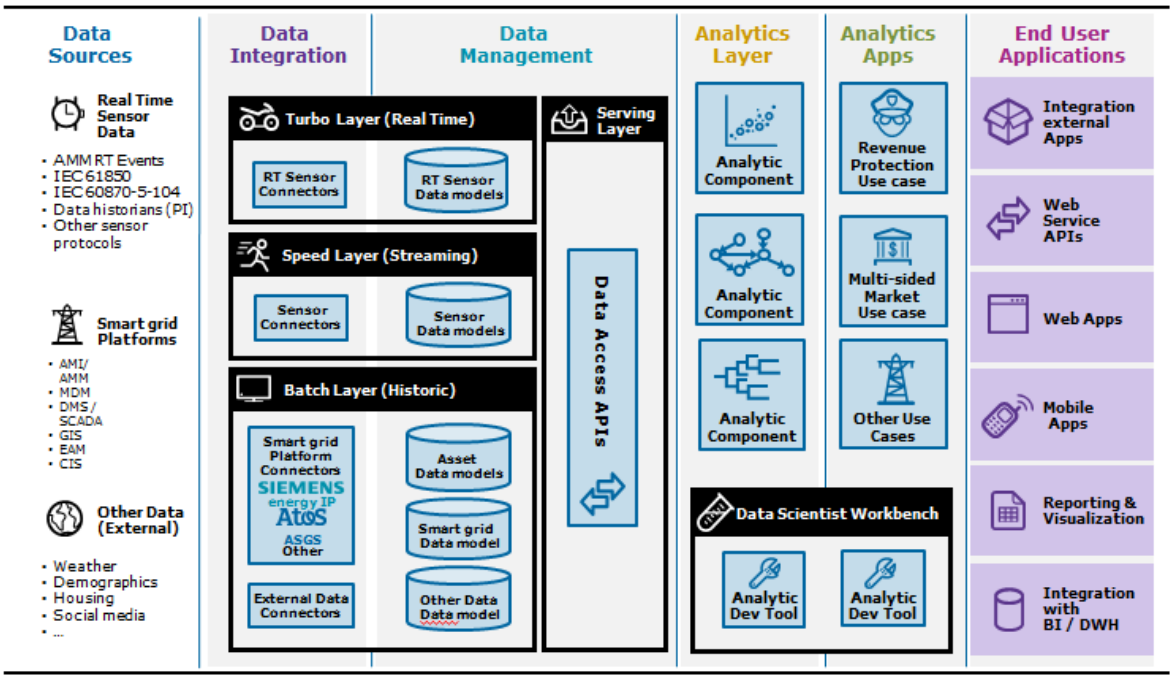
Architektúra IT a softvéru zobrazená na obrázku na vrstve Big Data IT poskytuje základ pre vývoj aktuálnych aplikácií zložených na dátach Twitter vo vrstve Aplikácia Big Data. V tomto zmysle je navrhovaná architektúra pripravená pre účel tejto časti (príklad architektúry). Avšak na charakterizáciu používania rôznych prvkov vrstvy aplikácie Big Data potrebujeme detailnejšie vysvetliť, ktorý typ aplikácií a nástrojov sa používa v každom z krokoch hodnotového reťazca dát, potrebný typ spracovania a použité nástroje.

3.4.2 Architektúra Big Data na spracovanie energetických dát (Tino)

Analytika priemyselných dát (Industrial Data Analytics) pre architektúru Smart Grid je spoločnou platformou pre rôzne prípady použitia v sieťových odvetviach, charakterizovanou rozdelením do troch rôznych vrstiev akvizície a spracovania dát:

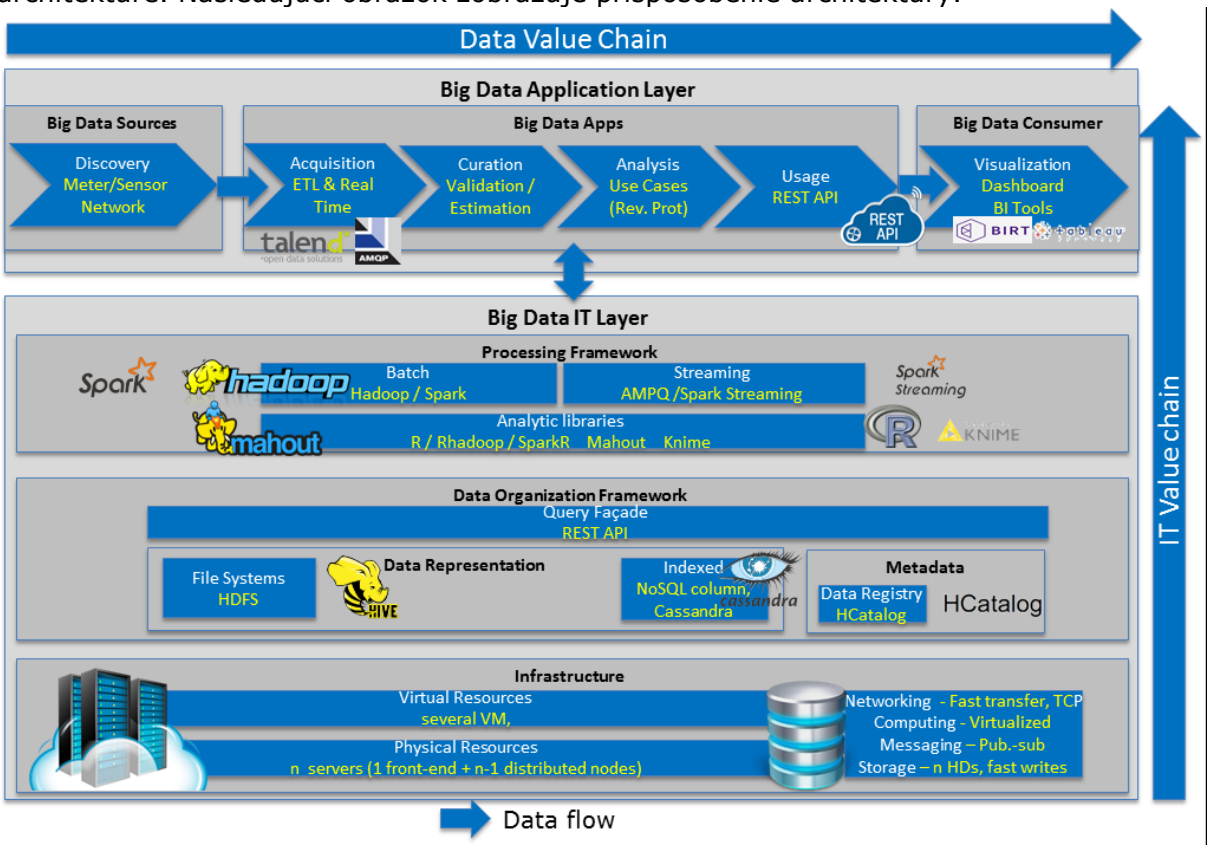
- Vrstva turbo (Turbo): akvizícia dát v reálnom čase z dát senzorov
- Rýchlostná vrstva (Speed): akvizícia streamových dát a platforiem smart grid.
- Dávková vrstva (Batch): Offline akvizícia a spracovanie dát pre externé zdroje informácií ako sú otvorené datasety, sociálne médiá, atď.

Obrázok dolu zobrazuje hlavné komponenty a vrstvy tejto generickej architektúry:



Obr. 13. Spoločná platforma pre rôzne prípady využitie v sieťových odvetviach.

Táto podrobná architektúra Big data pre oblasť energetiky sa dá mapovať na generickej architektúre. Nasledujúci obrázok zobrazuje prispôsobenie architektúry.



Obr. 14. Architektúra spracovania energetických dát.

Obdobne ako v predchádzajúcom scenári, sú hlavné prvky použité v tejto architektúre zobrazené na obrázku vysvetlené nižšie:

Objavenie dát

Ako sme videli, existuje rôznorodosť zdrojov dát, takže môže byť ťažké generalizovať Objavenie dát. Máme však dva široké prípady:

- Operačné tendencie majú tendenciu k tomu, že majú dôležitý komponent reálneho času a poskytujú protokoly zasielania správ pre výmenu informácií
- Informačné technológie sú obyčajne viac „dávkovo orientované“ a poskytujú rôzne typy API pre objavovanie dát, obyčajne založené na modeloch Enterprise SOA (založené na SOAP), avšak v poslednej dobe sa trh presúva smerom k architektúram založeným na REST. Aj tak sú protokoly „legacy“ File Transfer ešte stále časté.

Akvizícia dát

Podľa predchádzajúceho bodu máme nasledovné:

- U protokolov reálneho času existuje veľká rôznorodosť. V poslednom čase sa stávajú populárnymi protokoly zasielania správ z otvorených zdrojov ako AMQP alebo 0MQ.
- U IT systémov je bežným modelom prevádzky ETL (Extraction-Transformation-Load). Existuje niekoľko vyzretých platforiem ETL z otvorených zdrojov, ako Talend.

Čo sa týka **ukladania dát**, úložiská pre zavedenie Lambda architektúry sa dajú používať pre rôzne typy dát:

- Pre operačné dáta sa používa rýchla kľúčová hodnota alebo stĺpcová databáza, ako je Cassandra, na ukladanie dát, ku ktorým je potrebný častý prístup v dopytovaniach s nízkou latentnosťou (Speed Layer – rýchlostná vrstva).
- Pre historické dáta je používanie Hadoop HDFS pre dlhodobé ukladanie s Hive pre komplexné dopytovania a analýzy (Batch Layer – dávková vrstva).

Pre **framework spracovania** sa jednou z najpopulárnejších distribuovaných zostáv spracovania stáva Spark a vie zabezpečiť prístup k obom dátovým pamätiam, takže sa niekedy stáva „lepidlom“ pre typickú Vrstvu služieb (Service Layer)

Usporiadanie dát

V kontexte dát súvisiacich s energetikou sa dá koncepcia usporiadania dát sa dá integrovať do postupu Validácia-Odhad-Publikovanie (Validation-Estimation-Edition), alebo VEE. Dáta sa spracovávajú na nájdenie potenciálnych problémov (nesprávne dáta, neexistujúce dáta, alebo dáta, ktoré si vyžadujú pred použitím určité spracovanie.

Analýza dát

Ako sme videli, existuje mnoho rôznych prípadov použitia a existuje veľká rôznorodosť potenciálnych analytických metód. Nástroje môžu byť rôzne, ale obyčajne môžeme vidieť:

- Spark, so súvisiacimi knižnicami strojnej výučby (machine learning), ako sú Mahout alebo MLib, pre scenáre, kde je kritická výkonnosť, ako je streamové spracovanie.
- R, hoci nie je pre určité scenáre dostatočne výkonný, je veľmi populárnym jazykom z dôvodu svojej rozsiahlej programovej knižnice.

Použitie dát

Pri použití dát je jasný, jednoznačný trend: implementácia REST API pre prístup k heterogénnym dátovým bázam a k výsledkom spracovania. Jednou významnou špecializáciou REST pre Prístup k dátam je OData.

Vizualizácia dát

Na úrovni vizualizácie existujú dve hlavné možnosti:

- Vytvorenie na mieru šitých prehľadových stránok / portálov, založených na modeli REST, s využitím webových technológií, ako sú frameworky HTML5 a Javascript, ako Angular.js
- Použitie end-user BI koncového užívateľa a nástrojov reportingu. Niektorými populárnymi príkladmi sú Tableau (komerčné), BIRT (otvorený zdroj).

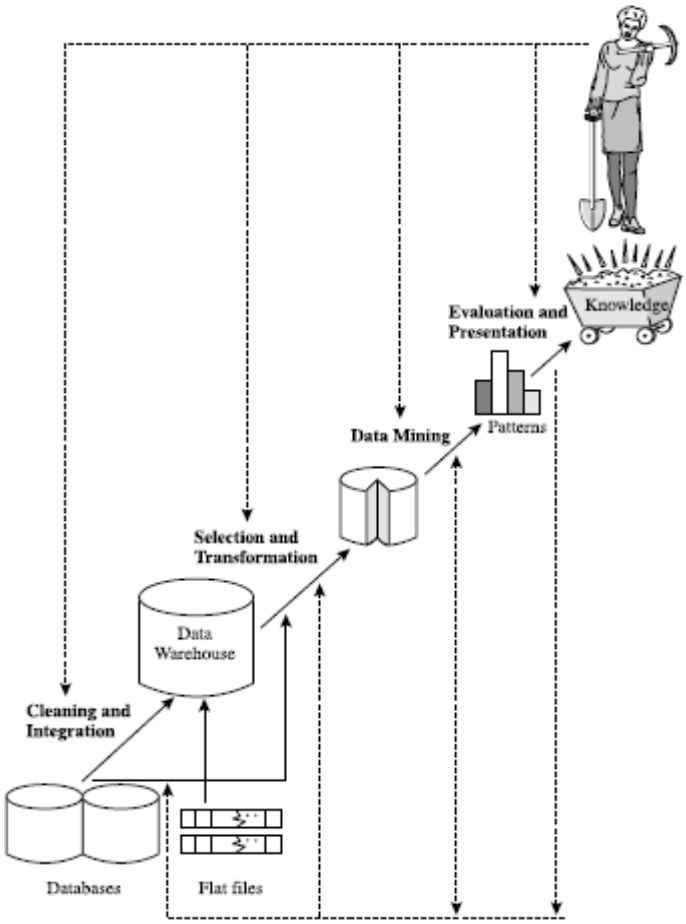
4 Výskum metód inteligentnej analýzy veľkých množín dát

Metódy pre analýzu dát a získavanie informácií z dát sú súhrnne označované ako dolovanie v dátach (angl. *data mining*) a sú súčasťou procesu objavovania znalostí (angl. *knowledge discovery from data*, skr. *KDD*). V rámci tejto vedeckej disciplíny bolo vyvinutých už mnoho metodológií pre dátovú analýzu [48]. Aktuálnou výzvou je sprostredkovanie existujúcich prístupov koncovým používateľom – dátovým analytikom, ktorí ich môžu využiť pre analyzovanie obrovských množstiev zozbieraných dát. Na základe vydolovaných informácií potom dokážu ľahšie a lepšie vykonávať rozhodnutia, ktoré budú prospešné pre príslušnú organizáciu.

Objavovanie znalostí (angl. *knowledge discovery from data*, skr. *KDD*) je proces extrakcie a sumarizácie užitočných informácií z veľkej množiny dát pomocou metód matematiky, štatistiky, umelej inteligencie, strojového učenia a ďalších metód.

Zasahuje sem viacero oblastí. Databázové systémy poskytujú platformu pre ukladanie a prácu s dátami pomocou dopytovacích jazykov a techniky škálovania pre prácu s veľkými objemami dát. Oblasť dátových štruktúr a algoritmov poskytuje techniky na návrh a analýzu algoritmov a dátové štruktúry. Teória pravdepodobnosti a štatistiky je značne využívaná v metódach pre dolovanie v dátach. Techniky strojového učenia (napr. rozhodovacie stromy, neurónové siete) sú tiež súčasťou mnohých algoritmov. Oblasť vyhľadávania informácií (angl. *information retrieval*) umožňuje dolovanie v textových dátach, webových zdrojoch a obsahuje metódy pre určenie miery podobnosti, ktoré sú využívané v algoritmoch aj pri vyhodnocovaní metód dolovania.

Proces (pozri obrázok 15) sa skladá z nasledujúcich krokov: (a) očistenie dát, (b) integrácia dát, (c) výber dát, (d) transformácia dát, (e) dolovanie v dátach, (f) identifikácia vzorov, (g) prezentácia znalostí.



Obr. 15. Proces objavovania znalostí [48].

Vo fáze očistenia dát sa upravujú dáta tak, aby boli použiteľné pre ďalšie spracovanie. Vykonávajú sa činnosti ako dopĺňanie alebo odstraňovanie neúplných záznamov, prevod záznamov na jednotný formát, odstraňovanie duplicít, príliš odchýlených hodnôt a pod. Účelom tejto fázy je spoznanie dát a ich príprava pre ďalšiu prácu.

Počas integrácie dát sa zoskupujú dáta z rôznych zdrojov do jednej množiny, z ktorej sa budú dolovať znalosti. Integrované dáta sú zvyčajne uložené do dátového skladu.

Ďalšími krokmi sú výber dát a transformácia dát. Účelom týchto krokov je pripraviť dáta tak, aby bolo možné v nich už dolovať. Výber dát a ich jednotlivých atribútov je ovplyvnený cieľom dolovania, t. j. tým, čo chceme z dát vydolovať a ako. Dáta sa transformujú do požadovanej podoby podľa plánovaného spôsobu dolovania. Pripravuje sa vhodný vstup pre zvolený algoritmus alebo metódu. Počas transformácie dát prebiehajú úlohy ako vyhladzovanie, agregácia, zovšeobecňovanie, normalizácia dát, konštrukcia atribútov a pod.

Predchádzajúce štyri kroky sa súhrne nazývajú predspracovane dát. Predspracovanie zaberá v procese objavovania znalostí najviac času, pretože má veľký vplyv na výsledok. V prípade veľkých dátových korpusov, vzhľadom na časovú zložitosť tejto časti procesu, nemusí byť na predspracovanie kladený tak veľký dôraz za cenu menej presných výsledkov.

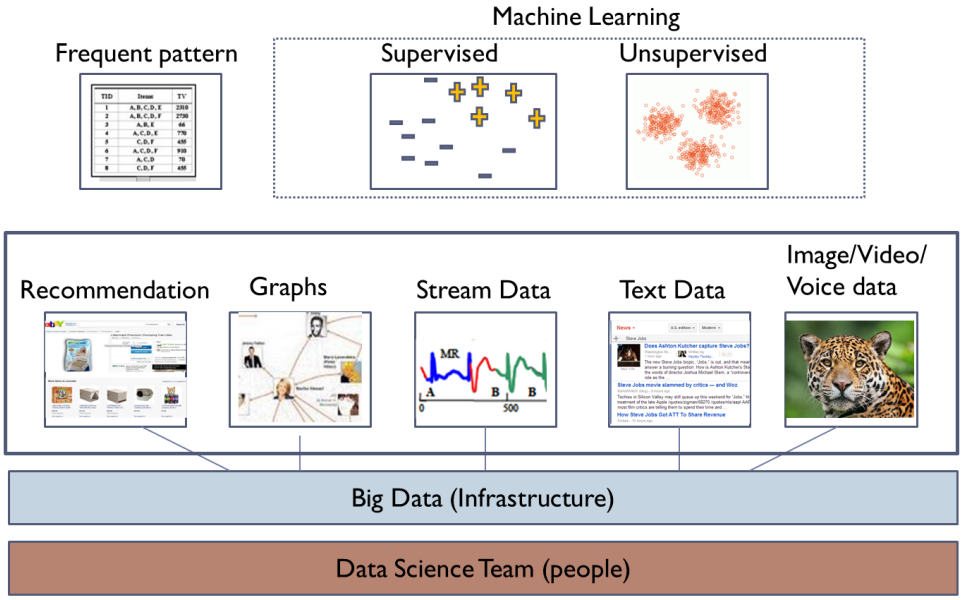
Hlavným krokom procesu objavovania znalostí je dolovanie v dátach. V tejto fáze je na predspracované dáta aplikovaný algoritmus alebo metóda, ktorá bola zvolená na základe typu vstupných dát (audio, video, text a pod.) a požadovaného výsledku dolovania. Dolovanie v dátach rieši štandardne dve základné skupiny úloh – prediktívne a deskriptívne. Predikcia na základe známych atribútov z dát predpovedá budúce hodnoty atribútov, ktoré nás zaujímajú. Deskriptívne úlohy objavujú v dátach vzory, ktoré charakterizujú dáta. Výstupom dolovania v dátach je zostrojený (prediktívny alebo deskriptívny) model.

Po vytvorení modelu metódou pre dolovanie v dátach je na rade jeho overenie. Pre overenie modelu sa zvyčajne využívajú odlišné dáta ako pri jeho vytváraní (tzv. trénovacia a testovacia množina). Model je možné ohodnotiť aj rôznymi metrikami, napr. vypočítanou chybou predikcie. Pre každú metódu pre dolovanie v dátach existuje vhodný spôsob overenia vytvoreného modelu.

Posledným krokom procesu objavovania dát je prezentácia vydolovaných znalostí. V tejto časti sa prezentuje vytvorený a overený model, ale najmä závery, ktoré vyprodukoval z našich dát a ktoré boli cieľom zadefinovanej úlohy. Pre väčšiu prehľadnosť a ľahšie pochopenie dát človekom sa využívajú vo veľkom rôzne techniky vizualizácií výsledkov a dát.

Oblasť dolovania v dátach (hlavný krok v procese objavovania znalostí) nemá presne definované podoblasti a spadá do nej mnoho rôznorodých techník a algoritmov. Niektoré zdroje odkazujú dolovaním v dátach na všeobecné metódy používané v tejto oblasti, napr. objavovanie vzorov, strojové učenie. Občas sa dolovaním v dátach myslia metódy zamerané na určitý typ dát, napr. dolovanie v texte, dolovanie v audiovizuálnych dátach, ktoré využívajú všeobecné metódy na špecifický typ dát. Navyše pojem dolovanie v dátach (angl. *data mining*) je v literatúre často zamieňaný s pojmom objavovanie znalostí (angl. *knowledge discovery from databases*) a sú chybne označované za identické pojmy. Preto existuje mnoho rozdelení metód pre dolovanie v dátach. Aby sme sa vyhli zmätočným označeniam, uvádzame rozdelenie [58], ktoré delí oblasť dolovania v dátach na štyri časti (pozri obrázok

16). V hornej vrstve sa nachádza základná metodológia, napr. strojové učenie, objavovanie častých vzorov. V druhej vrstve sa nachádzajú metódy, ktoré aplikujú základnú metodológiu na určité typy dát. Pod touto vrstvou sa nachádza infraštruktúra, kam patrí výskum nových technológií pre ukladanie a spravovanie veľkých dátových korpusov, napr. NoSQL databázy, Hadoop a pod. Posledná vrstva reprezentuje tím odborníkov, ktorí rozumejú všetkým predošlým vrstvám.



Obr. 16. Rozdelenie oblasti dolovania v dátach [58].

4.1 Základné techniky dolovania v dátach

Dolovanie v dátach zahŕňa päť hlavných techník:

- *sumarizácia* – opis dát pomocou štatistických metrík a vizualizácií vo forme grafov, tabuliek, viacdimenzionálnych kociek a pod.,
- *detekcia extrémov* (angl. *outliers detection*) – objavovanie nezvyčajných prvkov v dátovej množine, napr. podvodné bankové transakcie, využívané sú techniky ako štatistický test, metriky vzdialenosti a odchýlok atribútov prvkov,
- *objavovanie častých vzorov, asociácií, korelácií* – objavovanie často sa vyskytujúcich skupín prvkov, sekvencií, podgrafov, stromov a pod., patrí sem napr. technika asociačných pravidiel,
- *predikcia* – predpovedanie hodnoty želaného atribútu nového prvku na základe hodnôt jeho ostatných atribútov alebo jeho minulých hodnôt, podľa typu predpovedaného atribútu sa delí na klasifikáciu (nominálny atribút) a regresiu (spojitý atribút), patrí sem napr. technika rozhodovacích stromov, náhodných lesov, podporných vektorov (angl. *SVM*), neurónové siete, logistická regresia, lineárna regresia,

- *zhlukovanie* – objavovanie tried prvkov tak, aby prvky v triede si boli čo najviac podobné a prvky z rôznych tried boli čo najviac odlišné, zhluky môžu byť hierarchické alebo sa môžu vytvárať rozdeľovaním, pri čom sa využívajú vzdialenosti prvkov (napr. *k-means* algoritmus), hustota rozloženia prvkov alebo pravdepodobnosť výskytu prvku v danej triede (napr. *Expectation-Maximization* algoritmus).

Techniky predikcia a zhlukovanie využívajú vo veľkej miere algoritmy a vedomosti z oblasti strojového učenia, kde sa nazývajú učenie s učiteľom (predikcia) a učenie bez učiteľa (zhlukovanie).

4.1.1 Sumarizácia dát

Táto technika patrí medzi deskriptívne úlohy dolovania v dátach. Umožňuje spoznať a opísať dáta, resp. opísať nejaký koncept alebo triedu, ktorú dáta predstavujú. Opisy sa dajú odvodiť buď pomocou:

a) *charakteristiky dát* – zhrnutie dát z triedy, ktorú skúmame (cieľová trieda),

b) *diskriminácie dát* – porovnanie cieľovej triedy s jednou alebo viac kontrastnými triedami,

c) alebo oboma spôsobmi súčasne.

Charakteristika dát je zhrnutie všeobecných charakteristík a vlastností dát z nami vybratej triedy (cieľovej triedy). Dáta zodpovedajúce nami vytvorenej triede sú zvyčajne výsledkom určitého dopytu, napr. výrobky, ktorých predaj sa zvýšil za posledný rok o 10%.

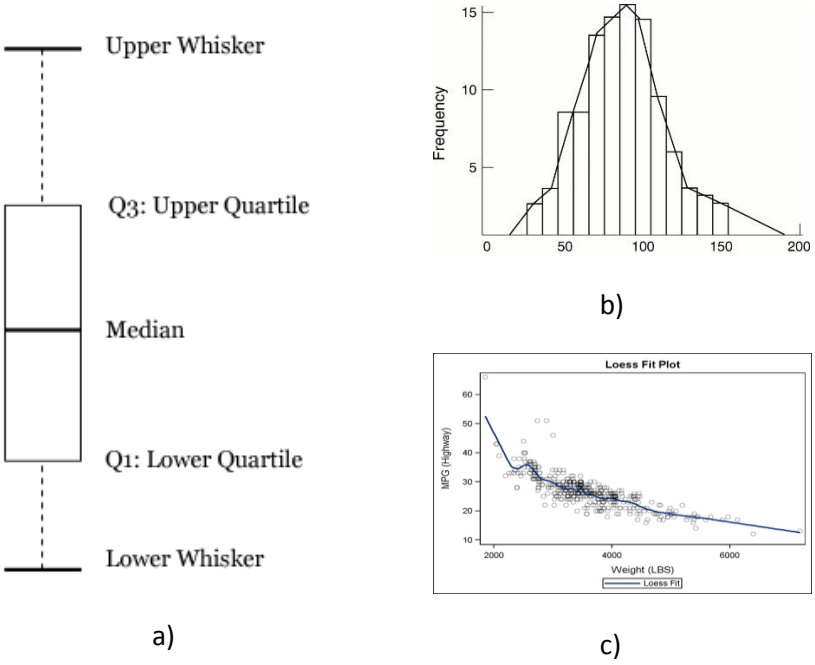
Diskriminácia dát je porovnanie charakteristík cieľovej triedy s charakteristikami kontrastných tried. Napríklad porovnanie výrobkov, ktorých predaj za posledný rok sa zvýšil o 10% s výrobkami, ktorých predaj za posledný rok sa znížil o 30%. Výsledný opis môže byť vyjadrený pomocou diskriminujúcich pravidiel, napr. 80 % výrobkov so zníženým predajom stúpla za posledný rok cena.

Základnou metódou charakteristiky dát je *deskriptívna štatistika*, čo je jednoduché zhrnutie na základe štatistických metrík a grafov. Tieto metriky opisujú rozloženie dát (pozri tabuľku 1). Metriky môžu byť buď holistické (je ich možné vypočítať len nad celým súborom dát, napr. medián, modus) alebo distributívne (výpočet možno distribuovať na podmnožiny dát a konečný výsledok získať z čiastkových výpočtov, napr. minimum, maximum, suma alebo počet). Algebrické metriky sú vypočítané pomocou algebrickej funkcie aplikovanej na jednu alebo viac distribuovaných metrík. Sú výpočtovo menej náročné. Ich výhodou je škálovateľnosť výpočtu v prípade veľkého objemu dát. Nevýhodou je citlivosť na extrémy v dátach (angl. *outliers*).

Tab. 1. Prehľad mier polohy a variability deskriptívnej štatistiky. Modrou farbou algebrické metriky, červenou holistické. (*n* je počet prvkov v dátach, n_i je absolútna početnosť i-teho prvku v dátach, Q_1 a Q_3 sú prvý a tretí kvartil – hodnota nachádzajúca sa za štvrtinou, resp. tromi štvrtinami zoradených hodnôt).

	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	rozptyl	$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i$	štandardná odchýlka	$\sigma = \sqrt{\sigma^2}$
	$\bar{x} = \frac{1}{2(x_{max} + x_{min})}$	variačné rozpätie	$R = x_{max} - x_{\min}$
		variačný koeficient	$v_x = \frac{\sigma}{x} \cdot 100\%$
	najčastejšia hodnota	medzikvartilové rozpätie	$IQR = Q_3 - Q_1$
	prostredná hodnota zoradených hodnôt		

Na grafický opis dát sa využíva škatuľový graf (angl. *boxplot*). Škatuľový graf (pozri obrázok 17) sa nazýva aj 5-číselná sumarizácia. Zobrazuje minimum, dolný kvartil, medián, horný kvartil a maximum. Hranice „škatule“ na grafe sú prvý a tretí kvartil. Vnútri sa nachádza medián vyznačený čiarou a zvyčajne vonku na predĺženej osi, tzv. „fúzoch“ sú minimum a maximum. Výška škatule zobrazuje metriku medzikvartilové rozpätie. Je pravidlo, že maximálna dĺžka „fúzov“ je jeden a pol násobok tohto rozpätia. Ak sa nachádzajú hodnoty za týmto rozpätím, sú označené ako extrémy (angl. *outliers*). Pre grafické zobrazenie rozloženia dát sa okrem škatuľového grafu štandardne využívajú histogram početnosti, polygón početnosti, kvantilový graf, Q-Q graf, diagram rozptýlenia a loess (skr. *local regression*) krivka, ktorá zobrazí trend, ktorý formujú dáta v diagrame rozptýlenia. Zhrnutie a opis dát pomáha objaviť šum a extrémy v dátach, a preto je používaný aj pri čistení dát.



Obr. 17. Príklady grafickej sumarizácie dát – a) škatuľový graf,
b) histogram a polygónom početnosti, c) loess krivka v diagrame rozptýlenia.

Druhou metódou charakteristiky dát je OLAP analýza (angl. *on-line analytical processing*). Vykonáva sa na dátach uložených v dátovom sklade, kde bývajú organizované podľa toho, čo opisujú (napr. zákazník, tovar). Poskytujú informácie z historického hľadiska a zvyčajne sú už sumarizované, t. j. poskytujú prehľady za určité obdobie. Dátový sklad je modelovaný tzv. multidimenzionálnou databázou. Každá dimenzia zodpovedá jednému alebo viac atribútom a každá bunka uchováva sumárnu hodnotu (napr. celkový zisk z predaja počítačov v New Yorku za prvý štvrťrok, pričom dimenzie predstavujú atribúty čas, miesto a typ výrobku). Fyzicky môže byť takáto databáza reprezentovaná relačnou databázou alebo multidimenziálnou dátovou kockou. Dátová kocka umožňuje čiastkové výpočty a rýchly prístup k sumarizovaným dátam. OLAP operácie využívajú znalosti o doméne, z ktorej pochádzajú dáta a na základe týchto vedomostí umožňujú prezentovať dáta na rôznych úrovniach abstrakcie. OLAP analýza sa vykonáva pomocou dvoch základných operácií na kocke:

- priblíženie a oddialenie (angl. *drill-down* a *roll-up*) – zmena úrovne abstrakcie jednej alebo viac dimenzií alebo pridanie/odobranie dimenzie, napr. zobraziť zisk nie za štvrťroky, ale mesiace; zobraziť zisky z predajní nie na základe miest, ale štátov, v ktorých sa nachádzajú,
- vykrojenie (angl. *slice* a *dice*) – vyrezať časť z kocky podľa hodnoty jednej (*slice*) alebo viac (*dice*) dimenzií, napr. zisk za prvý štvrťrok, pričom ostatné dimenzie (miesto, typ výrobku) ostávajú zachované.

OLAP operácie umožňujú aj rotovať osami kockami, transformovať 3D kocku na sériu 2D rovín, zoraďovať dáta, počítať sumy, priemery a pod. Pomocou OLAP môže používateľ

manipulovať, analyzovať dáta a vytvárať ich sumarizácie. Je to silný analytický nástroj, ktorý poskytuje štatistickú analýzu dát, ale aj analýzu vzorov v dátach a počítanie predpovedí.

Treťou metódou charakteristiky dát je AOI (angl. *attribute-oriented induction*). Je to technika, ktorá umožňuje automaticky vyberať atribúty a ich úroveň abstrakcie tak, aby čo najlepšie charakterizovali vybratú triedu – množinu dát. Vhodne vybraté atribúty a dáta vedú k lepším výsledkom metód pre dolovanie v dátach. AOI sa skladá z dvoch hlavných operácií:

- *odstránenie atribútu* – deje sa, ak atribút nadobúda veľa rozličných hodnôt a nedá sa zovšeobecniť alebo sa dá vyjadriť pomocou iných atribútov,
- *zovšeobecnenie atribútu* – deje sa, ak atribút nadobúda veľa rozličných hodnôt a existuje zovšeobecnenie atribútu, napr. v hierarchii konceptov ako ulica – mesto – štát., alebo rozdelením číselného atribútu na intervaly a pod.

Aby neprišlo k prílišnému zovšeobecneniu alebo naopak, je možné kontrolovať úroveň abstrakcie. Dva najbežnejšie používané spôsoby sú *kontrola prahu zovšeobecnenia atribútu* a *kontrola prahu zovšeobecnenia triedy*. Prvý spôsob stanovuje prah – maximálny počet hodnôt, ktoré môže nadobúdať atribút. Prah je stanovený buď individuálne pre atribúty alebo rovnaký pre všetky atribúty. Ak je prekročený nastáva odstránenie alebo zovšeobecnenie atribútu. Druhý spôsob stanovuje prah – maximálny počet záznamov, ktoré môže obsahovať naša trieda – množina dát. Ak je množina príliš veľká, zovšeobecnením niektorého z atribútov sa zoskupí viacero záznamov do jedného identického. Získaná dátová množina sa dá prezentovať graficky alebo ako pravidlá, ktoré poskytnú používateľovi sumarizáciu cieľovej triedy.

4.1.2 Detekcia extrémov

Je to ďalšia deskriptívna úloha dolovania v dátach. V rámci predspracovania a čistenia dát sa z dát odstraňujú extrémy (angl. *outliers*), aby boli dosiahnuté lepšie výsledky. Táto úloha sa však zameriava práve na vyhľadávanie takýchto nezvyčajných objektov v dátach, pretože pre niektoré oblasti môžu byť zaujímavé. Nazýva sa aj dolovanie extrémov (angl. *outlier mining*). Používajú sa naň prevažne štatistické testy, ktoré testujú, či objekt patrí do pravdepodobnostného modelu dát; miery vzdialenosti, pomocou ktorých sa identifikujú objekty príliš vzdialené od zhluku; a miery odchýlok, ktoré merajú nezvyčajné odchýlky v atribútoch.

Detekcia extrémov sa využíva v bankovníctve pre odhalenie podvodov, napr. odhaľovanie nezvyčajných transakcií a pohybov na účte pri odcudzení kreditnej alebo debetnej karty. V medicíne pomáha odhaľovať nezvyčajné výsledky testov, ktoré môžu predstavovať symptómy rôznych ochorení u pacientov. Ďalším využitím je sledovanie abnormálneho šírenia rôznych chorôb a epidémií kvôli zabezpečeniu očkovacích programov. V športe sa sledujú nezvyčajné športové výkony alebo vlastnosti športovcov. Najčastejším využitím dolovania extrémov je detekcia chýb v meraní a čistenie dát.

4.1.3 Objavovanie častých vzorov, asociácií a korelácií

Veľmi zaujímavou úlohou je skúmanie, aké prvky sa spolu často nachádzajú v dátach, napr. v nákupnom košíku zákazníkov, alebo aké sekvencie sa často opakujú, napr. aké kroky robí používateľ frekventovane v aplikácii. Objavovanie takýchto vzorov je ďalšia deskriptívna úloha dolovania v dátach. Vzory môžu byť buď skupiny prvkov, ktoré sa často spolu vyskytujú (angl. *frequent itemset*), sekvencie prvkov, ktoré sa často opakujú (angl. *sequential pattern*), alebo subštruktúry v prípade, že dáta sú vo forme grafu, stromu či mriežky (angl. *structured pattern*). Objavovaním vzorov je možné vydolovať zaujímavé súvislosti a asociácie medzi dátovými položkami.

Typickým príkladom dolovania frekventovaných skupín prvkov je *analýza nákupného košíka*, ktorej výsledky môžu byť využité v reklame a marketingu alebo v spôsobe umiestňovania produktov v obchode. Získanie asociácií medzi produktmi je možné pomocou techniky asociačných pravidiel.

Technika sa skladá z dvoch krokov:

1. nájdenie všetkých spoločných výskytov produktov v košíku,
2. nájdenie všetkých asociačných pravidiel.

Výskyt dvoch produktov v košíku je označený za asociáciu len ak sa produkty spolu vyskytujú spolu nad stanovenú minimálnu hranicu počtu výskytov. Frekventované výskyty produktov sú zvyčajne nájdené pomocou Apriori algoritmu alebo jednej z jeho modifikácií. Každá nájdená asociácia, resp. spoločný výskyt produktov v košíku, má určitú podporu (angl. *support*) a spoľahlivosť (angl. *confidence*), ktorými je vyjadrená užitočnosť a určitosť objaveného vzoru. Ak sú väčšie ako stanovené minimálne hodnoty, odvodí sa asociačné pravidlo v tvare $A \Rightarrow B$ (napr. *mlieko* $\Rightarrow$ *vajíčka*, teda ak si niekto kúpi mlieko, kúpi si aj vajíčka). Podpora vyjadruje podiel záznamov, v ktorých sa prvky A aj B nachádzajú spolu – vyhovujú pravidlu, zo všetkých záznamov. Spoľahlivosť vyjadruje podiel záznamov, ktoré vyhovujú pravidlu z tých, na ktoré sa dá pravidlo aplikovať, t. j. nachádza sa v nich len prvok A. Určuje presnosť pravidla.

Technika asociačných pravidiel môže byť rozšírená o ďalšie aspekty. V prípade, že pri dolovaní je zohľadnená aj hierarchia konceptov, z ktorej pochádzajú prvky pravidla (A,B), môžeme dolovať tzv. viacúrovňové asociačné pravidlá (angl. *multilevel association rules*). Môžu byť tak objavené asociácie medzi abstraktnejšími konceptmi namiesto konkrétnymi produktmi (napr. kto si kúpi fotoaparát, kúpi si aj pamäťovú kartu bez ohľadu na značku a typ). Druhým aspektom, ktorý možno pridať do dolovania pravidiel je ďalšia dimenzia, pretože doteraz spomínané pravidlá obsahovali len jednu dimenziu a mohli byť vyjadrené aj vo forme predikátov, napr.

$$kúpi(X, fotoaparát) \Rightarrow kúpi(X, pamäťová\ karta).$$

Pridaním ďalšej dimenzie, napr. vek nakupujúceho, možno dolovať tzv. viacdimenzionálne asociačné pravidlá, napr. *vek(X, 20-25)* $\wedge$ *kúpi(X, fotoaparát)* $\Rightarrow$ *kúpi(X, pamäťová karta)*, teda ak si 20-25 ročný človek kúpi fotoaparát, kúpi si aj pamäťovú kartu.

Pre dolovanie frekventovaných sekvencií sa používajú napr. AprioriAll algoritmus alebo GSP algoritmus. Na dolovanie subštruktúr sú využívané grafové algoritmy.

V praxi je objavovanie častých vzorov a asociácií využité napríklad pri vystavovaní produktov v obchodoch. Presunutie produktov, ktoré sú často kupované spolu, bližšie k sebe dokáže ešte viac zvýšiť ich predaj. Najznámejším príkladom je asociácia pivo a plienky z roku 1992, ktorá odhalila, že mladí otcovia, ktorí večer ostávali doma s deťmi, kupovali tieto dve položky spolu [108]. Amazon využíva tieto asociácie pre odporúčanie výrobkov a pri zobrazenom výrobku uvádza sekciu s výrobkami nedávno zakúpenými spolu so zobrazenou položkou. Sledovaním frekvencie predaja tovaru dokáže obchodník predpovedať, ktorými výrobkami má zásobovať obchod a ako efektívne využiť jeho obmedzený skladovací priestor. Objavené asociácie sú užitočné aj pri vymýšľaní rôznych marketingových kampaní, napr. zľava pri nakúpe viacerých produktov spolu, prípadne individuálne prispôsobovanie zľavových kupónov.

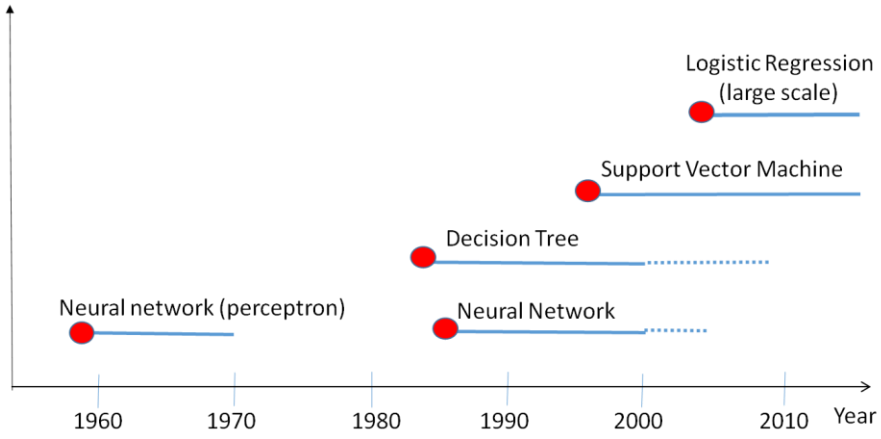
Ďalším využitím je analýza logov webového sídla, z ktorej možno odhaliť v akom poradí prechádza používateľ stránkami webového sídla a vytvoriť model klikania používateľa. Takéto informácie pomáhajú napríklad pri zlepšení prepojení stránok. Sledovaním činnosti prístrojov je možné odhaliť sekvencie udalostí, ktoré predchádzali pokazeniu prístroja a identifikovať tak príčiny alebo symptómy. V poisťovníctve je možné takto analyzovať poistné udalosti, určiť najčastejšie zdroje nehôd a na základe toho prispôsobovať ceny poistenia.

4.1.4 Predikcia a zhlukovanie

Obe tieto techniky pochádzajú z oblasti strojového učenia, kde sa nazývajú *učenie s učiteľom* (predikcia) a *učenie bez učiteľa* (zhlukovanie). Strojové učenie je súčasť oblasti umelá inteligencia. Zaoberá sa vývojom metód a systémov, ktoré sa učia z dát. Pri učení s učiteľom sú k dispozícii tzv. trénovacie dáta, z ktorých sa vytváraný model naučí určité znalosti a pravidlá, ktoré potom použije na ozajstných dátach. Patria sem metódy ako neurónové siete, rozhodovacie stromy, regresia, podporné vektory (angl. *support vector machines*, skr. *SVM*). Opakom je učenie bez učiteľa, kedy nie sú k dispozícii žiadne trénovacie dáta a model sa vytvára priamo na ozajstných dátach. Tento model slúži na opis dát.

Metódy strojového učenia sa vyvíjali spolu s umelou inteligenciou (pozri obrázok 18). V roku 1957 vznikli neurónové siete, konkrétne model perceptrónu [113]. Po zistení, že perceptrón nedokáže vyjadriť zložité funkcie, výskum v oblasti v neurónových sietí poľavil až do vzniku komplexných neurónových sietí v osemdesiatych rokoch. Neurónové siete našli uplatnenie v mnohých industriálnych odvetviach. Zhruba v tej dobe vznikol aj model rozhodovacích stromov, ktorý sa stal veľmi populárny, pretože je ľahko pochopiteľný a vysvetliteľný. Okolo

roku 1995 vznikol model podporných vektorov, ktorý je obľúbený dodnes. Od roku 2003 doteraz sa výskum orientuje na využitie metódy logistickej regresie, ktorá je vo veľkom využívaná na predikciu.



Obr. 18. Vývoj metód strojového učenia [58].

Súčasne s týmito metódami vznikali aj iné, ktoré ale nie sú dnes už tak populárne a využívané, napr. naivný Bayesov klasifikátor, Bayesovské siete, klasifikátor maximálne entropie. Mnohé ďalšie vznikajú aj v súčasnosti. Vývoj v oblasti strojového učenia je pravidelne predmetom konferencie ICML (International Conference on Machine Learning)[20].

4.1.5 Predikcia

Cieľom tejto prediktívnej úlohy dolovania v dátach je zostrojenie modelu, pomocou ktorého je možné budúce hodnoty atribútu na základe hodnôt ostatných atribútov. V závislosti od toho, či sa predpovedá nominálny alebo spojitý atribút, je aplikovaná buď technika *klasifikácie* alebo *regresie*.

Klasifikácia je technika, pomocou ktorej je zostrojený model, ktorý dokáže rozoznať triedu objektu. Dokáže klasifikovať objekty, ktorých trieda nie je známa na základe analýzy objektov z testovacej množiny, ktoré už majú priradenú triedu. Výsledný model je vyjadrený v tvare klasifikačných pravidiel, rozhodovacieho stromu, matematických formúl alebo neurónovej siete. Rozhodovacie stromy aj neurónové siete našli uplatnenie v bankovníctve a finančníctve, napr. schvaľovanie úverov, odhaľovanie podvodov, správa portfólií. Takisto sa využívajú v rôznych priemyselných procesoch a pri automatizovaní pošty (rozoznávanie adries). Pre klasifikáciu elektronickej pošty na spam a želané správy sa s obľubou využíva model podporných vektorov. V poslednom desaťročí sa stala významnou metódou klasifikácie logistická regresia, ktorou sa modeluje klikanie používateľov internetu. Model predpovedá, či používateľ klikne na určitú položku výsledkov vyhľadávania na stránke alebo nie. Vstupmi modelu sú informácie o používateľovi, obsah stránky, slovo, ktoré používateľ vyhľadával, atď. Tento model sa využíva pri poskytovaní výsledkov vyhľadávania

[20] http://icml.cc/

a umiestňovania reklám v rôznych veľkých internetových spoločnostiach, napr. Ebay, Amazon [112].

Regresia je štatistická technika pre dopĺňanie chýbajúcich hodnôt numerického spojitého atribútu alebo pre predpovedanie jeho budúcich hodnôt na základe dostupných dát. Modeluje vzťah medzi nezávislými známymi atribútmi a predpovedaným atribútom. Najpoužívanejšia je lineárna regresia (pozri rovnicu 1).

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1,2,\dots,n \tag{1}$$

Predpovedaný atribút *y* je vyjadrený pomocou nezávislého atribútu *x*, kde ε je náhodná chyba a *β* sú parametre modelu, ktoré chceme odhadnúť. Predpokladá sa, že náhodná chyba je tvorená nezávislými a rovnako rozloženými náhodnými premennými. Pre odhad parametrov modelu sa využívajú najčastejšie metóda najmenších štvorcov a metóda maximálnej vierohodnosti. V prípade, že predpovedaný atribút sa nechová lineárne, môže byť použitý iný typ závislosti, ktorý možno ľahko previesť na lineárny. Typy závislostí môžu byť lineárna, polynomická, logaritmická, exponenciálna alebo trigonometrická. V praxi predpovedaný atribút nezávisí len od jedného nezávislého atribútu, ale viacerých, a preto sa využíva viacrozmerná regresia. Okrem lineárnej regresie sa pre predpovedanie spojitého atribútu používajú aj iné metódy, napr. log-lineárne modely, Poissonova regresia, regresné stromy. Regresná analýza sa v praxi používa všade, kde potrebujeme odhadnúť presnú hodnotu, napr. predpovede cien akcií na burze, predpovedanie zamestnanosti / nezamestnanosti, vyčíslenie rizikovosti poskytnutia pôžičky, úveru a pod.

Zostavený model sa vyhodnocuje pomocou metrík na meranie jeho presnosti (angl. *accuracy*) a chyby (angl. *error*). Chyba je doplnkom presnosti. Pre klasifikátor sa dá vytvoriť chybová matica (angl. *confusion matrix*), ktorá zobrazuje počty správne klasifikovaných objektov a počty nesprávne klasifikovaných – zamenených objektov (*false positives* a *false negatives*).

Aby presnosť a chyba zobrazovali skutočné hodnoty, vyhodnocuje sa model na inej množine dát (testovacej množine) ako pomocou ktorej bol vytvorený (trénovacia množina). Na rozdelenie dát do testovacej a trénovacej množiny existuje niekoľko metód. Najpoužívanejšie sú tzv. *holdout* metóda, náhodné vzorkovanie, krížová validácia a *bootstrap* metóda. Pri *holdout* metóde sú dáta rozdelené v pomere jedna ku dvom. Menšia je testovacia množina. Dáta sú do množín vyberané náhodne tak, aby mali triedy približne rovnaké zastúpenie. Náhodné vzorkovanie je *k*-násobné opakovanie *holdout* metódy, pričom výsledná presnosť alebo chyba je priemer iterácií. Krížová validácia je podobná náhodnému vzorkovaniu. Rozdiel je, že dáta sa delia na *n* dielov, pretože prebehne *n* iterácii. Z dát *n-1* dielov je použitých v trénovacej množine a posledný diel ako testovacia množina a pri každej iterácii sa obmieňa diel použitý ako testovacia množina. *Bootstrap* metóda zostavuje trénovaciu množinu, ktorá má takú istú veľkosť ako všetky dáta (veľkosť *n*). Dáta do množiny vyberá náhodným výberom – *n* krát sa náhodne vyberie objekt z dát, pričom objekty

v trénovacej množine sa môžu opakovať. Zvyšné objekty nevybraté do trénovacej množiny idú do testovacej množiny. Do trénovacej množiny sa zvyčajne vyberie 63,2 % objektov zo všetkých dát.

4.1.6 Zhlukovanie

Patrí medzi deskriptívne úlohy dolovania v dátach. Cieľom je, tak ako pri klasifikácii, zatriediť objekty. Na rozdiel od klasifikácie, nie sú známe názvy tried, pretože nie je k dispozícii trénovacia množina dát. Preto sú výsledkom skupiny objektov, tzv. zhluky, z ktorých môžu byť názvy odvodené. Zhluky sú vytvárané na princípe maximalizovania podobnosti objektov v rámci jedného zhluku (angl. *intraclass similarity*) a minimalizovania podobnosti objektov medzi zhlukmi (angl. *interclass similarity*). Zhlukovacie techniky sa rozdeľujú na zhlukovanie:

- *rozdeľovaním* – je dopredu známy počet zhlukov (*k*), v prvom kroku sú dáta do zhlukov rozdelené náhodne a následne každou iteráciou sa objekty preskupujú tak, aby zhluky boli medzi sebou čo najviac odlišné a zoskupovali čo najpodobnejšie objekty, napr. *k-means* algoritmus, kde sú zhluky reprezentované priemernou hodnotou objektov; *k-medoid* algoritmus, kde sú zhluky reprezentované objektom, ktorý je čo najbližšie ku stredu zhluku, t. j. je najmenej odlišný od ostatných objektov v zhluku,
- *hierarchické* – hierarchicky rozdeľuje dáta na zhluky, môže začínať buď zvrchu, t. j. na začiatku existuje jeden zhluk tvorený všetkými dátami, ktorý sa postupne delí, pokým nie je splnená ukončujúca podmienka alebo každý objekt netvorí jeden zhluk; alebo zdola, t. j. na začiatku každý objekt tvorí jeden zhluk, ktoré sa postupne spájajú, pokým nie je splnená ukončujúca podmienka alebo netvoria všetky objekty jeden zhluk; nevýhoda je, že každý krok je nevratný, výhoda výpočtová zložitosť,
- *založené na hustote* – na rozdiel od techník rozdeľovaním, ktoré dokážu nájsť len zhluky guľovitého tvaru, zoskupujú objekty do zhlukov podľa hustoty, t. j. v susedstve objektu sa musí nachádzať nejaký minimálny počet ďalších objektov, aby bol zaradený do zhluku; takýto prístup umožňuje vytvárať zhluky ľubovoľného tvaru,
- *založené na modeli* – snažia sa napasovať dáta na určitý matematický model; sú založené na predpoklade, že dáta sú vygenerované podľa viacerých pravdepodobnostných distribúcií – každý zhluk má inú; príkladmi sú *EM algoritmus*, ktorý je rozšírením *k-means* algoritmu, v iteráciách upravuje počiatočné odhadnuté pravdepodobnostné rozdelenia, aby dáta čo najviac pasovali; a *SOM* technika (angl. *self-organizing feature maps*) a *Kohonenov algoritmus* je takisto rozšírením *k-means*, na ktorého začiatku je pár neurónov, ktoré súťažia o objekty, postupným získavaním objektov – rozrastaním sa vytvárajú mapy, objekt je pridaný k tomu neurónu, ktorého váhový vektor je najbližší k objektu,

- *pre dáta s veľkým počtom dimenzií* – tzv. kliatba dimenzionality sa dá riešiť pomocou *transformácie atribútov*, na čo slúžia techniky ako PCA (angl. *principal component analysis*) alebo SVD (angl. *singular value decomposition*), ktoré znížia počet dimenzií, pričom zachovajú vzdialenosti objektov v dátach; alebo *selekciou atribútov*, ktorá dokáže odstrániť atribúty irelevantné pre dolovanie; na tomto princípe je založená aj technika *zhlukovania podpriestorov* (angl. *subspace clustering*),
- *založené na ohraničeniach* – zhlukuje objekty podľa používateľských ohraničení, napr. veľkosť zhlukov, počet zhlukov, váha jednotlivých atribútov.

Pre overenie výsledku zhlukovania sa využívajú rôzne metriky, napr. Dunnov index, Davies-Bouldinov index, ktoré vyjadrujú vzdialenosti zhlukov a objektov v rámci zhluku. Príkladmi techník pre overenie sú siluetová validačná metóda a porovnanie so zlatým štandardom. V rámci porovnávania so zlatý štandardom sa vyhodnocujú metriky ako purity, normalizovaná vzájomná informácia, rand index a F-metrika. Zhlukovanie sa v praxi využíva napríklad pri objavovaní tematických okruhov v korpuse dokumentov, kedy sa texty zoskupujú do zhlukov podľa ich obsahu. Ďalším využitím zhlukovania je detekcia malware softvéru v počítači alebo objavovanie komunít v sociálnych sieťach.

4.2 Prognózovanie

Zamerali sme na vytváranie predpovedí, preto sme využívali pri analýze dát techniky dolovania v dátach, konkrétne predikciu, pre vytváranie prognóz.

Prognózovanie je oblasť, ktorá sa zaoberá tvorbou prognóz. Existuje veľa definícií, čo je to prognóza. Podľa slovenského prognostika Fedora Gála [39] je prognóza *„podmienená, o vedecké poznatky sa opierajúca výpoveď o budúcnosti nejakého objektu či javu“*.

Ďalšia definícia prognózy je [93]: „Prognóza je pohľad do budúcnosti, je to vedecky zdôvodnené tvrdenie o pravdepodobnom vývoji reálnych alebo neznámych, ale reálne možných faktorov (ekonomických, spoločenských, politických, technologických, prírodných a i.).“

Prognóza býva v hovorovej reči používaná ako synonymum pre slovo predikcia. My rozlišujeme tieto pojmy a pojmom predikcia označujeme techniku dolovania v dátach (strojové učenie s učiteľom), do ktorej spadajú metódy pre predpovedanie hodnôt nominálnych (klasifikácia) alebo spojitých atribútov (regresia).

Prognózovanie úzko súvisí s plánovaním a rozhodovaním. Prognózy ohraničujú možnosti pri rozhodovaní a uľahčujú tak proces rozhodovania a plánovania. Dobrá prognóza by mala byť *včasná*, t. j. mala by mať zvolený časový horizont tak, aby do času, kým sa naplní, bolo možné na ňu reagovať a urobiť potrebné opatrenia. Prognóza by mala byť aj prijateľne *presná*. Mali by sme poznať veľkosť chyby, s akou predpovedá, aby sme s ňou mohli plánovať. Všeobecne krátkodobé prognózy zvyknú byť presnejšie ako dlhodobé. Ďalej by

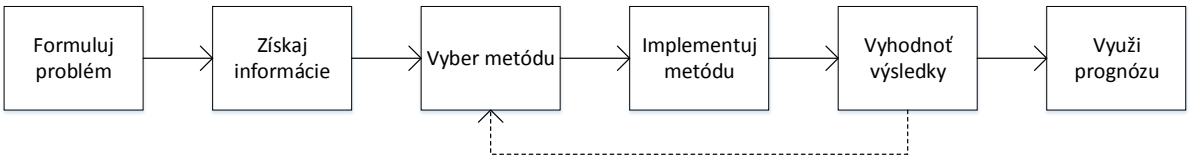
mala byť *spoľahlivá*. To znamená, že má poskytovať konzistentné výsledky a pracovať vždy na rovnakej úrovni presnosti. Na prognózu, ktorá raz predpovedá veľmi dobre, ale inokedy sú jej výsledky príliš nepresné, sa nemôžeme spoliehať. Prognóza by mala predpovedať v *zmysluplných jednotkách* (napr. jednotky objemu, dĺžky, meny), aby sme vedeli vytvárať čo najpresnejšie plány. Aby sa všetci v organizácii riadili rovnakými výsledkami, prognóza by mala byť dobre *zdokumentovaná*. Metóda prognózovania by mala byť *jednoduchá*, ľahko zrozumiteľná a použiteľná. *Náklady* na prognózovanie by nemali prevýšiť zisky, ktoré prognóza prinesie.

Základné oblasti v prognózovaní sú:

- *ekonomické prognózovanie* (angl. *economic forecasting*) – predpovede všeobecných ekonomických faktorov (inflácia, HDP, dane, zamestnanosť),
- *prognózovanie technológií* (angl. *technology forecasting*) – predpovede pravdepodobného možného vývoja v technológiách (napr. počet tranzistorov v čipoch v určitom roku),
- ***prognózovanie dopytu*** (angl. *demand forecasting*) – predpoveď koľko a aký bude dopyt po určitom produkte.

Okrem týchto oblastí sa robia prognózy napr. aj v politike, meteorológii, doprave...

Proces prognózovania je zobrazený na obrázku 19. Na začiatku sa určí účel prognózy a na aké dlhé obdobie chceme vytvoriť prognózu. Ďalšími krokmi sú získanie dát potrebných pre vytvorenie prognózy a výber vhodnej prognostickej metódy.



Obr. 19. Proces prognózovania [6].

Vhodná metóda závisí od zvoleného typu prognózy. Následne implementujeme metódu na dáta a vyhodnotíme presnosť prognózy. Ak prognóza nedáva dostatočne dobré výsledky, opakujeme aplikovanie metódy alebo zvolíme inú metódu. Pre lepšie prognózy sa môžu prognózy z rôznych metód aj kombinovať.

Prognózy delíme podľa časového horizontu na *dlhodobé* (dva a viac rokov), *strednodobé* (tri mesiace až dva roky), ***krátkodobé*** (do troch mesiacov).

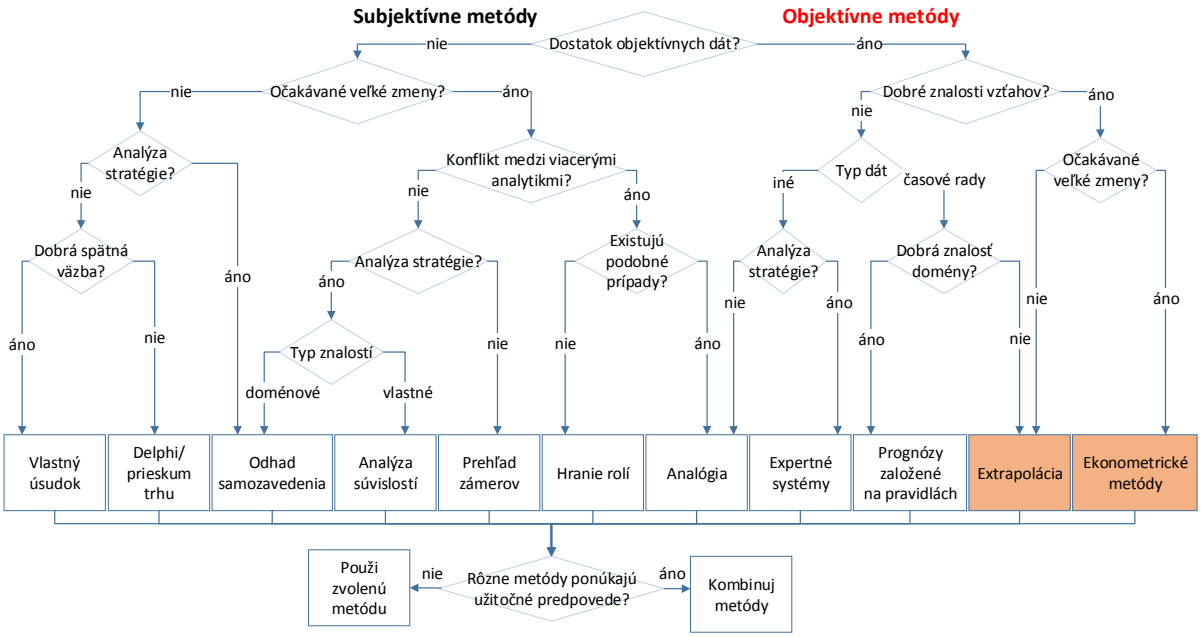
Všetky prognostické metódy poskytujú tzv. *bodovú* prognózu, čo je jedna odhadovaná hodnota na budúce časové obdobie. Tento odhad však skoro nikdy nie je pravdivý. Preto sa robí aj tzv. *intervalová* prognóza, čo je interval spoľahlivosti, v ktorom by sa mala nachádzať ozajstná hodnota. Interval sa určuje pomocou koeficientu spoľahlivosti ($1 - \alpha$). Hranice intervalu sú hodnoty, v ktorých pravdepodobnosť, že skutočná hodnota sa bude nachádzať

medzi nimi je rovná koeficientu spoľahlivosti (pozri rovnicu 2). Pravdepodobnosť má Studentovo rozdelenie a α sa nazýva hladina významnosti alebo riziko odhadu. Interval spoľahlivosti sa volí zvyčajne na úrovni 95% ($\alpha = 0{,}05$).

$$P(q_1 \leq Q \leq q_2) = 1 - \alpha \tag{2}$$

Podľa metódy, ktorou boli prognózy vytvorené ich klasifikujeme na *kvalitatívne* a *kvantitatívne*. Kvalitatívne prognózy sú produktom metód subjektívneho hodnotenia, napr. osobné hodnotenie, panelová zhoda, Delphi metóda. Kvantitatívne prognózy vznikli pomocou tzv. objektívnych metód, ktoré sú založené na matematickom modeli, napr. kauzálny model, model časového radu.

Presnosť prognostických metód závisí od množstva dostupných dát a od zmien v prostredí, v ktorom má prognóza fungovať. V prípade, že máme k dispozícii málo dát a v prostredí sa dejú len malé zmeny, subjektívne metódy prognózovania sú presnejšie, v ostatných prípadoch objektívne (pozri obrázok 20).



Obr. 20. Výber metódy prognózovania podľa Armstronga [6]. Tmavé sú zvýraznené čisto objektívne metódy. Ostatné závisia úplne alebo čiastočne na expertných vedomostiach.

Presnosť metód môžeme aj presne vyčísliť pomocou na to určených metrík. Ak sa rozhodujeme medzi viacerými prognózami, vyberáme väčšinou tú, ktorá má najmenšiu celkovú chybu pre všetky časové obdobia. Chyba prognózy v jednom časovom období sa meria ako rozdiel ozajstnej hodnoty, ktorá nastala y_t a predpovedanej hodnoty $\hat{y}_t$ (pozri rovnicu 3).

$$e_t = y_t - \hat{y}_t \tag{3}$$

Chyba celkovej prognózy pre všetky dáta (počet *n*) sa zvyčajne vyjadruje jednou z týchto metrík:

- stredná absolútna odchýlka (angl. *mean absolute error*, skr. *MAE*, pozri rovnicu 4),

$$MAE = \frac{\Sigma |e_t|}{n} \tag{4}$$

- stredná štvorcová chyba (angl. *mean square error*, skr. *MSE*, pozri rovnicu 5),

$$MSE = \frac{\Sigma e_t^2}{n-1} \tag{5}$$

- stredná absolútna percentuálna chyba (angl. *mean absolute percentage error*, skr. *MAPE*, pozri rovnicu 6),

$$MAPE = \frac{\Sigma \frac{|e_t|}{y_t} \times 100}{n} \tag{6}$$

- relatívna chyba prognózy nazývaná Theilov koeficient (angl. *Theil index*, pozri rovnicu 7).

$$TI = \frac{\Sigma e_t^2}{\Sigma y_t^2} \tag{7}$$

4.2.1 Subjektívne metódy prognózovania

Subjektívne, alebo kvalitatívne, metódy sú založené na ľudskom úsudku a vlastnom názore. Nezakladajú sa na žiadnych matematických modeloch. Rôzni jedinci môžu navrhnúť rôzne prognózy. Ich výhoda je, že dokážu v predpovedi zakomponovať najaktuálnejšie zmeny v prostredí a interné informácie, ktoré pozná prognostik. Nevýhoda je, že môžu byť zaujaté, a preto menej presné. Nazývajú sa aj odhadové, implicitné, neformálne, skúsenostné, intuitívne alebo pocitové.

Subjektívnych metód existuje veľa (pozri obrázok 20). Hlavní reprezentanti sú osobné hodnotenie, panelová zhoda, metóda Delphi a zákaznícky prieskum.

Osobné hodnotenie predstavuje predpoveď jedinca na základe jeho názoru a skúseností. V praxi sa používa najčastejšie, aj keď vôbec nemusí byť spoľahlivé. Je vhodné v prípadoch, kedy sa predpovedaná skutočnosť týka priamo zamestnancov, ktorí majú najviac informácií o súčasnej situácii, napr. či bude treba budúci mesiac opraviť výrobné zariadenie, či vystačia zásoby tovaru a pod. Problém nastáva, ak majú zamestnanci podstatne rozdielne názory.

Metóda panelovej zhody čiastočne eliminuje osobnostné postoje a predsudky, ale zachováva skúsenosti a vedomosti zamestnancov. Priebeh metódy spočíva v diskusii kolektívu odborníkov, ktorí sa zhodnú na budúcom možnom vývoji. Využíva sa v prípadoch, kedy nie je k dispozícií dostatok údajov, napr. pri zavádzaní nového výrobku na trh. Jej slabina je, že diskusia nie je nikým riadená a nie je zadefinovaný spôsob dosiahnutia zhody.

Tento nedostatok rieši metóda Delphi, ktorá je tiež založená na diskusii kolektívu odborníkov. V tomto prípade, ale členovia kolektívu nekomunikujú spolu naraz, ale svoje názory vyjadrujú pomocou dotazníka. Odpovede sú sumarizované a distribuované naspäť medzi odborníkov. Vtedy má každý člen možnosť revidovať odpovede vzhľadom na kolektív. Postup sa opakuje, pokým nenastane zhoda alebo po stanovený počet iterácií.

Zákaznícke prieskumy sú založené na dotazníku pre zákazníkov, v ktorom odpovedajú na otázky o svojich plánoch nakupovať výrobky. Na základe toho sú založené prognózy. Prieskumy trhu môžu poskytnúť cenné informácie a môže byť použitý s hociktorou z prognostických metód.

4.2.2 Objektívne metódy prognózovania

Objektívne, alebo kvantitatívne, metódy sú založené na matematických modeloch. Ich výhoda je, že sú konzistentné a uvažujú veľa informácií a dát pre jedno časové obdobie. Využívajú presne špecifikované procesy k analýze dát. Ich slabina je, že fungujú len tak dobre, aké dobré sú dáta do nich vstupujú. Závisia veľmi na kvalite dát a často krát ani nemáme k dispozícii merateľné dáta, nad ktorými by vedeli robiť výpočty. Delíme ich na:

- extrapolačné,
- ekonometrické.

Extrapolačné metódy predpokladajú, že štúdiom minulých hodnôt je možné predpovedať budúce hodnoty premennej. To znamená, že predpokladajú konzistentný vývoj z minulosti do budúcnosti. Patria sem metódy pre modelovanie časových radov. Extrémny prípad extrapolácie je naivný model, ktorý predpovedá na ďalšie časové obdobie hodnotu z posledného obdobia. Príkladmi iných metód sú dekompozičné metódy, Box-Jenkinsonova metodológia.

Ekonometrické, alebo kauzálne, metódy skúmajú vzťahy medzi príčinami a dôsledkami. Odvodzujú hodnoty závislej premennej od správania množiny nezávislých premenných, a tak v podstate odhadujú budúce hodnoty na základe udalostí v minulosti. Najpoužívanejšou ekonometrickou metódou je regresná analýza. Vzťah medzi nezávislými premennými je modelovaný pomocou lineárnej regresie.

5 Energetika

Jednou z oblastí, v ktorej sa využíva prognózovanie, a na ktorú sme sa primárne zamerali, je energetika. Jej úloha sa stala ešte dôležitejšou po legislatívnych zmenách, ktoré liberalizovali trh s elektrickou energiou a umožnili vstup konkurencie.

Trh s elektrickou energiou na Slovensku donedávna tvorili tri firmy, ktoré dodávali energiu: Západoslovenská energetika (ZSE), Stredoslovenská energetika (SSE) a Východoslovenská energetika (VSE). Od 1.7.2013 sa museli rozdeliť tieto firmy na dodávateľské a distribučné.

Odoberateľ elektriny v domácnosti získal *právo zvoliť si dodávateľa*. Liberalizácia trhu priniesla so sebou niektoré výhody, napr. väčší výber dodávateľa a viac konkurencie držia ceny na nižšej úrovni, lepšie služby pre odoberateľov, vyššiu bezpečnosť zásobovania.

Pre kontrolu trhu vznikli nezávislé regulačné úrady. Na Slovensku to je ÚRSO (Úrad pre reguláciu sieťových odvetví). Tento úrad určuje ceny za prenos a distribúciu elektriny. Jeho úlohou je aj zabezpečenie transparentných podmienok kontraktov medzi odoberateľmi a dodávateľmi a ochrana odoberateľov pred klamnými praktikami a dezinformovaním.

V súčasnosti sa na trhu s elektrinou podieľajú mnoho subjektov, ktoré môžeme rozdeliť do šiestich skupín podľa ich úlohy (pozri tabuľku 2).

Tab. 2. Účastníci na slovenskom trhu s elektrinou [139].

Štát	Ministerstvo HOSPODÁRSTVA SLOVENSKEJ REPUBLIKY, ÚRSO	MHSR, ÚRSO
Výroba	SLOVENSKÉ ELEKTRÁRNE, Enel, Pow-en, OZE ...a	elektrárne, obnoviteľné zdroje energie
Prenos a zúčtovanie	SEPS, OKTE	SEPS, a. s., OKTE, a. s.
Distribúcia	ZÁPADOSLOVENSKÁ DISTRIBUČNÁ, SSE-DISTRIBÚCIA, VSD lokálni distribútori	Západoslovenská distribučná, a. s., SSE-D, a. s., Východoslovenská distribučná, a. s., lokálni distribútori
Predaj	ZSE, SLOVENSKÉ ELEKTRÁRNE, Enel, GGE, STREDOSLOVENSKÁ ENERGETIKA, VE, energie2, magna ...a iní	ZSE Energia, a. s., Stredoslovenská energetika, a. s., Východoslovenská energetika, a. s., ČEZ Slovensko, s. r. o., Energie2, a. s., GGE, a. s., MAGNA ENERGIA, a. s. ...
Zákazníci	veľkí firemní zákazníci, stredne veľkí firemní zákazníci, domácnosti (podnikatelia)	veľmi vysoké napätie (VVN), vysoké napätie (VN), nízke napätie (NN)

Koncoví zákazníci majú zmluvu s dodávateľom elektriny a majú nainštalovaný merač elektrickej energie, pomocou ktorého dodávateľ monitoruje ich spotrebu. Zákazník nemusí byť vždy iba odoberateľ elektriny, ale môže ju do siete aj dodávať, napr. ak využíva obnoviteľné zdroje energie, akými sú solárne panely. Preto sa zákazníci v energetickom žargóne nazývajú *odberno-odovzdávacie miesta (skr. OOM)*.

Pre stabilné fungovanie elektrickej siete musí byť objem spotrebovanej elektriny a vyrobenej elektriny rovnaký. V skutočnosti vzniká na jednotlivých odberno-odovzdávacích miestach rozdiel medzi objednanou a skutočne spotrebovanou elektrinou. Tento rozdiel sa nazýva

odchýlka. Vo väčšine prípadov zodpovednosť za odchýlku nenesú zákazníci, ale preberá ju dodávateľ energie.

Odberno-odovzdávacie miesta, za ktorých odchýlky nesie zodpovednosť jeden dodávateľ elektriny sa súhrne nazývajú *bilančná skupina*. Dodávateľ je povinný zabezpečiť dodávky elektriny pre svoju bilančnú skupinu podľa uzavretých zmluvných podmienok so zákazníkmi. Zároveň ručí za to, že všetku elektrinu, ktorú objedná u výrobcov, jeho bilančná skupina spotrebuje.

Elektrická energia nie je komodita, ktorú je možné skladovať. Pre udržiavanie stabilného napätia v sieti sa využívajú tzv. regulátory. V praxi sú to prečerpávacie vodné elektrárne, ktoré majú schopnosť v čase prebytku elektrickej energie v sieti (napr. v noci) ju akumulovať v podobe vody v ich horných nádržiach. V čase nedostatku dokážu túto vodu prečerpať a vyrobiť chýbajúcu elektrickú energiu. Na Slovensku dohliada na spoľahlivosť a bezpečnosť dodávok elektrickej energie ÚRSO. Tento úrad má právomoci finančne sankcionovať dodávateľov elektrickej energie, ktorí nedodržiavajú regulatórne podmienky (t. j. spôsobia vysokú odchýlku v sieti) a je potrebné kvôli nim spúšťať regulátory.

Preto je dôležité pre dodávateľa elektriny dopredu poznať dopyt svojej bilančnej skupiny. Vytváraním presných prognóz o dopyte svojej bilančnej skupiny je dodávateľ schopný minimalizovať svoju odchýlku v sieti, výhodnejšie obchodovať s elektrinou na trhu a zvýšiť svoju konkurencieschopnosť. Pre vytváranie prognóz má dodávateľ k dispozícii údaje o spotrebe svojich zákazníkov z elektromerov nainštalovaných v odberno-odovzdávacích miestach.

Klasicky sa údaje z elektromerov zbierajú raz ročne (robí sa tzv. fyzický odpočet, sú to nepriebehové merania). Z týchto údajov nie je možné zistiť aká bola denná odchýlka na jednotlivých odberno-odovzdávacích miestach. Preto sa využíva metóda typových diagramov. *Typový diagram odberu* (skr. *TDO*) obsahuje relatívne priemerné hodinové hodnoty odberu. Existuje sedem typových diagramov popisujúcich ročnú spotrebu v rôznych zariadeniach (rodinné domy, bytovky, podniky, závody a pod.). Pre vyčíslenie dennej odchýlky sa nameraná ročná spotreba prepočítava podľa tohto diagramu.

Zákazník si v súčasnosti elektrickú energiu mesačne predpláca. Výška splátky sa určuje podľa jeho spotreby elektriny v minulom roku zistenej fyzickým odpočtom. Na konci roka sa na základe odpočtu ráta aj preplatok alebo nedoplatok za skutočne spotrebovanú elektrinu.

V záujme posilnenia práv a ochrany spotrebiteľa na trhu s elektrinou vydal Európsky parlament a Rada smernicu č. 2009/72/ES o spoločných pravidlách pre vnútorný trh s elektrinou. Táto smernica má podporiť aktívnu účasť všetkých účastníkov trhu a najmä zákazníkov. Ukladá všetkým členským štátom EÚ povinnosť zaviesť *inteligentné meracie systémy* do roku 2020. V slovenskej legislatíve je táto povinnosť ukotvená v zákone č. 251/2012 o energetike. Prevádzkovatelia distribučných sietí sú povinní inštalovať inteligentné merače podľa podmienok vo vyhláške MHSR č. 358/2013.

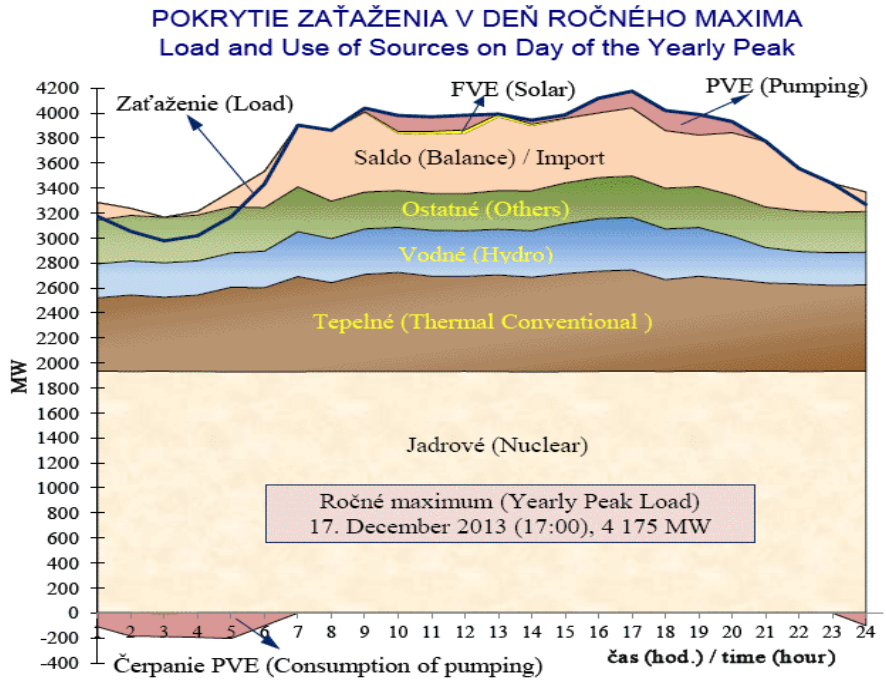
Inteligentné merače dokážu zaznamenávať údaje o spotrebe elektriny každých 15 minút. Presné informácie o spotrebe prinášajú mnoho výhod [139]:

- zníženie spotreby elektriny zákazníkom sledovaním vlastných údajov,
- *presnejšie plánovanie distribúcie elektriny*,
- presnejšie vyhodnocovanie výpadkov siete, čiernych odberov,
- fakturácia na základe presnej spotreby bez odhadov,
- ponuka produktov pre zákazníkov „šitých na mieru“,
- spresnenie výpočtu odchýlok siete.

Do roku 2020 by malo byť v Slovenskej republike inteligentnými meračmi vybavených 80 % odberno-odovzdávacích miest pripojených na regionálnu alebo miestnu distribučnú sústavu na napäťovej úrovni NN (nízke napätie) so spotrebou rovnou alebo väčšou ako 4MWh za rok. To predstavuje približne 600 tis. zákazníkov (podľa vyhlášky č. 358/2013 MHSR) zo všetkých na napäťovej úrovni NN – 2,38 mil. [139]. Pre dodávateľa energie to znamená príliv množstva údajov každú štvrťhodinu, pomocou ktorých môže predpovedať spotrebu elektriny na ďalšie obdobia. Pre analyzovanie takéhoto objemu údajov v reálnom čase sú potrebné automatizované prostriedky a metódy pre predikciu, ktoré budú zohľadňovať vlastnosti spotreby elektrickej energie a vytvárať čo najspoľahlivejšie krátkodobé prognózy.

5.1 Faktory ovplyvňujúce spotrebu elektrickej energie

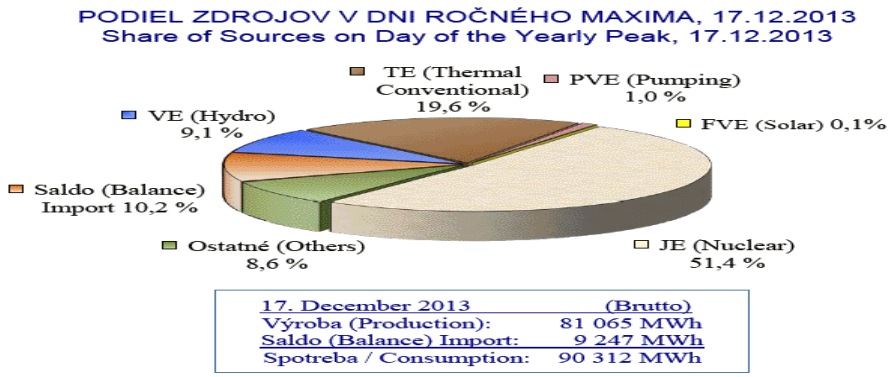
Na výrobe elektrickej energie na Slovensku sa podieľa viacero zdrojov (pozri obrázky 21 a 22). Viac ako polovica elektrickej energie je vyrábaná v jadrových elektrárňach. Jadrové elektrárne dokážu dodávať do siete stály výkon a pokrývajú základné zaťaženie siete. Nedokážu však pružne vykrývať okamžité výkyvy v dopyte. Preto sa využívajú ďalšie zdroje ako vodné a tepelné elektrárne, ktoré pokrývajú pološpičkové zaťaženie siete. Špičkové zaťaženie vykrývajú prečerpávacie vodné elektrárne, pretože turbíny vo vodných elektrárňach sa dajú rýchlo spúšťať a odstaviť, ich výkon sa dá dobre regulovať. Okrem využitia prečerpávacích vodných elektrární sa dajú okamžité výkyvy v dopyte pokryť aj obchodovaním s energiou a jej dovozom zo zahraničia.



Obr. 21. Pokrytie zaťaženia v dni ročného maxima spotreby elektriny (17.12.2013) [116] (PVE – prečerpávacie vodné el., FVE – fotovoltické el.).

Malou časťou sa na výrobe energie podieľajú aj obnoviteľné zdroje energie (OZE), napr. solárne panely, veterné elektrárne. Na Slovensku je budovanie OZE jednoduchšie a lacnejšie vďaka tomu, že dostatočne veľkú časť zaťaženia pokrývajú vodné elektrárne, ktoré dokážu rýchlo kompenzovať nepravidelné dodávky z OZE, ktoré sú ovplyvňované počasím. OZE môžu byť súčasťou odberno-odovzdávacieho miesta, a preto ich treba zohľadňovať pri vytváraní predpovedí spotreby elektrickej energie, najmä pri predpovediach na úrovni jedného odberno-odovzdávacieho miesta.

V súčasnosti na celkovú výrobu/spotrebu elektriny vplývajú OZE (bez veľkých vodných elektrární) len minimálne (pozri obrázok 22), no v budúcnosti sa očakáva nárast ich využívania. Dokument *Cestovná mapa pre obnoviteľnú energiu – COM(2006) 848,* ktorý schválila Európska komisia v roku 2007, navrhuje, aby sa EÚ zaviazala, že do roku 2020 sa budú OZE podieľať na výrobe elektrickej energie vo výške 20 %. Slovensko si v *Národnom akčnom pláne pre energiu z OZE* z roku 2010 stanovilo cieľ zvýšiť podiel OZE na výrobe elektriny z 19,1 % na 24 % do roku 2020. Slovensko využíva z OZE najmä vodnú energiu. Jedným z cieľov uvádzaných legislatívnych opatrení je znížiť produkciu CO_2.



Obr. 22. Podiel zdrojov elektrickej energie na výrobe v dni ročného maxima spotreby (17.12.2013) [116] (VE – vodné el., TE – tepelné el., PVE – prečerpávacie vodné el., FVE – fotovoltické el., JE – jadrové el.).

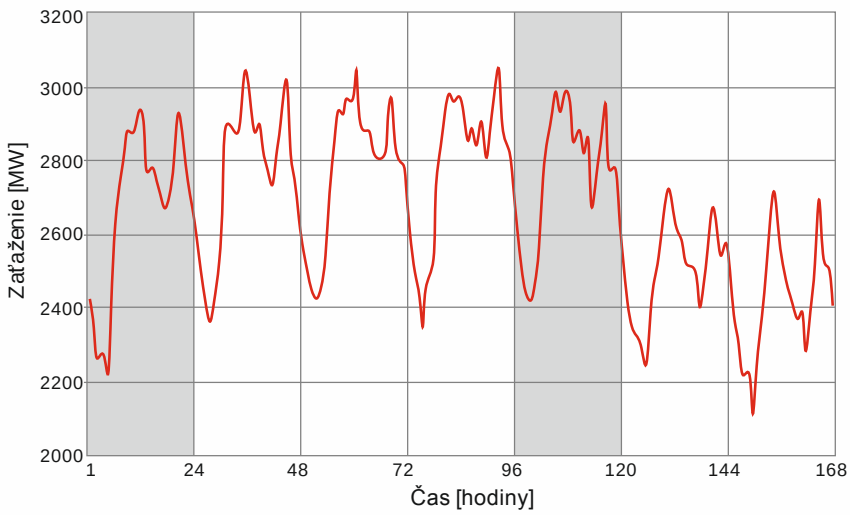
Ďalším faktorom, ktorý (negatívne) vplýva na spotrebu sú tzv. čierne odbery, ktoré sa síce prejavia na celkovej spotrebe, ale nie sú zachytené meračmi, nevystupujú teda v dátach o spotrebe a znižujú presnosť predpovede modelov vytvorených z údajov o meraní.

Najvýznamnejšie sa v údajoch o spotrebe elektrickej energie prejavujú sezónne faktory a to v troch cykloch – *denný*, *týždenný*, *ročný*. Môžeme sledovať, že spotreba sa mení počas dňa a počas noci, je iná počas pracovných dní, víkendov a sviatkov, a mení sa aj vzhľadom na ročné obdobie.

Počas denného cyklu je spotreba v noci nižšia ako cez deň. Ráno postupne stúpa, pred poludním nadobúda prvé maximum, ktoré po popoludňajšom poklese nasleduje večerné maximum. Počas týždenného cyklu sa podobne od pondelka spotreba zvyšuje, v strede týždňa dosahuje maximum a potom postupne klesá až do nedele. Časový rad spotreby elektriny sa často delí na štyri časti podľa spotreby v jednotlivých dňoch týždňa (pozri obrázok 23):

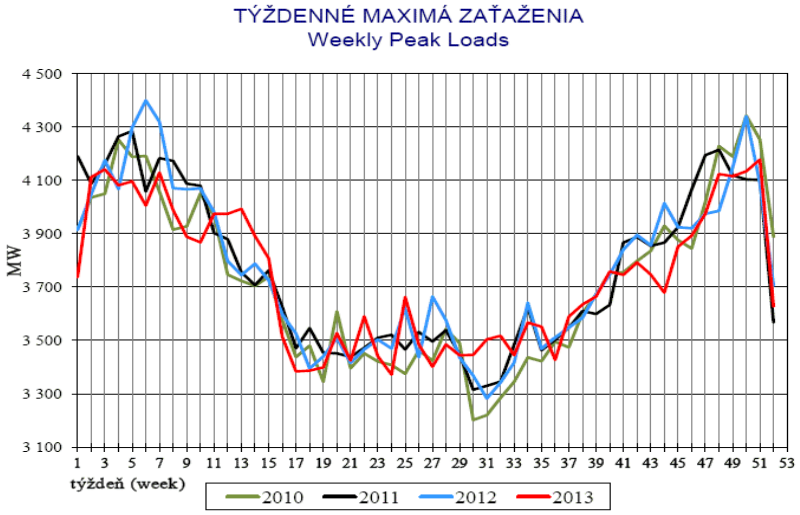
- pondelok – spotreba prudko narastá oproti víkendu,
- utorok až štvrtok – typická spotreba počas pracovného dňa,
- piatok – prudký pokles z pracovného dňa do pokojného víkendu,
- sobota a nedeľa – nízka spotreba oproti pracovnému týždňu.

Podobné denné priebehy spotreby elektriny ako cez víkend sú zaznamenávané aj počas sviatkov alebo prázdnin. Na denný priebeh má vplyv aj posun letného a zimného času.



Obr. 23. Týždenný diagram zaťaženia [92].

Počas ročného cyklu je v našom miernom pásme spotreba elektriny vyššia v zimných mesiacoch kvôli kratšiemu dennému svetlu a vykurovaniu. V letných mesiacoch môžeme sledovať lokálne maximá kvôli zvyšujúcemu sa využívaniu klimatizácií (pozri obrázok 24). Keďže spotreba elektriny silno závisí od dňa v týždni, musí sa pri súhrnných ročných alebo mesačných predpovediach zohľadňovať, akým dňom sa začína dané obdobie a aká je dĺžka roku alebo mesiaca.



Obr. 24. Týždenné maximá zaťaženia [116].

Medzi ostatné faktory, ktoré vplývajú na spotrebu elektrickej energie, patria počasie, spoločenské aktivity a ekonometrické faktory. Prejavy počasia v spotrebe závisia od teploty, vlhkosti, prípadne rýchlosti a smeru vetra. Pri nepriaznivom počasí môže stúpnuť vykurovanie elektrickou energiou, pri vysokých teplotách zase môže narásť využívanie klimatizačných jednotiek. Spoločenské aktivity predstavujú každodenné činnosti ľudí v závislosti od toho, či sa nachádzajú v domácnosti, na pracovisku alebo sa zabávajú vo voľnom čase [56]. Dlhodobo sa môžu v spotrebe prejaviť aj ekonometrické faktory, akými sú

HDP, inflácia či rast populácie [13], no je ich možné pozorovať len ak sú k dispozícii dáta za dlhší časový úsek (niekoľko rokov, desaťročí).

5.2 Monitorovanie spotreby na Slovensku

Na Slovensku a v Českej republike sa inteligentnému meraniu a vytváraniu modelov predpovedí venuje niekoľko firiem. Slovenská firma 2BridgZ[21] sa venuje prediktívnej analytike a ponúka riešenie aj z oblasti energetiky. Ich platforma pre predikciu v reálnom čase je založená na databázovom systéme SAP HANA (High-Performance, Analytic Appliance) od nemeckej softvérovej firmy SAP SE. Je to stĺpcovo orientovaná databáza, ktorá ponúka služby biznis analytiky v reálnom čase (dátový sklad).

Druhá slovenská firma Microstep - HDO s.r.o.[22] je dodávateľom riešení v oblasti energetiky a priemyselnej automobilizácie. Zameriava sa na systémy HDO (skr. hromadné diaľkové ovládanie), monitorovacie a riadiace systémy v priemysle. HDO umožňuje vysielať povely a signály za účelom zapnutia alebo vypnutia spotrebičov (napr. verejné osvetlenie, bojlery, akumulačné kúrenie) a prepínania taríf. Má mnohé funkcie ako inteligentné merače. Microstep ponúka širokú škálu softvérových prostriedkov pre energetické odvetvie vrátane softvéru pre vytváranie krátkodobých a strednodobých prognóz výroby a spotreby elektrickej energie v závislosti od teploty, rýchlosti vetra a slnečného osvitu. Systém Xenergie je univerzálna platforma pre prácu s energetickými dátami. Umožňuje napojenie na iné systémy a ich dáta, napr. prenosové sústavy (SEPS, ČEPS), operátora trhu (OKTE, OTE/CDS), meteorologické ústavy (SHMÚ, ČHMÚ), kurzy bánk. Pre predikciu a modelovanie záťaže siete obsahuje moduly Profile Forecaster, WhiteBox Forecaster a LoadModeler. Pri predikcii zohľadňujú tieto faktory: deň, teplota, osvit, ročné obdobie, dovolenka. Tento systém používajú poprední slovenskí a českí distribútori energie. V súčasnosti sa spoločnosť orientuje aj pre riešenia z oblastí inteligentných sietí.

Okrem softvéru pre obchodníkov s elektrinou sa niekoľko spoločností orientuje na zákazníkov. Zákazník má možnosť už dnes si nainštalovať inteligentný merač a sledovať vlastnú potrebu. Na základe toho mu môže monitorovacia spoločnosť odporučiť najvýhodnejšieho dodávateľa elektriny s vhodným tarifným systémom, ktorý najlepšie (najlacnejšie) pokrýva zákazníkovu spotrebu elektriny. Sledovanie vlastnej spotreby energie môže motivovať zákazníka zmeniť svoje správanie v domácnosti tak, aby výhodnejšie spotrebovával elektrinu. Týka sa to presunu činností ako púšťanie práčky, umývačky, ohrevu vody na časy, kedy je elektrina lacnejšia. Na Slovensku ponúka monitorovanie

[21] http://www.2bridgz.com/
[22] http://www.microstep-hdo.sk/ , http://www.smartgrids.sk/

s inteligentným meračom spoločnosť ESM Yzamer so službou *heat2go*[23], v Českej republike obdobné služby zabezpečuje firma *energomonitor*[24].

V oboch krajinách prebehla aj skúšobná prevádzka inteligentných meračov na vybratých odberno-odovzdávacích miestach. Na Slovensku najväčšia distribučná spoločnosť, Západoslovenská distribučná, realizovala doteraz dva pilotné projekty orientované na testovanie technológií. Vlastný pilotný projekt mal aj distribútor zo skupiny Stredoslovenskej energetiky [79]. Spoločnosť Východoslovenská distribučná zaviedla vlastný pilotný projekt s 300 inteligentnými meračmi v roku 2011. Jeho výsledky pomohli pri tvorbe slovenskej legislatívy. Závery potvrdili, že ročná úspora elektriny nebude na Slovensku vyššia ako 1,5 %. Zároveň sa potvrdilo, že záujem o informácie o spotrebe u zákazníkov s časom klesá a že najlepší spôsob ich poskytovania je prostredníctvom webového portálu [59].

V súčasnosti prebiehal pilotný projekt nasadzovania inteligentných meračov na Slovensku podľa podmienok vo vyhláške č. 358/2013 MHSR, ktorý trval do februára roku 2015. Skúšobná prevádzka trvala jeden rok a slúžila na zefektívnenie prevádzky a odskúšanie technológií distribútormi. Zároveň umožnila výber štandardných technológií, ktoré budú efektívne a neprekročia predpokladané náklady. Merače boli v tomto období u 1 % koncových odoberateľov (cca 6 000 ks). Po vyhodnotení skúšobnej prevádzky by sa malo pokračovať v zavádzaní inteligentných meračov u zvyšku odoberateľov podľa plánu do roku 2020.

Jeden z najväčších pilotných projektov v strednej Európe – WPP AMM (Wide Pilot Project) uskutočnila spoločnosť ČEZ v Českej republike v roku 2012. Merania prebiehali na 35 tis. meračoch. Na základe týchto meraní v spolupráci so spoločnosťou Mycroft Mind[25] a centrom CERIT-SC (Centre of Education and Research in IT – Scientific Cloud) Masarykovej Univerzity v Brne bola vytvorená simulácia 3,5 mil. odberných miest s meračmi. Tá mala overiť, či navrhované riešenie pre zber a ukladanie dát z meračov vydrží nápor pri reálnej prevádzke. Výsledky potvrdili, že áno a ukázali, aké parametre musí spĺňať konečné riešenie. V rámci tohto projektu bol navrhnutý a overený algoritmus pre predikciu spotreby akumulačných spotrebičov a navrhnuté a implementované metódy pre klasifikáciu odberných miest podľa spotreby akumulačných spotrebičov.

Pokroky v oblasti inteligentných sietí a meračov sú už štvrtým rokom diskutované na konferencii *Smart metering*[26], ktorú organizuje portál eFocus pod záštitou MHSR. Jej cieľom je prezentovať aktuálny vývoj v EÚ, praktické skúsenosti z členských štátov a výsledky z pilotných projektov nasadzovania inteligentných meračov. Medzi najvýznamnejšie konferencie na Slovensku, ktoré sa venujú vývojovým trendom v oblasti energetiky patria

[23] http://www.heat2go.sk/
[24] https://www.energomonitor.cz/
[25] http://www.mycroftmind.cz/
[26] http://konferencie.efocus.sk/konferencia/4.-rocnik-odbornej-konferencie-smart-metering-2014-ako-dalej

Energofórum[27], ktoré organizuje spoločnosť *sféra, a. s.*, takisto pod záštitou ministerstva, a Elektroenergetika[28], ktorú organizuje Technická univerzita v Košiciach.

5.3 Úlohy v doméne energetiky

V oblasti energetiky vzhľadom na charakter dát sú typickými úlohami:

- predikcia spotreby elektrickej energie pre distribučné spoločnosti, bilančné či iné skupiny alebo jednotlivcov,
- analýza vzťahu spotreby elektrickej energie a externých faktorov, napr. meteorologických či demografických údajov,
- dolovanie typických motívov a tvarov kriviek odberu elektrickej energie,
- zhlukovanie odberných miest podľa charakteristík odberu a
- klasifikácia odberateľov elektrickej energie alebo aj detekcia anomálií.

V našom výskume sa zamerali najmä na prvú úlohu, ktorá je z pohľadu dodávateľa elektrickej energie najdôležitejšia. Presné prognózy spotreby elektrickej energie sú základom pre výhodné obchodovanie na trhu s elektrinou. Dodávateľ dokáže pomocou nich minimalizovať svoju odchýlku v sieti, teda znížiť regulačné poplatky a včas predať nadbytočnú energiu v sieti.

Počas prác na tejto úlohe sme skúmali aj metódy zhlukovania, ktoré sú použiteľné pre zhlukovanie spotrebiteľov do skupín s podobnou krivkou odberu elektrickej energie. Zhlukovanie odoberateľov do skupín má potenciál pre presnejšie predikcie spotreby elektrickej energie týchto skupín.

Keďže sme využívali dáta z pracovného balíka 6, sú analyzované a opísané v spomenutom balíku. Súčasťou analýzy bolo aj skúmanie vzťahov s meteorologickými a demografickými údajmi. V rámci nášho výskumu sme využívali dáta o spotrebe elektrickej energie zo slovenských inteligentných meračov, z ktorých boli počas predspracovania vybraté odberno-odovzdávacie miesta s úplnými údajmi o spotrebe a boli rozdelené na regióny podľa prvých dvoch číslic poštového smerovacieho čísla.

Úlohám a výzvam v energetickej doméne z pohľadu dátovej analytiky sme sa venovali aj v našich publikáciách uverejnených na fórach zameraných aj na spracovanie a analýzu

[27] http://www.energoforum.sk/
[28] http://ee2013.fei.tuke.sk/

veľkých objemov dát, konkrétne v pracovnej dielni WIKT (Workshop on Intelligent and Knowledge oriented Technologies)[29,30] a na konferencii Data a Znalosti 2015[31,32,33].

6 Prehľad používaných analytických algoritmov a nástrojov pre spracovanie veľkých dát

Dáta o spotrebe elektrickej energie z inteligentných meračov majú tieto základné charakteristiky:

- *objem* – potencionálne nekonečný, lebo nové merania priebežne neustále pribúdajú, a preto ich treba aj priebežne spracovávať,
- *rýchlosť* – nové záznamy zo státisícov odberných miest pribúdajú každých 15 minút, kedy prebehne meranie spotreby,
- *premenlivosť* – spotreba elektrickej energie je ovplyvňovaná nielen časom, ale i mnohými ďalšími faktormi, napr. demografické, ekonomické a pod. Spotreba sa líši pre rôzne skupiny odoberateľov ako aj pre rôzne regióny Slovenska v závislosti na počty domácností/fabrík/obnoviteľných zdrojov atď. Dáta sa vyznačujú silnými dennými a týždennými sezónnymi cyklami a výkyvmi spotreby počas dní ako sviatky alebo závodné dovolenky.

Všetky tieto vlastnosti museli byť zohľadnené pri návrhu metód predikcie spotreby elektriny. Preto sme sa sústredili najmä na metódy pre predikciu *časových radov*, ktoré považujeme za najvhodnejšie pre tento typ dát.

Inkrementálny charakter pribúdania dát a ich potencionálne nekonečný objem si vyžadoval, aby pri ich spracovaní boli zohľadnené ďalšie obmedzenia typické pre spracovávanie prúdových dát, ktoré sa vyznačujú podobnými charakteristikami.

Prúdové dáta sa vyznačujú tromi hlavnými vlastnosťami:

- *objem* – mnohé zariadenia a aplikácie produkujú obrovské množstvá dát, napr. senzory, mobilné zariadenia, bankové transakcie, aktivita používateľov v sieti, na webových stránkach,

---

[29] KOSKOVÁ, G. et al.: Perspektívy modelovania a predikovania veľkých dát v energetike. In: *Proceedings of 9th Workshop on Intelligent and Knowledge Oriented Technologies*, 2014, pp. 57–62.
[30] ROZINAJOVÁ, V. et al.: Otvorené smery výskumu v oblasti dátovej analytiky. In: *Proceedings of 10th Workshop on Intelligent and Knowledge oriented Technologies 2015*, 2015.
[31] VRABLECOVÁ, P.: Inteligentná analýza veľkých objemov dát. In: *Proceedings of 9th Workshop on Intelligent and Knowledge Oriented Technologies*, 2014, pp. 126–129.
[32] LAURINEC, P., LUCKÁ, M.: Analýza vplyvu redukcie dimenzionality na zhlukovanie veľkých dátových množín. In: *Proceedings of Data a Znalosti 2015*, 2015, pp. 1–6.
[33] LÓDERER, M., ROZINAJOVÁ, V., EZZEDDINE, A.B.: Predikcia spotreby elektrickej energie založená na kombinácii predikčných metód. In: *Proceedings of Data a Znalosti 2015*, 2015.

- *rýchlosť* – dáta pribúdajú veľmi rýchlo, je nemožné ich analyzovať tradičným postupom objavovania znalostí a klasickými metódami pre dolovanie v dátach,
- *premenlivosť* – dáta, ale aj prostredie, ktoré ich ovplyvňuje, sa v čase menia, získané znalosti z dát majú časovo obmedzenú platnosť, preto aj modely predpovede sa musia v čase meniť.

Z týchto vlastností vyplývajú ohraničenia, ktorými sú limitované metódy pre analýzu prúdových dát. Ideálne vlastnosti metódy pre dolovanie prúdových dát sú [41]:

- má malú časovú zložitosť pre pravidelné spracovanie dátovej vzorky prúdu,
- používa nemenné množstvo pamäte,
- stačí jej jeden prechod dátami na zostrojenie modelu,
- je schopná sa prispôsobovať zmenám v dátach alebo v prostredí,
- vie zostrojiť skoro taký istý model ako by zostrojila klasická (neprúdová) metóda pre dolovanie v dátach.

Prúd predstavuje sekvenciu dát prichádzajúcu v čase. Na rozdiel od tradičného procesu objavovania znalostí nie je pri zostrojovaní prediktívneho modelu k dispozícii celá trénovacia množina dát, ale tá prichádza postupne v čase. To znamená, že model musí do seba *inkrementálne* zakomponovávať prichádzajúce dáta a predpovedá vždy na základe aktuálnych dát, ktoré dostal [43]. Hlavné rozdiely medzi tradičným dolovaním v databázach a dolovaním v prúdoch sú uvedené v tabuľke 3. Kvôli výpočtovým a pamäťovým obmedzeniam poskytujú metódy pre dolovanie v prúdoch často krát len odhady namiesto presných výsledkov. Vo všeobecnosti platí, čím je menšia chyba výsledku, tým väčšie sú nároky na čas a pamäť. Posledným podstatným rozdielom je, že prúdové dáta sú distribuované, t. j. prichádzajú súčasne z viacerých uzlov nejakej siete (napr. senzorov, mobilných zariadení, počítačov). Preto metódy pre dolovanie v prúdoch musia byť robustné voči komunikačným výpadkom ako sú chýbajúce dáta a oneskorenia.

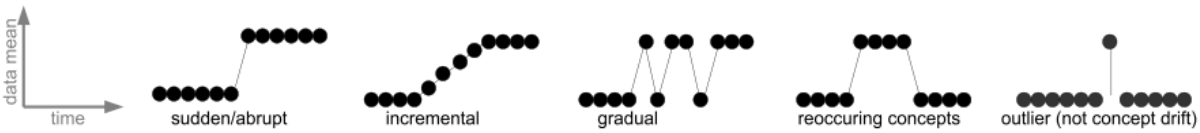
Tab. 3. Hlavné rozdiely medzi dolovaním v databáze a v prúde [38].

	databázy	*dátové prúdy*
prístup k dátam	náhodný	sekvenčný
počet prechodov dátami	mnohonásobný	jeden
čas spracovania	neobmedzený	obmedzený
dostupná pamäť	neobmedzená	obmedzená nemenná
výsledky	presné	približné
distribuované	nie	áno

Prúdové metódy operujú v dynamicky sa meniacom prostredí. V reálnych aplikáciách sa často menia vlastnosti veličiny (konceptu), ktorú sa snažíme predpovedať. Preto sa od metód okrem inkrementálneho spracovania prichádzajúcich dát vyžaduje, aby boli *adaptívne* a vedeli sa v čase prispôsobovať dynamickým zmenám predpovedaného konceptu. Tým pádom poskytovali väčšiu presnosť predpovede. Pri zmene v prúde by sa mali vedieť adaptovať bez nového trénovania modelu predpovede. Zmeny sú zvyčajne nečakané a objavujú sa náhle. Na obrázku 25 sú znázornené vzory typicky sa objavujúcich zmien v prúde. Zmena pozorovanej veličiny (angl. *concept drift*) môže byť [40]:

- *náhla* – napr. ak senzor v sieti je nahradený iným s inou kalibráciou,
- *inkrementálna* – napr. ak sa senzor v sieti postupne kazí a znižuje sa jeho presnosť,
- *pozvoľná* – napr. spotreba elektriny v domácnosti pri prechode na iný tarifný program, kedy sa postupne menia návyky využívania spotrebičov,
- *opakujúca sa* – napr. spotreba elektriny v domácnosti počas sviatkov alebo dovolenky, náhla zmena sa opakuje v čase, ale nie cyklicky ako sezónnosť, nevieme, kedy nastane.

Extrémy v dátach (angl. *outliers*) sú prirodzenou súčasťou prúdových dát a vznikajú pri chybách merania, chýbajúcich dátach a pod. Nie sú však považované za zmenu v prúde, t. j. nie je kvôli nim potrebné meniť model predpovede. Metódy predikcie prúdových dát by mali byť dostatočne robustné voči extrémom v dátach, ale zároveň by sa mali prispôsobovať dynamickým zmenám v prúde. Mali by vedieť správne rozoznávať zmeny od extrémov.



Obr. 25. Zmeny objavujúce sa v prúde (angl. *concept drifts*) [40]. Na x-ovej osi je čas, na y-ovej priemer prichádzajúcich hodnôt. Posledný príklad – extrém v dátach (angl. *outlier*) nie je zmena prúdu.

6.1 Predikcia časových radov

Metódy predikcie spotreby elektrickej energie sú špeciálnym prípadom dolovania v prúdoch. V závislosti od toho, aký typ metódy používajú na *učenie* modelu, sa delia na:

- *regresnú analýzu,*
- *analýzu časových radov,*
- *umelú inteligenciu* (napr. neurónové siete, fuzzy logika, podporné vektory),
- *hybridné metódy* aplikujú kombináciu predchádzajúcich.

6.1.1 Regresné modely

Regresia je jednou z najpoužívanejšou technikou v predikcii. Cieľom regresie je určiť koeficienty funkcie, ktorá najlepšie opisuje vzťah medzi vysvetľovanou/závislou premennou (dependent variable) a jednou alebo viacerými vysvetľujúcimi/nezávislými premennými (indipendent variables). Koeficienty funkcie vyjadrujú citlivosť závislej premennej od zmeny nezávislých premenných. Regresia môže byť definovaná nasledovným vzťahom [5]:

$$y = X\beta + \varepsilon$$

Kde y je n×1 vektor závislých premenných, X je matica n×p nezávislých premenných, β je p×1 vektor koeficientov funkcie a ε je n×1 vektor náhodných chýb.

Hodnoty najlepších koeficientov môžu byť vypočítané pomocou minimalizácie chyby medzi meranými Y a predikovanými hodnotami závislej premennej y (priemerná štvorcová chyba) [100]:

$$\min E[(Y - \hat{y})^2]$$

Po výpočte koeficientov môže byť regresná funkcia použitá na predikciu.

V súčasnosti existuje niekoľko typov regresných modelov, ktoré majú rôzne vlastnosti a používajú sa na nájdenie najlepšej krivky, ktorá opisuje použité dáta s najmenšou chybou:

- Lineárny regresný model – obsahuje iba jednu vysvetľujúcu premennú.
- Viacnásobný regresný model – obsahuje dve a viac vysvetľujúcich premenných, čo umožňuje modelovať zložitejšie vzťahy.
- Robustný lineárny model – by sa mal používať v prípade keď sa v dátach objavujú extrémy (hodnoty, ktoré sa výrazným spôsobom odchyľujú od všeobecného trendu v dátach).
- Polynomiálny regresný model – je model, ktorého nezávislé premenné nadobúdajú mocninu viac ako 1.
- Nelineárny regresný model – funkcia modelu nie je lineárna, preto musí byť výpočet priemernej štvorcovej chyby vykonaný iteratívne.
- Iné.

Najpoužívanejšou štatistickou metódou pre krátkodobé prognózovanie spotreby elektriny je regresná analýza. Patrí do kategórie ekonometrických alebo kauzálnych metód, pretože sa využíva v ekonometrii pre vysvetlenie ekonomických javov. Výsledkom regresnej analýzy je vzťah medzi vysvetľovanou veličinou (v našom prípade spotrebou elektriny) a vysvetľujúcimi veličinami (napr. teplota, ročné obdobie, deň, čas). Keďže spotreba elektriny závisí od mnohých faktorov, využíva sa viacrozmerná regresia. Parametre modelu sú určené zvyčajne

metódou najmenších štvorcov. Pri modelovaní spotreby sa používajú premenné počasia ako teplota a vlhkosť [57, 94, 121], binárne premenné signalizujúce sviatky [103] a oddelenie premenných závislých od počasia [61, 105]. V posledných rokoch sa experimentuje s nelineárnou viacrozmernou regresiou [131], neparametrickou regresiou (jadrové odhady a vyhladzovanie) [8] a pravdepodobnostnou lineárnou regresiou [53]. V [8] bola použitá pri predikcii skupina modelov, každý reprezentoval jeden typ správania sa spotreby.

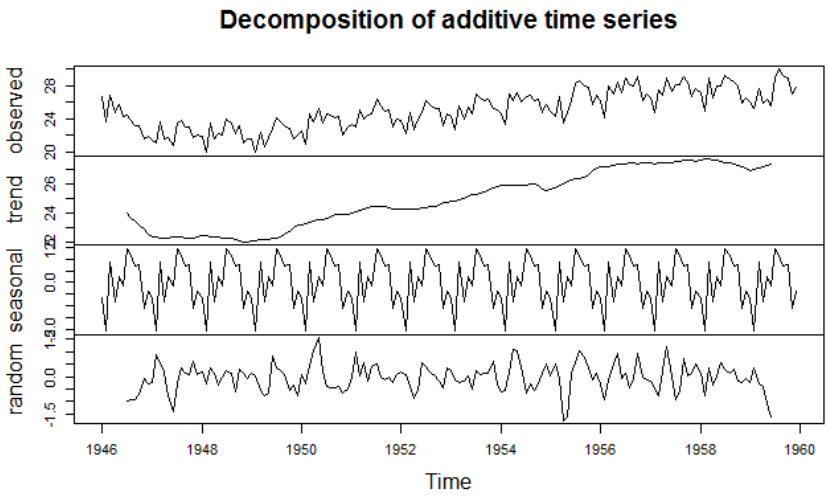
6.1.2 Analýza časových radov

Analýza časových radov[34] pozostáva z viacerých metód skúmajúcich hodnoty radov v snahe získať zmysluplné štatistické informácie a charakteristiky skúmaného radu. Najčastejšie využívaný prístup v analýze časových radov je založený na vyhladzovaní historických dát s cieľom oddeliť základné vzory v dátach od náhodností a šumu. Základné vzory je potom možne použiť na predikovanie budúcich hodnôt. Identifikované vzory môžu byť rozdelené do menších podvzorov, ktoré lepšie opisujú faktory ovplyvňujúce hodnoty v časovom rade. Tento postup sa nazýva dekompozícia časového radu. Dekompozičné metódy sa snažia identifikovať štyri hlavné zložky časového radu [18]:

- Trendová zložka (T) v časových radoch vyjadruje hlavnú dlhodobú tendenciu vo vývoji radu v čase. Ak časový rad neobsahuje trend, tak sa nazýva stacionárny. Naopak, ak obsahuje trendovú zložku, tak sa nazýva nestacionárny.
- Sezónna zložka (S) je opakujúca sa krátkodobá fluktuácia hodnôt pozorovanej premennej s periódou jeden rok alebo kratšou. Sezónna zložka môže byť napríklad denná, týždenná, mesačná, atď.
- Cyklická zložka (C) popisuje strednodobé zmeny spôsobené činiteľmi, ktoré pôsobia určitú dobu a vracajú sa v opakovaných periódach.
- Náhodná zložka (E) je veličina, ktorá sa nedá popísať žiadnou z funkcii času. Príčiny náhodných vplyvov sa označujú ako neočakávané a ťažko predvídateľné. Navyše sa predpokladá ich vzájomnú nezávislosť.

Ukážka dekompozície časového radu:

[34] Časový rad je postupnosť hodnôt sledovanej premennej, meranej v čase, o ktorej sa predpokladá, že má vnútornú štruktúru, ktorú možno odhaliť.



Obr. 26. Dekompozícia časového radu na jednotlivé zložky. Rozložený dátový rad zachytáva počet pôrodov za mesiac v meste New York od roku januára 1946 do decembra 1959[35].

Vytvorenie presnej predpovede si často vyžaduje oddeliť niektorú zo zložiek(T, S, C, E) z časového radu a jej samostatnú predikciu. Na základe vzťahu medzi zložkami radu sa najčastejšie využívajú dva modely:

Aditívny model

$$Y_t = T_t + C_t + S_t + E_t \quad \textbf{for } t = 1, 2, 3, \ldots, T$$

Multiplikatívny model

$$Y_t = T_t * C_t * S_t * E_t \quad \textbf{for } t = 1, 2, 3, \ldots, T$$

Aditívny model sa najčastejšie používa v prípade sú jednotlivé zložky modelu navzájom nezávislé. Naopak, multiplikatívny model je vhodné použiť v prípade, keď existuje medzi zložkami časového radu závislosť, napríklad, rast trendovej zložky pôsobí na rast sezónneho výkyvu.

Metódy analýzy časových radov sú extrapolačné metódy, t. j. predpovedajú budúcnosť len na základe minulosti, nepoužívajú vysvetľujúce premenné ani medzi nimi nehľadajú vzťahy. Najvýznamnejšie sú Box-Jenkinsonova metodológia (ARIMA modely), exponenciálne vyrovnávanie a stavový model / Kalmanov filter.

Box-Jenkinsonova metodológia [19] obsahuje niekoľko modelov – autoregresívny AR(p), model kĺzavých priemerov MA(q), všeobecný model ARMA (p,q) a integrovaný model ARIMA(p,d,q). Najviac využívaný pre modelovanie spotreby je model ARIMA a jeho sezónny variant SARIMA (angl. *seasonal ARIMA*). Model SARIMA sa zapisuje *ARIMA(p,d,q) × (P,D,Q)*$_m$. Nesezónne zložky sú definované v prvej zátvorke a sezónne v druhej. Parameter *m* označuje počet periód v sezóne. Parametre (p, P) určujú počet pozorovaní pre autoregresívny model,

[35] http://robjhyndman.com/tsdldata/data/nybirths.dat

parametre (d, D) označujú stupeň diferencovania potrebný na to, aby bol rad stacionárny a parametre (q, Q) určujú počet posledných pozorovaní pre model kĺzavých priemerov. Model môže byť rozšírený aj o viac sezónnych cyklov. Pre predpoveď spotreby sa uvažovalo veľa sezónnych ARIMA modelov rôznych typov [45, 89, 96, 120, 123]. Niektoré práce uvažujú v modeli aj vplyv počasia [24, 135], ale jednorozmerné metódy sú pokladané za postačujúce pre krátkodobé predpovede, pretože počasie sa v takých krátkych intervaloch mení len postupne a to sa odráža dostatočne v údajoch o spotrebe [127]. Takéto metódy dokážu operovať aj v podmienkach, kde meteorologické dáta nie sú k dispozícii.

Exponenciálne vyrovnávanie je široko využívané [62], pretože je to jednoduchý a robustný prístup. Jeho základné varianty sú – jednoduché, Holtove, Wintersove. Posledný variant uvažuje aj sezónnosť, ktorá je podstatná pri modelovaní spotreby elektriny. Preto bol tento model rozšírený o dvojitú (dennú, týždennú) [125] a trojitú (dennú, týždennú, ročnú) sezónnosť [128]. Model s dvojitým sezónnym cyklom bol nedávno úspešne aplikovaný pre predpoveď elektriny v Brazílii [122]. Model bol doplnený o ošetrenie špeciálnych dní (sviatkov) a vzťah spotreby a počasia. Pravidlá ošetrujúce tieto stavy boli navrhnuté aj v práci [7]. Model obstál najlepšie aj v nedávnom porovnaní modelov exponenciálneho vyrovnávania pre predpovedanie spotreby elektriny v Malajzii [2]. Nedávne práce rozšírili tento prístup o možnosť definovať rôzne denné cykly pre rôzne dni, napr. víkendy, pracovné dni [44] alebo časti dní, napr. noc, doobedie pracovného dňa [124].

Stavový model a Kalmanov filter [70] sú využívané pre zmenšenie chyby predpovede. Dáta spracováva ako prúd a periodicky sa aktualizuje podľa chyby predpovede tak, aby ju minimalizoval. V kroku predikcie vytvára odhady stavu v nasledujúcom období aj s vyčíslenou istotou, v kroku aktualizácie je spozorovaný ozajstný stav, vyhodnotená chyba a aktualizovaný odhad budúceho stavu ako vážený priemer posledných odhadov. Vážený je podľa istôt odhadov. Spotrebu elektrinu monitoruje ako náhodný proces. Pre presné odhady je treba viac ako 3-10 rokov údajov. Nevýhodou je, že táto metóda je citlivá na extrémy a šum v dátach. Kalmanov filter bol úspešne použitý pre predpovedanie spotreby elektriny [96, 105, 130].

6.1.3 Umelá inteligencia

Neurónové siete sú jednou z metód strojového učenia. Populárne pre predpovedanie spotreby elektriny boli najmä v 90-tych rokoch [50]. Boli navrhnuté ako kombinácia metód analýzy časových radov a regresnej analýzy. Očakávalo sa, že pomocou neurónových sietí bude možné modelovať nelineárne vzťahy medzi spotrebou a počasím. Využitiu metód umelej inteligencie na predikciu je v posledných rokoch venovaná väčšia pozornosť ako štatistickým metódam.

Expertné systémy [67, 109, 110] predpovedajú na základe vedomostí, ktoré majú uložené ako súbor pravidiel. S každým novým údajom môžu odvádzať nové pravidlá. Pri predpovedi elektriny sú pravidlá odvádzané napríklad z údajov o teplote, type dňa, spotrebe z predošlého dňa.

Fuzzy logika môže byť využitá v expertných systémoch [136], ale i v neurónových sieťach [66, 102]. Výhodou fuzzy dedukčných systémov je, že oproti expertným systémov nepotrebujú toľko informácií od doménového experta.

Podporné vektory (angl. *support vector machines*, skr. *SVM*) boli tiež úspešne využité ako technika predikcie spotreby elektriny [64], ale pre strednodobé prognózy. Model SVM bol rozšírený, aby vedel odhadovať nelineárne vzťahy medzi premennými, na tzv. regresiu podporných vektorov (angl. *support vector regression*, skr. *SVR*) [55].

Viac prác, ktoré sa venujú metódam predpovede spotreby využívajúc umelú inteligenciu – najmä neurónové siete a SVR, a metódam pre ich optimalizáciu, napr. evolučné algoritmy a optimalizácia rojom častíc (angl. *swarm particle optimization*), je opísaných v aktuálnych prehľadoch [53, 56].

6.1.4 Hybridné metódy

V snahe vylepšiť presnosť prognóz, mnohé metódy kombinovali rôzne prístupy. Proces predikcie niektorých metód sa skladá z viacerých krokov, pričom v každom kroku je použitá iná technika, napr. iná technika pre identifikáciu parametrov modelu a iná na postavenie modelu, počiatočná predpoveď jedným spôsobom a doladenie predpovede technikou založenou na fuzzy logike, alebo jedna technika na predspracovanie dát a druhá na predikciu [53].

Výber vhodnej metódy na predikciu spotreby elektriny závisí od kontextu, v akom má byť metóda použitá. Najvýznamnejšie kritéria zahŕňajú dĺžku prognózy, množina dostupných premenných – historické dáta, meteorologické údaje, úroveň doménových a odborných znalostí o metódach predikcie, očakávané zmeny v povahe dát, zložitosť (zrozumiteľnosť) modelu a výpočtovú či pamäťovú zložitosť. Výhody a nevýhody základných metód krátkodobého prognózovania spotreby sú zosumarizované v tabuľke 4.

Tab. 4. Výhody a nevýhody základných metód pre krátkodobé prognózovanie spotreby elektriny [52].

	výhody	**nevýhody**
viacrozmerná regresia	• ľahko interpretovateľný model, • ľahká na implementáciu, aktualizáciu a automatizáciu, • dobrá presnosť	• spolieha sa na vysvetľujúce premenné, • potrebuje vysvetľujúce premenné, určenú funkciu a aspoň dva roky trénovacích dát
ARIMA	• málo parametrov, ktoré treba odhadovať, • nepotrebuje veľa historických dát na trénovanie, • dobrá presnosť pre krátke intervaly	• malá presnosť na dlhšie intervaly, • vysoké náklady na implementáciu, aktualizáciu a automatizáciu, • ťažko interpretovateľný model kĺzavých priemerov

neurónové siete	• potrebné minimálne vedomosti o štatistike a doméne, • dobrá presnosť počas obyčajných dní	• vysoká výpočtová náročnosť, • veľa parametrov, • ťažko interpretovateľné, • nízka presnosť pri nezvyčajnom počasí

Porovnanie jednorozmerných metód krátkodobého prognózovania spotreby elektriny [123], ktoré používajú na predpoveď len historické dáta o spotrebe, vyhodnotil ako najlepšiu metódu exponenciálne vyrovnávanie rozšírené o dvojitú sezónnosť. V nedávnej štúdii metód s trojitým sezónnym cyklom sa ukázala ako najpresnejšia kombinácia ARIMA modelu a exponenciálneho vyrovnávania [127]. Uvádzané prieskumy porovnávali navzájom iba metódy analýzy časových radov. Bohužiaľ sme sa nestretli s empirickou štúdiou, ktorá by porovnávala komplexne všetky prístupy k predikcii – regresiu, analýzu časových radov aj umelú inteligenciu.

6.2 Zhlukovanie

Okrem úlohy predikcie sme skúmali aj metódy zhlukovania, ktoré sú využiteľné pre vylepšenie výsledkov predikcie[36] zhlukovaním spotrebiteľov s podobnou krivkou odberu elektrickej energie. Zanalyzovali sme základné metódy zhlukovania ako aj metódy použiteľné na veľké objemy dát a prúdy dát. Zaoberali sme sa aj vplyvom redukcie dát na zhlukovanie – vplyv na výsledok a výpočtovú náročnosť zhlukovania. Publikovaný výstup z tejto oblasti je opísaní v časti 7.7.

6.2.1 Základné metódy zhlukovania

6.2.1.1 K-means

Najznámejším zhlukovacím algoritmom je určite K-means, jeho cieľom je rozdeliť N pozorovaní do k zhlukov, v ktorom každé pozorovanie patrí do zhluku s najbližším centroidom, ktorý je reprezentantom zhluku. Optimalizačné kritérium roztriedenia objektov do zhlukov je založené na minimalizácii súčtu štvorcov euklidovskej vzdialenosti medzi každým objektom zhluku a prislúchajúcim centroidom.

Tento problém roztriedenia je vo všeobecnosti výpočtovo veľmi ťažký (tzv. NP-ťažký). Avšak existujú veľmi účinné heuristické algoritmy, ktoré roztriedenie „dobre“ zvládnu. Jedným z nich je jednoduchý Lloydov algoritmus [84].

Princíp Lloydovho algoritmu:

Zvolíme náhodne počiatočné roztriedenie do *k* zhlukov. Vypočítame centroidy zhlukov a každý objekt spojíme s centroidom, ktorý je k nemu najbližšie. Potom vytvoríme nové zhluky

[36] LAURINEC, P., LUCKÁ, M.: Analýza vplyvu redukcie dimenzionality na zhlukovanie veľkých dátových množín. In: *Proceedings of Data a Znalosti 2015*, 2015, pp. 1–6.

tým spôsobom, že objekty priradíme k najbližším centroidom. Vypočítajú sa nové centroidy. Takto sa to opakuje, až kým nedostaneme dvakrát rovnaké usporiadanie do zhlukov, alebo kým nie je dosiahnutý maximálny počet iterácií.

Výhody algoritmu K-means sú takéto:

- Ľahko pochopiteľný a realizovateľný.
- Rýchlo dokonverguje k „dobrému“ riešeniu pri konečnom počte iterácii.

Nevýhody algoritmu K-means sú tieto:

- Rôzne iniciálne roztriedenia do zhlukov môžu viesť k rôznym finálnym roztriedeniam. Preto je vhodné algoritmus spustiť viackrát pri rôznych začiatočných roztriedeniach.
- Zhlukovanie závisí od jednotiek, v ktorých meriame objekty. Ak sú premenné v rozličných mierkach, alebo sú veľmi odlišné vzhľadom na ich rozsah, potom sa odporúča štandardizácia dát.
- Nie je vhodný, ak zhluky majú nekonvexné tvary.
- Premenné musia byť reálne vektory.

6.2.1.2 K-medoids

Metóda K-medoids je veľmi podobná metóde K-means. Namiesto centroidov sa tu používajú medoidy. Medoid je najstrednejší objekt zhluku, alebo inak povedané, najlepší reprezentant zhluku. Práve preto môžeme používať len vzájomné vzdialenosti (resp. nepodobnosti) medzi objektami. Cieľom je nájsť zhlukovanie, ktoré minimalizuje súčet nepodobností medzi objektom v zhluku a prislúchajúcim medoidom.

Rovnako, ako pri K-means, tento problém je náročné vypočítať a preto sa používajú vhodné heuristické algoritmy. Jedným z nich je algoritmus ,,Rozdeľovanie okolo medoidov'' (*angl.* Partitioning around medoids, skr. PAM) [111].

Princíp algoritmu PAM:

Náhodne zvolíme *k* objektov ako začiatočných medoidov. Kým nie je presiahnutý maximálny počet iterácií, alebo ak už nie je zmena v medoidoch, tak opakujeme nasledovné kroky. Vypočítame vzdialenosť každého objektu k jeho najbližšiemu medoidu a spočítame sumu nepodobností. Nasledovne zameníme každý medoid za každý nemedoid a spočítame sumu nepodobností. Vyberieme usporiadanie s najmenšou sumou nepodobností.

Výhody algoritmov typu K-medoids sú takéto:

- Ľahko pochopiteľný a realizovateľný.
- Rýchlo dokonverguje k „dobrému“ riešeniu pri konečnom počte iterácii.
- Väčšinou je menej citlivý na ,,odľahlé pozorovania''.
- Povoľuje používanie nepodobností objektov (vhodný napr. na zhlukovanie textových dokumentov a pod.).

Nevýhody algoritmov typu K-medoids sú tieto:

- Rôzne iniciálne roztriedenia do zhlukov môžu viesť k rôznym finálnym roztriedeniam. Preto je vhodné algoritmus spustiť viackrát pri rôznych začiatočných roztriedeniach.
- Zhlukovanie závisí od jednotiek v ktorých meriame objekty. Ak sú premenné v rozličných mierkach, alebo sú veľmi odlišné vzhľadom na ich rozsah, potom sa odporúča štandardizácia dát.

6.2.1.3 Analýza zhlukov založená na normálnom modeli

Modernejšou metódou zhlukovania je analýza zhlukov založená na normálnom modeli [36]. Hlavným predpokladom tejto metódy je, že vektory pozorovaní $x_1, x_2, \ldots, x_n$ charakterizujúce objekty, sú realizáciami náhodných vektorov $X_1, X_2, \ldots X_n$, a tie majú p-rozmerné normálne rozdelenie $X_i \sim N_p(\mu_{\gamma i}, \Sigma_{\gamma i})$, kde $\gamma \in \Gamma_{n,k}$, $\mu_1, \ldots, \mu_k \in \mathbb{R}^p$ (stredné hodnoty) a $\Sigma_1, \ldots, \Sigma_k \in \mathbb{S}_{++}^p$ (kovariančné matice) sú parametre nášho modelu. Pod symbolom $\mathbb{S}_{++}^p$ budeme rozumieť p-rozmerný priestor pozitívne definitných matíc. Predpoklad normálneho rozdelenia náhodných vektorov umožňuje používať klasicky známe štatistické postupy, čo je veľká výhoda. Optimálne roztriedenie do zhlukov nájdeme pomocou metódy maximálnej vierohodnosti. Čiže maximalizujeme vierohodnosť súčinu hustôt p-rozmerného normálneho rozdelenia. Výsledná optimalizačná funkcia (kritérium) pozostáva z determinantu kovariančných matíc zhlukov. Ďalšou veľkou výhodou tejto metódy je, že ňou môžeme meniť všetky vlastnosti rozdelenia medzi zhlukmi. Pod vlastnosťami rozdelenia zhluku rozumieme jeho orientáciu, tvar a veľkosť. Kľúčom k tomuto je reparametrizácia kovariančných matíc Σj, pokiaľ ide o jeho rozklad za pomoci vlastných čísel. Vyriešenie tohto optimalizačného problému je NP-ťažké. Využíva sa naň väčšinou EM-algoritmus (angl. Expectation Maximization) [49] alebo aj napríklad genetický algoritmus [76].

Najväčšia výhoda tejto metódy je, že dokáže nájsť zhluky v zhlukoch, prekrývajúce sa zhluky a rôzne eliptické tvary. Nevýhodou je väčšia výpočtová náročnosť a zložitosť metódy.

6.2.1.4 DBSCAN

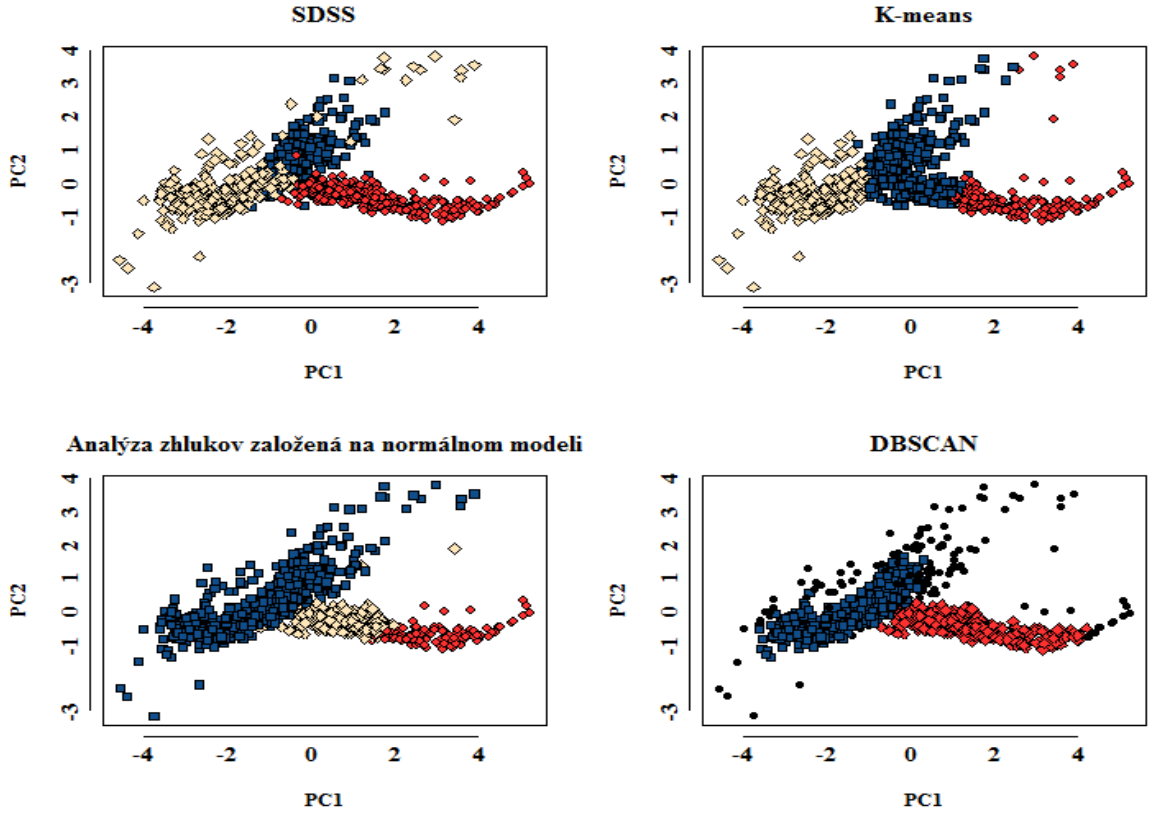
DBSCAN (Density Based Spatial Clustering of Applications with Noise) je algoritmus založený na hustote objektov [31]. Vlastnosti tohto algoritmu sú:

- Schopný identifikovať zhluky ľubovoľných tvarov.
- Schopný zhlukovania pri zašumených dátach.
- Netreba vyberať dopredu počet zhlukov.

Opisuje zhluky ako oblasti s vysokou hustotou bodov. Neformálna definícia takéhoto zhluku by mohla vyzerať nasledovne. Pre bod v zhluku, susedstvo toho bodu, s istým daným polomerom, musí obsahovať istý minimálny počet bodov.

Na obrázku 27 je zobrazené porovnanie výsledkov troch metód analýzy zhlukov (K-means, analýza zhlukov založená na normálnom modeli a DBSCAN) na dátovej množine SDSS

s počtom pozorovaní 1084 a počtom dimenzií 4. Zobrazenie do dvojrozmerného priestoru je vykonané pomocou metódy hlavných komponentov.



Obr. 27. Porovnanie troch metód analýzy zhlukov na dátovej množine SDSS.

6.2.2 Zhlukovanie veľkých objemov dát

S narastajúcim počtom dát majú zhlukovacie algoritmy problém s komplexnosťou a tým spojenou škálovateľnosťou a rýchlosťou [118].

Postupne si predstavíme algoritmy, ktoré sa snažia vysporiadať s veľkými dátami. Ich rozdelenie, ktoré ide podstate aj v chronologickom poradí ich vzniku, je nasledovné:

- Metódy vykonávané na jednom počítači (procesore)
  - Techniky založené na odberu vzoriek
  - Redukcia dimenzionality
- Metódy vykonávané na viacerých počítačoch (procesoroch)
  - Paralelné techniky
  - MapReduce

6.2.2.1 Metódy analýzy zhlukov vykonávané na jednom procesore

Techniky založené na odbere vzoriek sú prvými pokusmi ako zlepšiť rýchlosť a škálovateľnosť. Ich názov je odvodený od odberu vzoriek, nakoľko namiesto zhlukovania na celom dátovom súbore, ho vykonávajú na vzorke dátovej množiny a výsledky potom

zovšeobecňujú pre celý dátový súbor. Predstavíme tri metódy takéhoto zhlukovania. Sú to CLARA , CLARANS a CURE .

CLARA. CLARA (angl. Clustering Large Applications) je rozšírením metódy K-medoids (teda algoritmu PAM) pre veľké počty dát. Tento prístup vykonáva analýzu zhlukov na vybratej vzorke z dátovej množiny a potom priraďuje všetky objekty v dátovej množine k vytvoreným zhlukom. Kroky algoritmu sú nasledovné:

1. Výber vzorky z *n* objektov a ich priradenie do *k* zhlukov.
2. Priradenie každého objektu v dátovej množine k najbližšiemu zhluku.
3. Uloženie priemernej vzdialenosti medzi objektami a ich príslušných zhlukov.
4. 5-krát sa zopakuje tento proces a vyberie sa výsledok s minimálnymi priemernými vzdialenosťami.

CLARANS. CLARANS (Clustering Large Applications based on Randomized Sampling) bol navrhnutý na vylepšenie efektívnosti algoritmu CLARA. Cieľom je nájsť také lokálne optimum prehľadaním celého súboru, pre ktoré je rozdiel len v každej iterácií, že prezrie len vzorku susedov v okolí medoidu. CLARA teda vykonáva odber vzoriek na začiatku a to ovplyvňuje (obmedzuje) celý prehľadávací proces, pričom v CLARANS, odber vzoriek je vykonávaný dynamicky pre každú iteráciu v prehľadávacej procedúre. Pozorovania ukázali, že CLARANS je efektívnejší ako CLARA.

CURE. CURE (angl. Clustering Using REpresentatives) je algoritmus, ktorý predpokladá, že objekty sa nachádzajú v euklidovskom priestore. Okrem toho nepredpokladá nič o tvare zhlukov, tieto nemusia byť normálne rozdelené a môžu mať rôzne ohyby, S-tvary alebo aj prstene. Namiesto reprezentácie zhlukov pomocou centrov a medoidov, teda jedným bodom, používa kolekciu reprezentatívnych bodov, ako už aj názov implikuje. V princípe je CURE algoritmus hierarchického zhlukovania. Použije každý objekt ako jeden zhluk a postupne spája dva existujúce zhluky, kým nezostane k zhlukov. Proces vyberania dvoch zhlukov pre spojenie v každom kroku je založené na výpočte minimálnej vzdialenosti medzi všetkými možnými dvojicami reprezentatívnych bodov z dvoch zhlukov. Fázy a kroky CURE algoritmu sú tieto:

Prvá fáza:

1. Vyberie sa malá vzorka dát a rozdelí sa do zhlukov v hlavnej pamäti. Na zhlukovanie sa použije hierarchické zhlukovanie.
2. Vyberie sa malá množina bodov z každého zhluku tak, aby sa stali reprezentatívnymi bodmi. Tieto body sú vybrané tak, aby boli čo najďalej od seba vzdialené.
3. Každý reprezentatívny bod sa presunie o pevný zlomok vzdialenosti medzi jeho pozíciou a centroidom toho zhluku. Dobrým zlomkom je napríklad 20% zo vzdialenosti.

Druhá fáza:

Spoja sa dva zhluky, ak majú pár reprezentatívnych bodov, jeden z každého zhluku, dostatočne blízko. Používateľ môže vybrať vzdialenosť, ktorou definuje blízkosť. Tento krok sa opakuje, kým nie je nájdených viac dostatočne blízkych zhlukov.

Tretia fáza:

Posledný krok CURE je priraďovanie bodov. Každý bod *b* je načítaný zo sekundárneho úložiska a porovnaný s reprezentatívnymi bodmi. Body *b* sú priradené k zhlukom s reprezentatívnymi bodmi, ktoré sú im najbližšie.

6.2.2.2 Paralelné metódy analýzy zhlukov

S prichádzajúcimi dátami s veľkosťou terabajtov až petabajtov už techniky vykonávané na jednom počítači nepostačujú. Je teda potrebné algoritmy spúšťať na viacerých strojoch. Táto technika dovoľuje rozdeliť veľké množstvo dát na menšie časti, ktoré sa spúšťajú na rozdielnych strojoch .

Techniky zhlukovacích algoritmov vykonávaných na viacerých strojoch sa dajú rozdeliť do dvoch kategórií:

- Neautomatické distribuovanie – paralelné
- Automatické distribuovanie – MapReduce

Paralelné a distribuované zhlukovacie algoritmy nasledujú takýto všeobecný postup:

1. V prvej etape sú dáta rozdelené na vzorky a distribuované na jednotlivé procesory.
2. V druhej etape každý procesor vykoná individuálne zhlukovanie na vzorke dát.
3. Potom je vykonané všeobecné zhlukovanie na jednom procesore (počítači).
4. Poslednou etapou je vylepšenie individuálneho zhlukovania a opakovanie 2. a 3. kroku.

Nižšia presnosť distribuovaných algoritmov môže byť spôsobená dvoma dôvodmi. Prvý dôvod je možný, ak sa na rôznych procesoroch použije rôzny zhlukovací algoritmus. Druhým je, keď sa aj použije rovnaký algoritmus na všetkých procesoroch, tak rozdelenie dát na vzorky môže zmeniť výsledky zhlukovania.

DBDC. DBDC je paralelný zhlukovací algoritmus založený na hustote. Teda hustota bodov v každom zhluku je väčšia ako mimo zhluku, kde sa nachádzajú oblasti šumu s nízkou hustotou. DBDC je algoritmus ktorý sa riadi klasickým cyklom pre paralelné zhlukovanie. V individuálnych zhlukovacích etapách sa použije definovaný algoritmus na zhlukovanie a potom pre všeobecné zhlukovanie sa použije algoritmus DBSCAN vykonávaný na jednom procesore, ktorý dokončí výsledky. Výsledky ukázali, že DBDC vytvára zhluky rovnakej kvality ako jeho sériový predchodca, ale je 30-krát rýchlejší.

G-DBSCAN. Nový spôsob paralelného počítania sa nedávno otvoril a v to podobe využitia sily spracovania s GPU namiesto CPU. G-DBSCAN je GPU založený paralelný algoritmus pre

zhlukovanie založený na hustote . Táto metóda sa líši tým, že používa indexovanie založené na grafe pre vytvorenie väčšej flexibility a väčších možností paralelizovania.

G-DBSCAN má dva kroky a obidva sú paralelizované. V prvom kroku sa skonštruuje graf. Každý objekt reprezentuje uzol a hrana medzi dvoma objektmi sa vytvorí, ak ich vzdialenosť je nižšia alebo rovná ako preddefinovaný prah. Ak je tento graf hotový, druhým krokom je identifikácia zhlukov. Na to sa používa algoritmus BFS (angl. Breadth First Search), ktorý prehľadá celý graf.

Výsledky ukazujú, že G-DBSCAN je rýchlejší 112 násobne, ako jeho sériová implementácia.

K-means založené na MapReduce (PK-means). PK-means je distribuovaná verzia zhlukovacieho algoritmu K-means . K-means náhodne zvolí k objektov z dátovej množiny v začiatočnom kroku a potom opakuje dve fázy. Prvá priradí k najbližším zhlukom každý objekt a po skončení priraďovania, v druhom kroku, sa prepočítajú centre zhlukov s priemeru objektov.

PK-means distribuuje výpočet medzi strojmi použitím MapReduce na zrýchlenie a zoškálovanie procesu. Individuálne zhlukovanie, ktoré zahŕňa prvú fázu sa odohráva v map fáze a všeobecné zhlukovanie vykonáva druhú fázu v reduce fáze. PK-means poskytuje presné (rovnaké) výsledky, ako jeho sériový príbuzný K-means, len v oveľa rýchlejšom čase.

MR-DBSCAN. Nedávno navrhnutý algoritmus je MR-DBSCAN, ktorý je škálovateľnou MapReduce verziou DBSCAN algoritmu . Tri hlavné nevýhody paralelného DBSCAN algoritmu MR-DBSCAN doplňuje (opravuje):

1. Nie je úspešný zvládať zaťaženie medzi paralelnými uzlami (angl. nodes).
2. Tieto algoritmy sú limitované v škálovateľnosti, lebo všetky kritické pod-procesy nie sú paralelizované.
3. Ich architektúra a návrh ich limituje k menšej prenositeľnosti na vznikajúce paralelné vzory.

V MR-DBSCAN bola navrhnutá nová metóda rozdelenia dát založená na znížení výpočtových nákladov a všetky pod-procesy sú plno paralelizované.

MapReduce využívajúci GPU. GPU je mnohonásobne efektívnejšia ako CPU. Kým CPU tvorí niekoľko procesorových jadier, GPU sú zložené z tisícov jadier, ktoré ich robia oveľa výkonnejšími a rýchlejšími ako CPU. MapReduce so spojením s CPU reprezentuje veľmi efektívny rámec pre distribuované počítanie, ale ak sa použije s GPU namiesto CPU, tak tento rámec môže ešte vylepšiť škálovateľnosť a rýchlosť. GPMR je MapReduce rámec na použitie viacerých GPU. Hoci v zhlukovacích aplikáciách tento rámec ešte nebol implementovaný, ale rast veľkosti dát naliehajú na vedcov, aby využili viac škálovateľné algoritmy, čiže tento rámec môže byť do budúcna správnym riešením.

6.2.2.3 Techniky založené na predspracovaní dát: redukcia dimenzionality

Okrem počtu pozorovaní (objektov) je dôležitým faktorom aj dimenzionalita dát. Čím väčšiu dimenzionalitu dáta majú, tým je doba vykonávania dlhšia. Techniky odberu vzoriek redukujú dátovú množinu, ale neponúkajú riešenie pre mnohodimenzionálne dáta.

Redukcia dimenzií môže zlepšiť, okrem zníženia časovej náročnosti, aj presnosť zhlukovacích algoritmov. Princípom redukcie je koncentrácia maximálneho počtu informácií (variability v dátach) do minimálneho počtu dimenzií. Zanedbať je možno tie premenné, ktoré majú v sebe málo informácií (variability) a sú v dátovej množine v podstate zbytočné. Ďalšou výhodou redukcie dimenzionality je vizualizácia mnohorozmerných dát v dvojrozmernom alebo trojrozmernom priestore (tzv. graf skóre). Vizualizácia môže pomôcť odhaliť napríklad možné odľahlé pozorovania (angl. outliers), počty a tvary zhlukov, a názorne prezentovať výsledky roztriedenia do zhlukov.

Opíšme najznámejšiu techniku redukcie dimenzionality a to analýzu hlavných komponentov.

Analýza hlavných komponentov. Úlohou analýzy hlavných komponentov (angl. Principal Components Analysis, skr. PCA) [80] je nájdenie náhodného vektora $Z = (Z_1, \ldots, Z_r)^T$, ktorý má nižšiu dimenziu ako náhodný vektor $X = (X_1, \ldots, X_p)^T$. Snaha je, aby náhodný vektor Z obsahoval čo najviac potrebných informácií z predchádzajúceho vektora X.

V analýze hlavných komponentov sa podobne ako v analýze zhlukov založenej na normálnom modeli, využíva rozklad kovariančnej matice pomocou vlastných čísel.

Predpokladajme, že náhodný výber $\mathbf{X}_1, \ldots, \mathbf{X}_n$ má rozdelenie $N_p(\mu, \mathbf{\Sigma})$. Z realizácií týchto náhodných vektorov skonštruujeme dvojrozmerné vektory skóre nasledovným spôsobom:

$$\boldsymbol{Z}_k = (\hat{u}_1^{(n)}, \hat{u}_2^{(n)})^T (\boldsymbol{X}_k - \overline{\boldsymbol{x}}), \quad k = 1, \ldots, n,$$

kde $(\hat{u}_1^{(n)}, \hat{u}_2^{(n)})^T$ sú odhady vlastných vektorov. Tieto odhady sú vlastné vektory výberovej kovariančnej matice:

$$\hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{x}})(\boldsymbol{X}_i - \overline{\boldsymbol{x}})^T$$

Počet hlavných komponentov, ktoré sa zanedbávajú, môže byť aj iný. Najjednoduchším pravidlom je ponechanie toľkých hlavných komponentov, aby v dátach zostalo približne 80 – 95% variability.

Nelineárne metódy redukcie dimenzionality. Doteraz spomenuté metódy redukcie dimenzionality predpokladali lineárne vzťahy medzi premennými, alebo ich normálnosť (pozorovania pochádzajú z normálneho rozdelenia). Tieto predpoklady nie sú vo väčšine praktických prípadoch splnené. Nelineárne metódy oproti vyššie uvedeným metódam hľadajú v dátach „nenormálne“ vzory.

Jednou takouto metódou redukcie dimenzionality je analýza nezávislých komponentov (ICA) [63]. Princípom tejto metódy je transformovať mnoho-dimenzionálne dáta do nezávislých (tzv. negaussovských) dimenzií. Tieto dimenzie sa nazývajú *latentné premenné*.

V poslednej dobe je veľmi používanou metódou t-SNE (t-distributed Stochastic Neighbor Embedding) [86]. Princípom tejto metódy je priradiť každej dvojici pozorovaní pravdepodobnosť zo Studentovho t-rozdelenia a transformovať dáta na základe ich entropie. t-SNE dokáže transformovať vysoko dimenzionálne dáta do 2D tak, aby možné skupiny v dátach boli čo najviac oddelené. Menšou nevýhodou tejto metódy je jej stochastickosť, čiže každým spustením algoritmu dostaneme (trocha) iný výsledok.

Nevýhodou väčšiny metód redukcie dimenzionality, ktoré majú model silne založený na štatistike, je ich výpočtová náročnosť v prípade, ak pracujeme s veľkými dátami. Preto v mnoha aplikáciách analýzy zhlukov na mnoho dimenzionálne dáta sa používa náhodná projekcia . Náhodná projekcia, ako je z názvu zrejmé, vyberie náhodne dimenzie z dátovej množiny a tie sa potom použijú na zhlukovanie. Dimenzie sa pravdaže nemusia vyberať len náhodne, ale aj napríklad pomocou nejakej variančnej analýzy (čím viac variancie sa nachádza v premennej, tým je väčšia šanca, že sa v nej nachádzajú dôležité informácie).

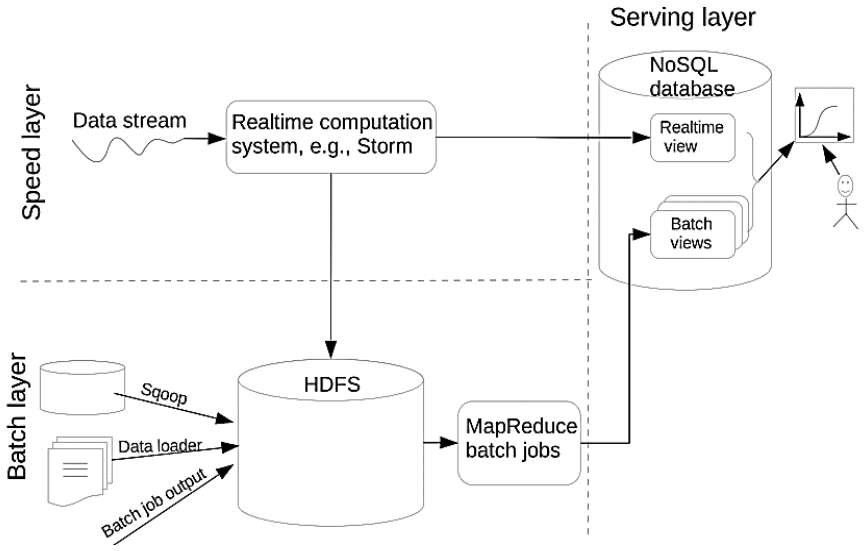
6.3 Nástroje na spracovanie veľkých objemov dát

V súčasnosti mnoho spoločností vyjadruje potrebu pre spracovávanie veľkých objemov dát v reálnom čase [10]. Preto v posledných rokoch vzniklo mnoho prostriedkov pre ich správu a analýzu. Spracovanie dát v reálnom čase sa skladá väčšinou z dvoch hlavných krokov:

- *integrácia* – proces transformácie dát do dátového skladu vo vhodnom formáte, tento proces sa nazýva aj ETL (angl. *extraction-transformation-load*),
- *analýza* – získanie znalostí z dátového skladu.

Dáta sa zvyčajne integrujú dávkovo v určených intervaloch, napr. denne, týždenne, mesačne. Pri spracovaní v reálnom čase musí tento proces prebiehať kontinuálne z rôznych paralelných prúdov dát. Pre umožnenie analýzy v reálnom čase vzniká mnoho metód a prostriedkov pre automatizáciu tradičných metód pre dolovanie v dátach. Veľký dôraz sa kladie na časovú zložitosť metód, ktorá by mala byť čo najmenšia, aby analytik mohol dostávať okamžité odpovede.

Prechodom z dávkového spracovávania veľkých objemov dát na spracovanie v reálnom čase sa musela zmeniť architektúra dátových systémov. Všeobecné riešenie pre orchestráciu nástrojov pre integráciu a analýzu dát poskytol Marz [88], ktorý navrhol *Lambda architektúru*. Tá definuje všeobecný systém, na ktorom je možné vykonávať dopyty nad všetkým dátami (*„query=function(all data)“*). Takéto dopytovanie vo všeobecnosti vyžaduje veľa výpočtových a pamäťových zdrojov. Tento problém rieši architektúra tak, že výsledky sa počítajú vopred ako množina pohľadov a dopyty sa vykonávajú až nad týmito pohľadmi.



Obr. 28. Lambda architektúra [83].

Lambda architektúra (pozri obrázok 28) sa skladá z troch vrstiev – dávkovacia vrstva, vrstva paralelného rýchleho spracovania a vrstva obsluhy.

Dávkovacia vrstva počíta pohľady na zozbierané dáta. Tento proces prebieha nekonečne dokola. Pohľady sú zastarané o čas, ktorý trvá ich výpočet. Tento čas vypĺňa vrstva paralelného rýchleho spracovania, ktorá spracováva najnovšie prichádzajúce dáta a počíta pohľady v skoro reálnom čase. Všetky dopyty sú vybavované vrstvou obsluhy, ktorá berie dáta z pohľadov vytvorených dvomi predošlými vrstvami. Výsledky z oboch typov pohľadov sú spojené dokopy ako ponúknuté ako výsledok dopytu. Výhodou tejto architektúry je, že prichádzajúce dáta sú ukladané nespracované, a preto môžu byť neskôr pridané nové pohľady na ne alebo môžu byť existujúce pohľady znova prepočítané, ak sa zistili nejaké chyby.

V skutočnosti sú jednotlivé vrstvy implementované ako rôzne nástroje. Vrstva paralelného rýchleho spracovania narába s prúdom dát, preto využíva nástroje ako Strom, S4 alebo Spark. Dávkovacia vrstva potrebuje ukladať obrovské množstvá dát a je zvyčajne implementovaná pomocou technológii založených na výpočtovom rámci MapReduce [30] (napr. Cascade, Pig, Hive).Vrstva obsluhy potrebuje vedieť rýchlo vybavovať dopyty nad pohľadmi. Pohľady predstavujú oveľa menší objem dát voči všetkým nespracovaným dátam, ktoré prúdia do systému. Preto sa tu využívajú NoSQL databázy, ktoré sú veľmi rýchle (napr. Memcache, Redis) a v prípade potreby dokážu narábať aj s veľkými objemami dát (HBase, Cassandra, ElephantDB, MongoDB).

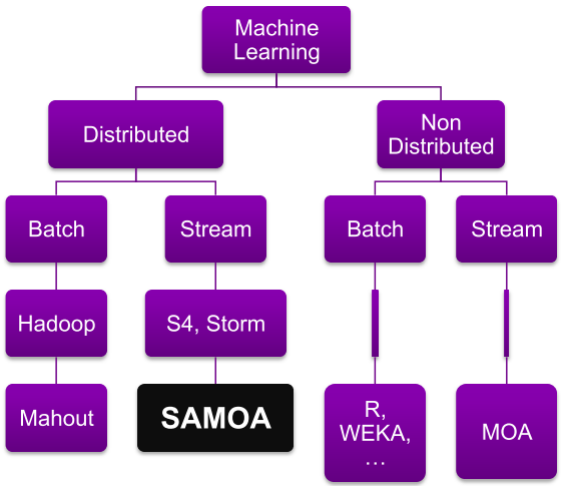
Nástroje na prácu s veľkými objemami dát sa rozdelili vo všeobecnosti na dve skupiny – tie, ktoré využívajú MapReduce a tie, ktoré nie. Od predstavenia výpočtového rámca MapReduceho využilo mnoho spoločností v komerčných (Informatica, IBM Websphere, Oracle Data Warehouse) aj open source (Pig, Hive, Cascade) produktoch a stal sa štandardom. MapReduce využíva množstvo uzlov v zhlukoch na paralelné spracovávanie,

a tým významne skracuje čas vykonávaných operácií. Druhú skupinu nástrojov pre správu veľkých objemov dát predstavujú tzv. NoSQL databázy. Patria sem databázy, ktoré nepoužívajú na ukladanie dát relácie, ale iné štruktúry ako dvojice kľúčov a hodnôt, stĺpce, grafy alebo dokumenty (MongoDB, ElephantDB, Cassandra). Ich výhodou je, že dokážu niektoré operácie vykonávať rýchlejšie ako relačné databázy.

Viaceré nedávne prieskumy ponúkajú prehľad súčasných nástrojov na spracovanie veľkých objemov dát v reálnom čase [15, 34, 83, 98]. My uvádzame pár open source nástrojov, ktoré sa sústreďujú na spracovanie a dolovanie v prúdových dátach. Nástroje sa delia podľa ich schopnosti pracovať distribuovane a paralelne počítať na viacerých uzloch, čo znižuje čas výpočtu pri väčších objemoch dát. Ďalej sa členia podľa toho, či dáta spracovávajú dávkovo alebo prúdovo (pozri obrázok 29).

Hadoop a Mahout. Najznámejším nástrojom spájaným s veľkými objemami dát je *Hadoop*. Je to open source implementácia výpočtového rámca MapReduce od spoločnosti Apache. Jeho nevýhoda je, že nie je vhodný na spracovanie v reálnom čase, pretože model MapReduce je navrhnutý tak, aby bol škálovateľný a robustný voči chybám, ale nie je optimalizovaný pre efektívne čítanie a zápis (angl. *I/O operations*) [78]. Je ideálny pre dávkové spracovanie dát, preto je využitý napríklad v dávkovacej vrstve Lambda architektúry. Samotný Hadoop neobsahuje algoritmy strojového učenia. Implementuje ich knižnica *Mahout*. Algoritmy využívajú distribuované počítanie na rámci Hadoop.

Viacero nástrojov – Hadoop Online [27], HStreaming[37], HaLoop [22], DEDUCE [73], sa snažilo odstrániť nedostatok Hadoopu a prispôsobiť ho pre spracovanie prúdových dát, no ponúkajú len obmedzenú podporu pre prúdové počítanie.



Obr. 29. Taxonómia nástrojov na dolovanie v dátach [98].

Storm. Nástroj Storm[38] je distribuovaný výpočtový rámec vhodný pre spracovanie prúdových dát. Je to prostredník medzi databázou a zdrojom dát. Jeho tvorcom je Marz, ktorý vytvoril aj

[37] www.adello.com

Lambda architektúru. Momentálne je vyvíjaný spoločnosťou Apache. Spracováva prúdy podobne ako Hadoop dávkové dáta. V praxi je často používaný s databázou HBase, ktorá je postavená na Hadoop, ale má optimalizované operácie čítania a zapisovania. Storm neobsahuje algoritmy strojového učenia a slúži len na efektívnu manipuláciu s prúdovými dátami.

S4. Platforma S4[39] (Simple Scalable Streaming System) bola vytvorená spoločnosťou Yahoo!, teraz je vyvíjaná spoločnosťou Apache. Je to distribuovaný výpočtový rámec ako Storm, ale líši sa decentralizovanou architektúrou. Všetky výpočtové uzly v zhlukoch sú si rovné. Motiváciou pre vytvorenie S4 bola absencia výkonnej platformy, ktorá automaticky zvláda paralelné počítanie, komunikáciu medzi uzlami a nezaťažuje tým programátora aplikácie, ktorá potrebuje spracovávať prúd dát [99]. S4 neobsahuje algoritmy strojového učenia.

MOA. Nástroj MOA [16] (Massive Online Analysis) je určený na dolovanie v prúdoch. Obsahuje súbor algoritmov strojového učenia pre úlohy ako klasifikácia, regresia, zhlukovanie a detekcia extrémov. Obsahuje aj metódy pre detekciu zmien v prúde a overenie vytvorených modelov. Je vytvorený na Waikatskej Univerzite na Novom Zélande, kde vznikol aj známy softvér pre dolovanie v dátach WEKA [47]. MOA je napísaná v programovacom jazyku Java a dokáže pracovať len na jednom uzle.

R stream a RMOA. R[40] je softvér pre štatistické výpočty a tvorbu grafov. Obsahuje mnoho balíkov s rôznymi algoritmami pre dolovanie v dátach. Vyvíja a používa ho široká komunita. Rovnako ako MOA je určený pre prácu na jedinom uzle. Nedávno bol predstavený balík *stream* [46], ktorý umožňuje pracovať aj s prúdovými dátami v prostredí R. Zatiaľ implementuje len algoritmy pre zhlukovanie z prúdových dát, no je rozšíriteľný o ďalšie úlohy dolovania v prúdoch. Projekt *RMOA*[41] umožňuje volať funkcie nástroja MOA pre prácu s prúdmi z prostredia R. Momentálne sa sústredí iba na klasifikáciu.

SAMOA. Platforma SAMOA [98], vyvíjaná spoločnosťou Yahoo!, je určená pre distribuované strojové učenie z prúdových dát. Umožňuje prepojenie algoritmov strojového učenia z nástroja MOA so systémom spracovávajúcim prúdy (angl. *stream processing engine*, skr. *SPE*), akým je Storm alebo S4. V budúcnosti má ambície sa stať používanou a vyvíjanou širokou komunitou, podobne ako je teraz softvér R. Nové algoritmy bude možné vyvíjať a dopĺňať formou balíkov.

---

[38] http://storm.apache.org/
[39] http://incubator.apache.org/s4/
[40] http://www.r-project.org/
[41] https://github.com/jwijffels/RMOA

7 Vývoj nových analytických algoritmov

7.1 Kombinácia predikčných modelov

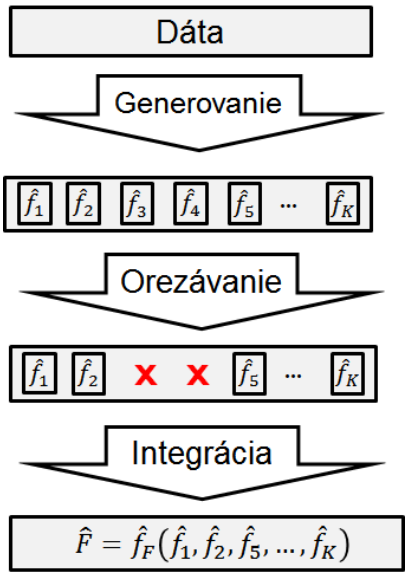
Kombinácia súboru modelov (Ensemble Learning) je metóda z oblasti strojového učenia, ktorá sa začala používať najmä v posledných dvoch dekádach. Jej základnou myšlienkou je, že vhodnou kombináciou výstupov viacerých predikčných modelov môže vzniknúť výsledok, ktorý je presnejší ako výsledok najpresnejšieho modelu použitého v kombinácii [91].

Kombinácia súboru modelov má v súčasnosti niekoľko rôznych implementácii, ako napríklad Bagging [20], Boosting [37], Random Forests [21], Nonlinear Weighted Ensemble Learning [1], Cocktail Ensemble [138] a iné.

Vo všeobecnosti má táto metóda tri hlavné výhody:

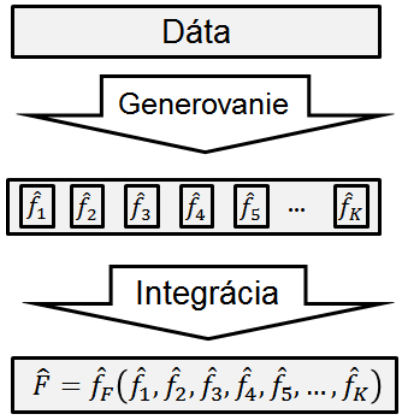
1) Výberom vhodnej kombinácie modelov je možné zvyšovať predikčnú schopnosť modelov a súčasne znižovať ich nedostatky, čo vedie k zlepšeniu výslednej predpovede.
2) V niektorých prípadoch nie je zrejmé, ktorý predikčný model poskytne optimálny výsledok. Preto môže byť použitie kombinačnej stratégie výhodným riešením.
3) Kombinácia výsledkov viacerých modelov môže znížiť chyby predpovede spôsobené nesprávnymi predpokladmi, chýbajúcimi dátami alebo chybnými hodnotami v dátach.

Postup predikcie pomocou súboru metód sa skladá z troch hlavných krokov (pozri Obr. 30). Prvým krokom je vygenerovanie(natrénovanie) sady predikčných modelov ($\hat{f}_1, \hat{f}_2, \hat{f}_3, \dots, \hat{f}_K$). Podľa typu použitých modelov sa rozlišuje heterogénny prístup (keď sú použité predikčné modely rôzneho typu) a homogénny prístup (keď sú použité predikčné modely rovnakého typu, napr. iba neurónové siete). Druhým krokom je orezávanie, ktorého úlohou je výber najvhodnejších modelov z existujúcej sady, s cieľom zvýšiť celkovú presnosť výslednej predikcie. Tretím krokom je integrácia výstupov predikčných modelov do jedného výsledku $\hat{F}$. Na tento účel sa používajú najrôznejšie váhovacie algoritmy $\hat{f}_F$, ktoré sa snažia vypočítať optimálne váhy pre lineárnu kombináciu výstupov jednotlivých predikčných modelov.



Obr. 30. Hlavné časti postupu predikcie pomocou súboru metód [91].

Medzi najkritickejšie problémy spojené s použitím metódy kombinácie súboru modelov patrí zabezpečenie rôznorodosti predikčných modelov a výber vhodného kombinačného prístupu [91]. Naša navrhovaná metóda rieši obidva problémy. Vychádza z heterogénneho prístupu, ktorý implicitne zaručuje rôznorodosť sady základných modelov. Navyše, nie je potrebné použiť žiadne orezávanie, čím sme znížili výpočtovú a časovú zložitosť metódy (pozri Obr. 31).



Obr. 31. Postup predikcie pomocou nami navrhnutého prístupu bez nutnosti použitia orezávania sady modelov.

Klasické prístupy váhovania, ako napríklad jednoduchý priemer alebo metóda General Ensemble Model (GEM) [106] nedosahujú najpresnejšie výsledky. Navyše sa nevedia vysporiadať s problémom multikolinearity[42], ktorá sa môže prirodzene vyskytnúť medzi

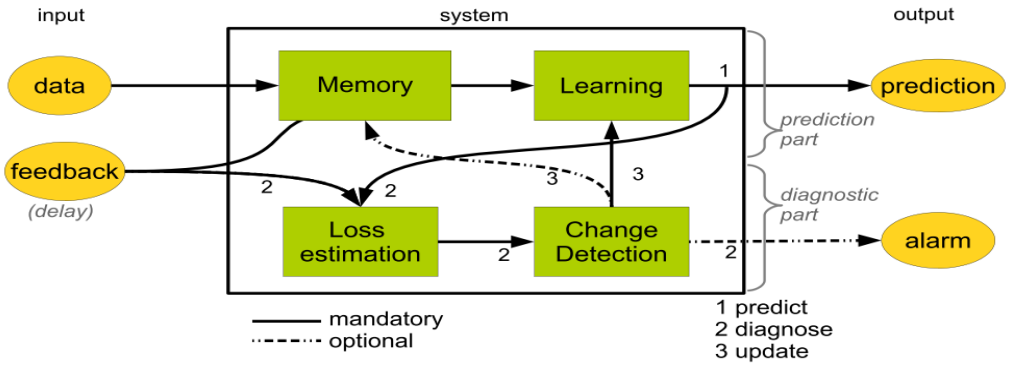
[42] Existencia vzťahu lineárnej závislosti medzi vysvetľujúcimi premennými, ktorá spôsobuje veľké štandardné chyby pri operáciách s inverznými maticami.

výstupmi predikčných modelov, čo spôsobuje zhoršenie výslednej predpovede. Komplexnejšie váhovacie prístupy ako napríklad metóda založená na oblasti expertízy alebo dynamické váhovanie, vyžadujú veľké množstvo trénovacích dát.
Výpočet optimálnych váh môže byť považovaný za optimalizačnú úlohu, preto sme si na jej riešenie zvolili biologicky inšpirované metaheuristiky (napr.: genetický algoritmus, umelá včelia kolónia, optimalizácia rojom častíc) a robustnú metódu založenú na mediánovej chybe modelov. Uvedené prístupy sa vedia vysporiadať s problémom multikolinearity, sú ľahko pochopiteľné a nie sú výpočtovo náročné. Ich použitím sme prekonali výsledky klasických kombinačných prístupov.

7.2 Inkrementálne a adaptívne modely

Na základe požiadaviek na metódy predikcie prúdových dát v reálnom čase (včasné prispôsobovanie sa zmenám, správna identifikácia zmien od extrémov, krátky výpočtový čas a konštantná pamäťová náročnosť) bol formálne definovaný všeobecný proces metódy skladajúci sa z troch krokov (pozri obrázok 32) [40]:

1. *predikcia* – keď prídu nové dáta, vytvorí sa predikcia pomocou aktuálneho modelu,
2. *diagnostika* – po príchode skutočných hodnôt sa vyhodnotí presnosť,
3. *aktualizácia* – na základe diagnostiky sa aktualizuje model predpovede.



Obr. 32. Všeobecný proces metódy predikcie prúdových dát v reálnom čase [40] (1 – predikcia, 2 – diagnostika, 3 – aktualizácia; plné čiary – povinné, prerušované čiary – nepovinné kroky).

Systém pre spracovanie prúdových dát v reálnom čase sa v tomto procese skladá zo štyroch modulov. Pamäťový modul určuje, ktoré z prichádzajúcich dát pôjdu do učiaceho sa modulu. Učiaci sa modul obsahuje algoritmus na predikciu, ktorý vytvára model. Modul na odhad strát vyhodnocuje presnosť predikcie a odosiela informácie do modulu na detekciu zmien v prúde. Ten aktualizuje model predikcie vytvorený v učiacom sa module

Na základe štyroch modulov systému pre dolovanie prúdových dát môžu byť rozdelené jednotlivé metódy pre dolovanie v prúdoch podľa toho, akú techniku využívajú v jednotlivých moduloch, t. j. ako spracovávajú dáta z pohľadu pamäťových nárokov, ako sa učia model predikcie, ako ho vyhodnocujú a ako sa prispôsobujú zmenám v prúdoch. Modulárny pohľad

na adaptívne metódy predikcie prúdových dát v reálnom čase umožňuje ich ľahkú klasifikáciu a poukazuje na možnosti kombinovania techník jednotlivých modulov pre vytvorenie čo najlepšej metódy na predikciu prúdových dát s ohľadom na doménu, v ktorej má byť využitá.

Z hľadiska učenia môžeme algoritmy rozdeliť do štyroch skupín [40]:

- **offline (pretrénovanie)**

  Tieto algoritmy sa učia z celej trénovacej množiny dát. Počas celého trénovania majú prístup k celej dátovej množine, môžu z nej náhodne čítať údaje a čítať ich viackrát. Je to vlastne tradičné objavovanie znalostí, ktoré sme spomínali v časti 4. Naučený model predikcie je nasadený až keď skončí trénovanie.

- **inkrementálne**

  Inkrementálne algoritmy sa trénujú aj počas toho ako sú už nasadené. Dáta spracovávajú tak ako prichádzajú, čiže jeden-po-jednom alebo dávku-po-dávke, napr. ak by prichádzali naraz dáta za uplynulý deň. Čítajú dáta len raz. Skupina inkrementálnych algoritmov s čiastočnou pamäťou si časť dát uchováva a tie môže čítať aj viackrát.

- **online**

  Online algoritmy sú natrénované na neúplnej dátovej množine a počas toho ako sú nasadené sa opakovane aktualizujú s každými novo prichádzajúcimi dátami.

- **prúdové**

  Prúdové algoritmy sú online pre veľmi rýchlo prichádzajúce dáta. Stačí im jeden prechod dátami a majú obmedzený čas a pamäť na spracovanie jednotky prichádzajúcich dát.

Návrhom inkrementálneho a adaptívneho postupu spracovania veľkých objemov dát sme sa zaoberali aj v našich výstupoch, ktoré boli publikované na lokálnych fórach zameraných aj na dátovú analytiku a veľké objemy dát, konkrétne pracovná dielňa WIKT[43] a študentská konferencia IIT.SRC[44], ktorá je pravidelne organizovaná na Fakulte Informatiky a Informačných Technológií Slovenskej Technickej Univerzity v Bratislave.

Adaptívne modely

[43] VRABLECOVÁ, P.: Inteligentná analýza veľkých objemov dát. In: *Proceedings of 9th Workshop on Intelligent and Knowledge Oriented Technologies*, 2014, pp. 126–129.
[44] VRABLECOVÁ, P.: Power demand forecasting from stream data with concept drifts. In: *Proceedings of 11th Student Reasearch Conference in Informatics and Information Technologies*, 2015, pp. 131–132.

Hlavnou nevýhodou modelovania spotreby elektrickej energie tradičnými metódami je, že natrénovaný model je po nasadení do systému nemožné zmeniť. Ak predpokladáme, že údaje o spotrebe budú do systému kontinuálne prúdiť a predpoveď sa bude vykonávať v reálnom čase, musí byť model schopný udržať si svoju presnosť aj pri fluktuáciách v prúde dát. Preto musí byť adaptívny a vedieť prispôsobovať sa možným zmenám (pozri začiatok kapitoly). To, že adaptívne metódy a predikcia v reálnom čase sú potrebné si ľudia uvedomovali už v 90-tych rokoch [32, 101, 105] ešte pred definíciou veľkých údajových korpusov. No ešte v roku 2002 bolo konštatované, že tejto oblasti sa venuje málo výskumníkov [4]. Keďže od 90-tych rokov sú najpopulárnejším nástrojom na predikciu neurónové siete a metódy umelej inteligencie, väčšina adaptívnych mechanizmov pochádza z tejto oblasti.

Práce na tému *adaptívne prognózovanie spotreby elektrickej energie* za posledných dvadsať rokov (cca od 1990) môžeme rozdeliť do nasledujúcich štyroch skupín.

Periodická aktualizácia parametrov. Metódy spadajúce do tejto kategórie modelujú spotrebu kombináciou modelov – pre lineárnu a nelineárnu časť spotreby elektriny. Počas predikcie sú opakovane aktualizované ich parametre alebo váhy, s akými vstupujú do predpovede tak, aby sa minimalizovala chyba. ARIMA modely boli takto využité s metódou maximálnej vierohodnosti [90], metódou vážených rekurzívnych najmenších štvorcov (angl. WRLS) [32] či metódou pre odhad minimálnej priemernej štvorcovej chyby (angl. MMSE) [64]. V [105] pomocou Kalmanovho filtra predpovedali lineárnu časť kompozitného modelu, ktorého parametre boli prispôsobované metódou exponenciálnych rekurzívnych vážených štvorcov. Digitálne filtre použité na predikciu takisto minimalizovali vektor chýb rekurzívnou metódou [87]. Z oblasti umelej inteligencie sme sa stretli s dvomi prácami, v ktorých váhy neurónovej siete boli aktualizované v reálnom čase [23, 65]. Pomocou metódy rekurzívnych najmenších štvorcov bola minimalizovaná aj chyba predpovede aditívneho modelu [9] a hybridného modelu semi-parametrickej regresie [85].

Adaptívny model. Parametre modelov nie sú aktualizované mechanizmom pre korekciu chýb, ale adaptácia je ukotvená v samotnom modeli. V [69] bol vytvorený adaptívny ARIMA model, ktorý používal spotrebu elektriny za predošlý deň ako ďalšiu vysvetľujúcu premennú. Typickým príkladom adaptívneho modelu je exponenciálne vyrovnávanie s hladkým prechodom (angl. *smooth transition exponential smoothing*), ktorého vyrovnávacia konštanta bola nahradená prechodovou funkciou [126].

Pravidelné pretrénovanie. Alternatívou k automatickému prispôsobovaniu sú metódy, ktoré ho simulujú pravidelným pretrénovaním modelu s novými dátami. Táto stratégia je využívaná najmä v modeloch neurónových sietí, ktoré sú často opisované ako adaptívne. Táto vlastnosť však súvisí s adaptívnym učením modelu, ktorý je postupne trénovaný na dodaných dátach. Počas predikcie je model už nemenný. Metódy, ktoré optimalizujú adaptívne učenie neurónových sietí, počítajú s tým, že model bude často pretrénovaný, a preto sa snažia tento proces čo najviac zefektívniť. Príkladmi tohto prístupu sú práce [11,

42, 54, 81, 83, 97, 134], kde pretrénovanie prebieha na dennej, týždennej, mesačnej báze, ale aj vždy po vykonaní predpovede na ďalšie obdobie v hodinových intervaloch [137].

Súbory modelov ošetrujúce každý prípad. Poslednou skupinou adaptívnych metód sú tie, ktoré predpovedajú pomocou veľkej skupiny dopredu natrénovaných modelov [33, 117]. Výber modelu, pomocou ktorého sa bude predpovedať v danej chvíli, závisí od vstupných parametrov ako typ dňa a hodina.

7.3 Inkrementálna predikcia časového radu heterogénnym súborom modelov

V tejto časti opíšeme nami navrhnutý heterogénny inkrementálny model učenia súboru metód (angl. ensemble learning) na predikciu časových radov[45]. Prístup učenia súboru metód bol zvolený z dôvodu jeho schopnosti rýchlo sa adaptovať na náhle zmeny v rozdelení skúmanej premennej a z dôvodu, že má potenciál byť presnejší ako samostatné modely predikcie. Keďže sa zameriavame na predikciu časového radu spotreby elektrickej energie, v súbore metód musia byť modely, ktoré zvládajú rôzne typy sezónností. Musia byť vypočítané modely zahŕňajúce ročnú sezónnosť. Takéto modely stačí prepočítať raz za rok (dlhodobá predpoveď). Modely kopírujúce denné sezónnosti potrebujú dáta dĺžky jeden deň až viac a môžu byť počítané inkrementálne (krátkodobá predpoveď). Potenciálom navrhnutého modelu učenia súboru metód je použitie viacerých jednoduchých metód, ktoré sa dajú prirodzene paralelizovať, a jeho schopnosťou adaptácie na inkrementálne učenie. Toto predurčuje túto metódu vhodnú na veľké prúdy údajov.

Základné modely vchádzajúce do učenia súborov sú dvoch typov – regresné modely a modely založené na analýze časových radov.

Modely krátkodobej predikcie boli vybrané nasledovné:

- Sezónny naivný model náhodnej prechádzky.
- Autoregresný model.
- Trojité Holt-Wintersovo exponenciálne vyrovnávanie skombinované s diskrétnou waveletovou transformáciou.
- Sezónna dekompozícia časového radu založená na lokálnej regresii *loess* spolu s kombináciou s ARIMA modelom.
- Sezónna dekompozícia časového radu založená na lokálnej regresii *loess* spolu s kombináciou s exponenciálnym vyrovnávaním.
- Viacnásobný lineárny regresný model.
- Jednoduchá neurónová sieť s jednou skrytou vrstvou.
- Regresia založená na podporných vektoroch skombinovaná s diskrétnou waveletovou transformáciou.

[45] KOSKOVÁ, G. et al.: Incremental ensemble learning for electricity load forecasting. *Acta Polytechnica Hungarica*, 2015 [prijaté].

Modely dlhodobej predikcie boli vybrané nasledovné:

- Model kĺzavých priemerov.
- Model kĺzavých mediánov.
- Trojité exponenciálne vyrovnávanie s ARIMA chybami zvládajúce dvojitú sezónnosť.

Modely krátkodobej predikcie sa trénovali na trénovacej množine dĺžky 4 až 10 dní. Model učenia súborov je použitý na predikciu spotreby na jeden deň dopredu. Nech *h* je počet denne dostupných pozorovaní. V dni *t,* model vytvorí *h* predikcií váženým priemerom predikcií od *m* základných modelov. Ak sú k dispozícii reálne pozorovania na daný deň, tak je vypočítaná chyba predikcie. Na základe vypočítanej chyby sú váhy aktualizované a každý základný model *i=1, ..., m* je pretrénovaný na dávke dát s_i.

Nech Y^t je matica predikcií z *m* základných metód pre nasledujúcich *h* pozorovaní v čase *t*:

$$Y^t = \begin{pmatrix} y_1(t) & \cdots & y_m(t) \\ \vdots & \ddots & \vdots \\ y_{h1}(t) & \cdots & y_{hm}(t) \end{pmatrix} = \begin{pmatrix} y_1(t) & \ldots & y_m(t) \end{pmatrix}$$

a $w^t = (w_1^t \quad \ldots \quad w_m^t)^T$ je vektor váh pre *m* základných metód v dni *t*. Váhy w_j^t sú začiatočne stanovené na hodnotu 1. Váhy a prislúchajúce predikcie sú skombinované na vytvorenie predikcie $y^t = \begin{pmatrix} y_1(t) & \ldots & y_m(t) \end{pmatrix}^T$. Predikcia učenia súboru metód pre *k*-te (*k = 1, …, h*) pozorovanie je vypočítané vzťahom:

$$y_k(t) = \left(\sum_{j=1}^{m} \left[y_j(t) \, w_j(t) \right]\right) / \left(\sum_{j=1}^{m} w_j(t)\right)$$

kde w^t je vektor obsahujúci váhy preškálované na hodnoty z intervalu $\langle 1, 10 \rangle$.

Keď sú pozorovanie pre deň *t* dostupné, tak vektor váh môže byť prepočítaný. Z predikčnej matice Y^t a vektora *h* práve dostupných pozorovaní $y^t = \left(y_1^t \quad \ldots \quad y_h^t\right)^T$, je vypočítaný vektor

chýb $e^t = \left(e_1^t \quad \ldots \quad e_h^t\right)^T$ pre *m* metód. Chyba každého modelu sa vypočíta na základe vzťahu $e_j(t) = \text{median} \left(|y_j(t) - y(t)|\right)$. Vektor chýb je použitý na aktualizovanie vektora váh pre základné modely metódy učenia súborom metód. Váha pre *j*-tu metódu je vypočítaná vzťahom:

$$w_j^{t+1} = w_j^t \frac{\text{median}(e^t)}{e_j^t}.$$

Výhoda navrhnutého typu váhovania je v jeho robustnosti a schopnosti zotaviť sa z účinkov základných metód. Je robustná z dôvodu použitia mediánu absolútnych chýb a mediánu chýb metód. Metóda preškálovania, ktorá nedopustí aby nejaká metóda mala váhu nula, umožňuje modelu zotaviť sa z príslušných základných metód v prítomnosti náhlych zmien.

Metódu sme implementovali v prostredí R. Vyhodnocovanie bolo vykonávané na desiatich zosumovaných dátových množinách, ktoré kopírovali rozloženie odberných miest podľa PSČ. Metódou učenia súborov sme jednotlivé metódy predčili o 8-12%. Používali sme miery vyhodnocovania presnosti predikcií ako MAE, RMSE, sMAPE, MdAE a Wilcoxonov rank sum test (z dôvodu zamietnutia normálneho rozdelenia vektora chýb).

7.4 Aplikácia biologicky inšpirovaných metód pre vylepšenie adaptívnej predikcie súborom modelov

Vo výskume[46] sme sa zamerali na zlepšenie predikčných schopností navrhnutého heterogénneho inkrementálneho modelu učenia súboru metód na predikciu časových radov v situáciách, keď sa v časovom rade vyskytujú postupné alebo náhle zmeny (concept drift).

Navrhnutý model využívajúci váhovaciu metódu založenú na mediánovej chybe nedosahoval očakávanú presnosť predikcie. Preto sme sa v ďalšom výskume rozhodli využiť biologicky inšpirované metaheuristiky, ktoré podľa odborných zdrojov dosahujú výborné výsledky pri riešení optimalizačných úloh medzi, ktoré sa zaraďuje aj výpočet optimálnych váh. Z biologicky inšpirovaných algoritmov sme vybrali Genetický algoritmus (GA), ktorý patrí do skupiny Evolučných algoritmov a Optimalizáciu pomocou roja častíc (PSO), ktorý patrí do skupiny Rojovo inteligentných algoritmov. GA aj PSO sú stochastické optimalizačné algoritmy založené na populácii jedincov. Hlavný rozdiel spočíva v spôsobe výmeny informácii v populácii. Jedince v PSO si navzájom vymieňajú informáciu o najlepšom objavenom riešení, ktorú následne využívajú na zlepšenie aktuálneho riešenia. V genetickom algoritme si jedince nevymieňajú žiadne informácie o existujúcich riešeniach, iba redukujú jedince s najhorším riešením.

Genetický algoritmus je stochastická optimalizačná metóda inšpirovaná procesom evolúcie [51]. Jej základom je predpoklad, že iba najlepšie adaptované jedince dokážu prežiť a úspešne sa rozmnožiť. Implementovaný genetický algoritmus pracuje s populáciou jedincov, v ktorej každý jedinec predstavuje potenciálne riešenie. Vlastnosti každého jedinca sú kódované vo forme chromozómov. V našej implementácii je to d-dimenzionálny vektor reálnych hodnôt/váh. Každý jedinec má svoju fitnes hodnotu vypočítanú fitness funkciou, ktorá na základe vlastností jedinca, odráža jeho vhodnosť pre dané prostredie (resp. jeho vhodnosť pre riešenie danej úlohy).

Na začiatku algoritmu vygenerujeme náhodnú populáciu N jedincov. V každej iterácii sa všetky jedince v populácii navzájom porovnajú a potom sa vyberie 50% najlepších jedincov (prirodzený výber) do nasledujúcej generácie. Druhá polovica novej generácie vzniká procesom kríženia. Vyberajú sa dvojice jedincov (rodičia), ktorých chromozómy sa použijú v procese kríženia. Týmto spôsobom vzniknú dva nové jedince, ktoré sú zaradené do novej

[46] GRMANOVÁ, G. et al.: Application of Biologically Inspired Methods to Improve Adaptive Ensemble Learning. In: *7th World Congress on Nature and Bilogically Inspired Computing*, 2015.

populácie. Keďže nepracujeme s textovým reťazcami ale s celými číslami, museli sme prispôsobiť metódy kríženia rodičov:

$$dieťa_1 = rodič_1 * u + rodič_2 * (1 - u)$$

Kde u je náhodne vybrané číslo z rovnomerného rozdelenia. Druhé dieťa (jedince) vzniká z rovnakých rodičov ale s novým náhodne vybraným číslom u.

Pri výbere rodičov do procesu kríženia sme implementovali ruletový výber, ktorý uprednostňoval jedince s vyššou hodnotou fitnes.

Okrem kríženia sme v GA implementovli operáciou mutácie, kedy sa jedincovi s určitou pravdepodobnosťou náhodne zmenila jedna časť chromozómu. Vďaka tomu mohol nadobúdať nové vlastnosti, ktoré sa v populácii nenachádzajú. Operátor mutácie sme museli prispôsobiť zvolenej reprezentácii chromozómov. Hodnoty v i-tej časti chromozómu bola menené rovnicou:

$$X^i_{novy} = X^i_{aktualny} + N(0, 0.05^2)$$

kde $X^i_{aktualny}$ je i-ta zložka chromozómu mutovaného jedinca X. Celý proces sa opakuje kým nie je dosiahnutý maximálny počet generácii, alebo sa nenájde dostatočne dobré riešenie.

Druhou zvolenou metaheuristikou je PSO, ktorá čerpá inšpiráciu v pohybe kŕdľov vtákov a húfovov rýb [72]. Metóda využíva kolektívnu inteligenciu kŕdľa. Každý jedinec skúma časť stavového priestoru a hľadá optimálnu polohu (riešenie optimalizačnej úlohy). V tomto zoskupení existuje vedúci jedinec, ktorý predstavuje doteraz najlepšie objavené riešenie. Ostatné jedince nasledujú toto globálne riešenie, pričom si pamätajú svoje doteraz najlepšie objavené lokálne riešenie.

V každej iterácií je okamžitá rýchlosť jedincov upravená podľa pôsobiacich síl (globálnej a lokálnej). Globálna sila pôsobí v smere najlepšieho objaveného riešenia. Lokálna sila pôsobí v smere najlepšieho predchádzajúceho riešenia:

$$x^i_{k+1} = x^i_k + v^i_{k+1}$$

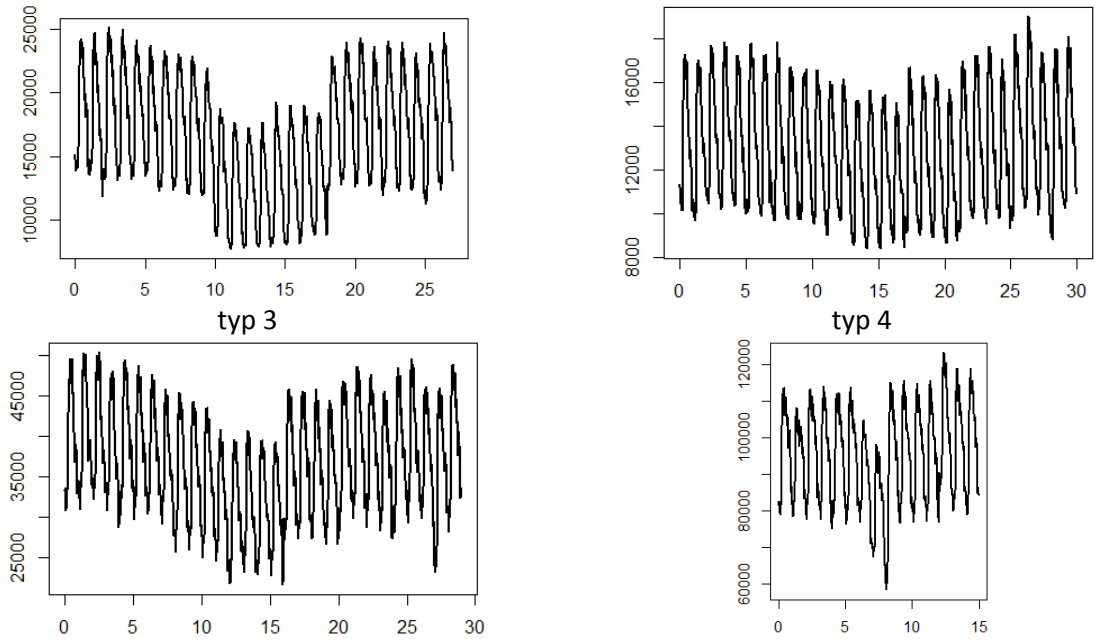
$$v^i_{k+1} = v^i_k + c_1 r_1 (p^i_k - x^i_k) + c_2 r_2 (p^g_k - x^i_k)$$

kde x^i_k je aktuálna poloha častice, v^i_k je rýchlosť častice, c_1 a c_2 sú akceleračné parametre, ktoré sme nastavili na hodnotu 2 a vektory r_1 a r_2 sú náhodne čísla z intervalu $(0,1)$.

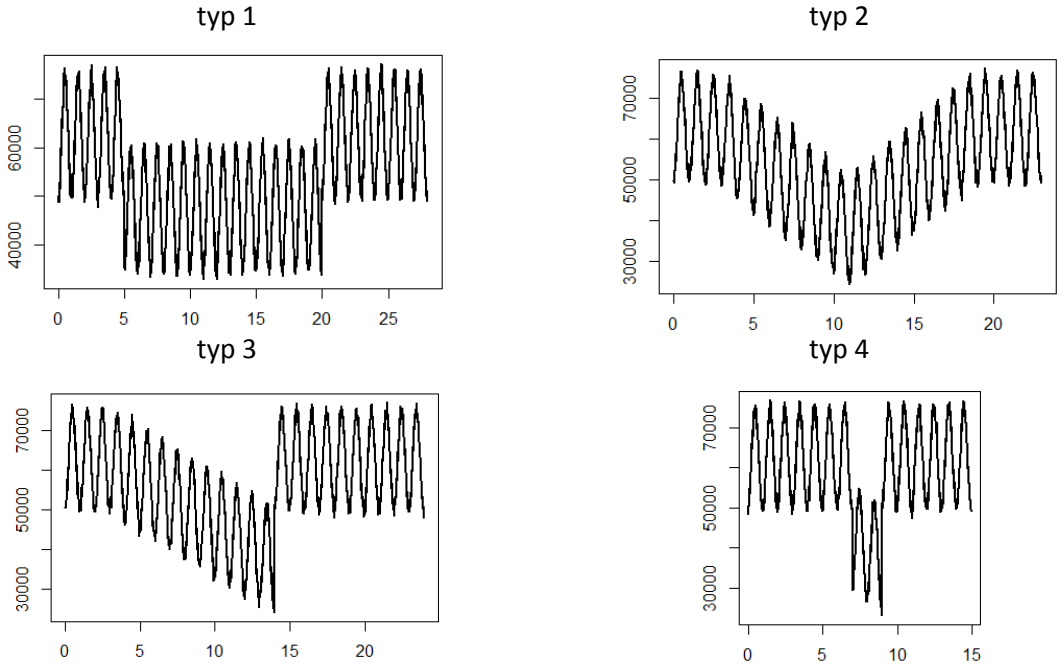
V našej implementácii sme obmedzili počet častíc, ktoré si medzi sebou vymieňajú informácie, čím sme zabezpečili aby častice vedeli uniknúť z lokálneho minima.

V dostupných dátach o spotrebe elektrickej energie sme identifikovali 4 najčastejšie sa vyskytujúce vzory zmien v odbere.

typ 1 typ 2



Následne sme vygenerovali sadu syntetických dát s dvomi úrovňami zašumenia: $\mathcal{N}(0,1000^2)$ a $\mathcal{N}(0,3000^2)$. Syntetické dáta sme použili na overenie vplyv zašumenia na presnosť predikcie navrhnutého inkrementálneho modelu.



Datasety obsahujúce vzory zmien v odbere sme použili na natrénovanie súboru 13 základných predikčných metód. Následne sme ich výstupy použili na trénovanie navrhnutého inkrementálneho modelu s využitím rôznych váhovacích metód (GA, PSO a MB). Pri trénovaní modelu sme používali 3, 4 a 5 dňové veľkosti trénovacieho okna.

Výsledky jednoznačne preukázali, že biologicky inšpirované algoritmy prekonávajú svojou presnosťou metódu MB. Priemerná predikčná chyba PSO a GA bola na reálnych datasetoch o 5,31 % menšia ako chyba MB. Na syntetických dátach bola chyba biologicky inšpirovaných algoritmov lepšia až o 5,80 %. Najlepšie výsledky boli dosiahnuté metódou PSO, ktorej priemerná chybovosť bola o 7,16 % nižšia ako priemerná chybovosť MB a o 2,84 % nižšia v porovnaní s metódou GA. Na syntetických datasetoch boli dosiahnuté výsledky veľmi podobné. Experimenty preukázali negatívny vplyv zašumenia syntetických datasetov. V niektorých prípadoch sa predikčná schopnosť testovaných metód GA, PSO a MB zhoršila o 3%. Vo všeobecnosti môžeme prehlásiť, že navrhnutý heterogénny inkrementálny model založený na učení súboru metód váhovaných biologicky inšpirovanými algoritmami, dokáže znížiť chybu predikcie v časových radoch, ktoré obsahujú postupné ale i náhle zmeny.

Po úspešnom aplikovaní biologicky inšpirovaných váhovacích algoritmov (GA – genetický algoritmus a PSO – optimalizácia rojom častíc) sme analyzovali a implementovali ďalšie potenciálne úspešné algoritmy. Výsledky sme publikovali v našom výstupe na konferencii Data a Znalosti[47]. Vo výskume sme sa zamerali na predovšetkým na rojovo orientované a ekologicky orientované algoritmy, pretože poskytujú mechanizmy na výmenu informácií medzi členmi populácie (agentmi) počas behu algoritmu, čo urýchľuje nájdenie optimálneho riešenia. Po dôkladnej analýze odbornej literatúry sme zo zoznamu [35] takmer 40 rojovo inteligentných algoritmov vybrali dva: umelú kolóniu včiel (ABC – Artificial Bee Colony) a optimalizáciu pomocou svorky sivých vlkov (GWO – Grey wolf optimizer). Uvedené metódy obsahujú mechanizmy na vyvažovanie plošného (diversification) a zosilneného (intensification) prehľadávania , ktoré sú potrebné na nájdenie riešenia v optimalizačných úlohách.

Umelá kolónia včiel – *Artificial Bee Colony* (ABC)

Metóda je inšpirovaná správaním včelích kolónií. Každá kolónia sa skladá z troch typov včiel, ktoré hľadajú zdroje potravy (nektár) resp. riešenia optimalizačnej úlohy [71]. Každý zdroj/riešenie je reprezentované ako $X_i = (x_1, x_2, x_3, \dots, x_d)$ d-dimenzionálny vektor hodnôt. Množstvo nektáru vyjadruje vhodnosť nájdeného riešenie, vyhodnoteného na základe fitness funkcie.

Prvým typom včiel sú včely zamestnané včely/zberačky (*employment bees*), ktorých úlohou je udržiavať nájdené riešenia a hľadať v jeho najbližšom okolí nové riešenie. Pohyb včely V_{ij} v okolí zdroja je definovaný nasledovne:

$$V_{ij} = X_{ij} + \phi_{ij} \times (X_{ij} - X_{kj})$$

[47] LÓDERER, M., ROZINAJOVÁ, V., EZZEDDINE, A.B.: Predikcia spotreby elektrickej energie založená na kombinácii predikčných metód. In: *Proceedings of Data a Znalosti 2015*, 2015.

kde X_k je náhodne vybrané riešenie, X_i je aktuálne riešenie, $j \in \{1,2,\ldots,N\}$ sú náhodne vybrané indexy a ϕ_{ij} je náhodne vybrané číslo z intervalu $[-1,1]$.

Ak je nové riešenie V_i lepšie ako aktuálne riešenie X_i, tak si ho včela zapamätá a staré riešenie zabudne. Ak v určitej vopred stanovenej dobe (limit) nedôjde k zlepšeniu udržiavaného riešenia, robotníčka zanechá toto riešenie a stáva sa z nej prieskumníčka.

Druhým typom sú včely nezamestnané včely (*onlooker bees*), ktoré čakajú v úli na informácie od zamestnaných včiel. Každá nezamestnaná včela si vyberá existujúci zdroj potravy, ktorého okolie bude prehľadávať. Výber zdroja je založený na ruletovom mechanizme (roulette wheel selection):

$$P_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n}$$

Zdroje s väčším množstvom nektáru sú vyberané s vyššou pravdepodobnosťou. O svojich výsledkoch informujú ostatné včely po návrate do úľa.

Posledným typom včiel sú včely prieskumníčky (*scout bees*), ktoré prehľadávajú priestor, bez ohľadu na existujúce objavené riešenia. Prieskumníčky sú nenáročné na prehľadávanie a môžu rýchlym (náhodným) spôsobom objaviť dobré riešenie. Poloha nového zdroja je určená vzťahom:

$$X_i[j] = X_{min}[j] + rand(0,1)(X_{max}[j] - X_{min}[j])$$

Kde $j \in \{1,2,\ldots,N\}$ sú náhodne zvolené indexy. Hodnoty $X_{min}[j]$ a $X_{max}[j]$ predstavujú dolné a horné ohraničenie prehľadávaného priestoru.

Optimalizáciu pomocou svorky sivých vlkov – *Grey wolf optimizer (GWO)*

GWO algoritmus je jeden z najnovších rojovo inteligentných algoritmov, ktorý bol publikovaný v roku 2014 [95]. Je inšpirovaný hierarchiou a loveckými schopnostiam svorky vlkov. Sivé vlky žijú v malých svorkách s pevne definovanou hierarchiou vládnutia, ktorá im umožňuje prežiť vo veľmi drsnom prostredí. Každý svorka sa skladá zo štyroch typov vlkov: alfa (α), beta (β), delta (δ) a omega (ω). Alfa predstavuje vodcu svorky, ktorý rozhoduje, ktorá korisť sa bude loviť. Beta a delta vlky pomáhajú alfa vlkovi pri rozhodovaní a pri riadení aktivít v svorke. Zvyšné vlky s najnižším postavením (ω) pomáhajú vyššie postaveným vlkom pri love. Počas lovu sú omega vlky riadené príkazmi vlkov alfa, beta a delta, ktoré ich smerujú k dobrej koristi (dobrému riešeniu optimalizačnej úlohy). Lov vždy pozostáva z troch hlavných častí: hľadanie koristi, obkľúčenie a napadnutie/skolenie.

Publikovaný GWO algoritmus implementuje hierarchiu vlkov a ich loveckú techniku, v ktorej hľadanie koristi predstavuje exploráciu a obkľúčenie spolu s napadnutím predstavujú zosilnené prehľadávanie v okolí riešenia. Každý vlk reprezentuje jedno riešenie v d-dimenzionálnom priestore riešení.

Algoritmus začína s N náhodne rozmiestnenými vlkmi a končí po vykonaní preddefinovaného počtu iterácii. NA začiatku každej iterácie sa v svorke vyberú tri najlepšie vlky (alfa, beta a delta) na základe hodnoty fitness funkcie. Zvyšné vlky sa stávanú omegami. Počas iterácie omega vlky upravujú svoju polohu na základe polôh naposledy vybraných vedúcich vlkov. Poloha vybraného vlka j v iterácii t je upravená nasledujúcimi rovnicami, ktoré najprv určujú dĺžku trasy, ktorú môže daný vlk prejsť a smer tejto trasy na základe aktuálnej polohy a polôh troch vedúcich vlkov:

$$\vec{D}_\alpha = |\vec{C}_1.\vec{X}_\alpha - \vec{X}| \qquad \vec{X}_1 = \vec{X}_\alpha - \vec{A}_1.\vec{D}_\alpha$$

$$\vec{D}_\beta = |\vec{C}_2.\vec{X}_\beta - \vec{X}| \qquad \vec{X}_2 = \vec{X}_\beta - \vec{A}_2.\vec{D}_\beta$$

$$\vec{D}_\delta = |\vec{C}_3.\vec{X}_\delta - \vec{X}| \qquad \vec{X}_3 = \vec{X}_\delta - \vec{A}_3.\vec{D}_\delta$$

Finálna pozícia vlka v Čase t je určená rovnicou:

$$\vec{X}_{t+1} = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3}$$

Kde $\vec{D}$, $\vec{C}$ a $\vec{X}$ sú d-dimenzionálne vektory, $\vec{X}_\alpha$ je pozícia alfa vlka, $\vec{X}_\beta$ je pozícia beta vlka a $\vec{X}_\delta$ je pozícia delta vlka. Vektory $\vec{C}_1$, $\vec{C}_2$, $\vec{C}_3$, $\vec{A}_1$, $\vec{A}_2$ a $\vec{A}_3$ obsahujú náhodne zvolené hodnoty a $\vec{X}$ je aktuálna pozícia vlka.

Vektory koeficientov $\vec{C}$ a $\vec{A}$ sú vypočítané nasledovnými rovnicami:

$$\vec{A} = 2\vec{a}.\vec{r}_1 - \vec{a}$$

$$\vec{C} = 2.\vec{r}_2$$

kde hodnoty vektora $\vec{a}$ klesajú lineárne z 2 na 0 počas behu algoritmu. Vektory $\vec{r}_1$ a $\vec{r}_2$ obsahujú náhodné hodnoty z intervalu [0, 1].

Vektory $\vec{A}$ a $\vec{C}$ slúžia na vyvažovanie plošného a zosilneného prehľadávania. Plošné prehľadávanie nastáva keď je hodnota $\vec{A}$ väčšia ako 1 alebo menšia ako -1. Vektor $\vec{C}$ môže inicializovať plošné prehľadávanie v prípade keď je jeho hodnota väčšia ako 1. Naopak, zosilnené prehľadávanie nastáva keď je hodnota |A|<1 a C<1. Hodnota vektora $\vec{A}$ klesá lineárne počas behu algoritmu, čo zaručuje plynulý prechod z plošného na zosilnené prehľadávanie. Avšak $\vec{C}$ sa generuje vždy náhodne, preto môže plošné a zosilnené prehľadávanie nastať v ľubovoľnom bode optimalizačného procesu, čo je veľmi užitočný mechanizmus na únik z lokálneho minima. Optimalizácia končí, keď algoritmus dosiahne predefinovaný počet iterácii. Nositeľom najlepšieho riešenia je alfa.

Tretí vybraný algoritmus je zo skupiny ekologický inšpirovaných algoritmov, ktoré sú inšpirované vnútrodruhovou a ako aj medzidruhovou komunikáciou a interakciou s prostredím. Ekologicky inšpirované algoritmy sú výpočtovo náročnejšie ako rojovo inteligentné algoritmy, pretože musia počítať aj s medzidruhovou komunikáciou a možným vplyvom prostredia.

Optimalizácia založená na biogeografii – *Biogeography-Based Optimization* (BBO)

Metóda čerpá inšpiráciu z ostrovnej biogeografie [119]. Základným princípom je myšlienka, že rýchlosť zmeny počtu živočíšnych druhov na ostrove je výrazne závislá na rovnováhe medzi počtom imigrujúcich a emigrujúcich druhov. Živočíšne druhy sa sťahujú z jednej lokality do inej s cieľom nájsť vhodné podmienky (SIV – *Suitability Index Variable*). Za lokalitu budeme v našom prípade považovať vektor hodnôt SIV, ktorý predstavuje možné riešenie optimalizačného problému.

Za dobré riešenia sú považované lokality s vysokým indexom vhodnosti habitatu (HSI - *Habitat Suitability Index*). Tieto oblasti sú obývané veľkým množstvom živočíšnych druhov. Naopak zlé riešenia sú považované za lokality s nízkym HSI a obýva ich menšie množstvo druhov. Rýchlosť imigrácie a emigrácie živočíšnych druhov medzi lokalitami slúži na prenos informácií a zmenu indexu SIV jednotlivých lokalít.

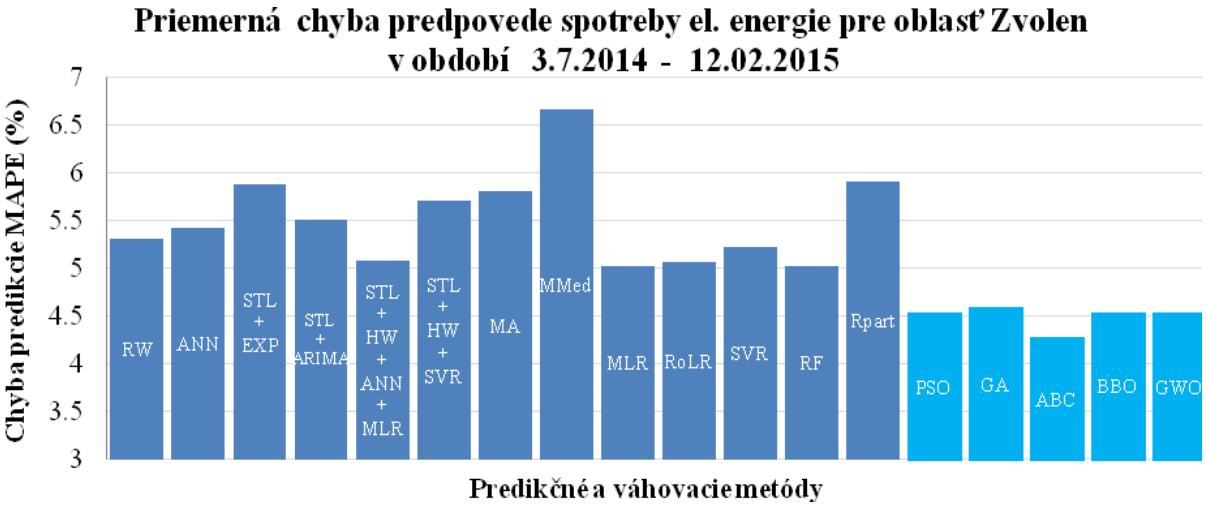
Experimenty s uvedenými algoritmami boli vykonané nad dátami zo slovenských inteligentných meračov, ktoré merali spotrebu elektrickej energie v 15-miutových intervaloch. Dáta boli použité na natrénovanie 13 základných predikčných modelov, ktoré sú opísané v predchádzajúcej kapitole. Výsledky (predikcie) základných predikčných modelov boli následne kombinované uvedenými biologicky inšpirovanými algoritmami.

Nastavenia algoritmov sú uvedené v nasledujúcej tabuľke:

Algoritmus	Názov parametru	Použitá hodnota
ABC	populácia počet zdrojov potravy počet iterácii spôsob výberu zdroja potravy limit	30 15 100 ruleta 195
BBO	populácia počet iterácii typ migrácie/imigrácie elitizmus modifikácia prostredia zachovanie poradia vyhodnocovania	50 200 lineárny 10 1 TRUE
GA	populácia	150

	naturálna selekcia kríženie mutácia elitizmus počet iterácii	50% najlepších jedincov vznikne 50% jedincov $N(0, 0.05^2)$ 5% 200
GWO	populácia počet iterácii rýchlosť koristi $\vec{a}$ $\vec{r_1}$ a $\vec{r_2}$	80 500 lineárne z 2 na 0 (0, 1)
PSO	populácia počet iterácii r_1, r_2 akcelerácia c_1 a c_2 počet informátorov poradie vyhodnocovania	40 100 (0, 1) 2 3 Náhodné

Takmer na všetkých testovaných datasetoch dosiahol najlepšie výsledky algoritmus ABC. Algoritmy PSO a GWO dosahovali podobné výsledky. V nasledujúcom grafe je znázornená priemerná absolútna percentuálna chyba MAPE pre oblasť Zvolen. V grafe sú zobrazené chyby základných predikčných modelov a chyby biologicky inšpirovaných váhovacích algoritmov.



Obr. 33. Chyby predpovede rôznych predikčných modelov.

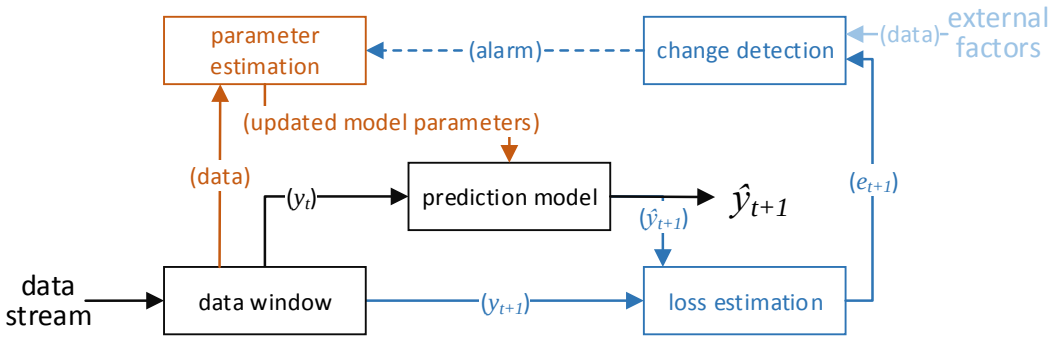
7.5 Analýza prúdu dát pre predpoveď spotreby elektrickej energie

V publikačných výstupoch[48,49] sme sa zamerali adaptívny a inkrementálny algoritmus na predikciu spotreby elektrickej energie. Algoritmus je schopný detegovať zmeny (angl. concept drift) v prichádzajúcich dátach a prispôsobovať sa im.

[48] VRABLECOVÁ, P. et al.: Stream Data Analysis for Power Demand Forecasting. *International Transactions on Electrical Energy Systems*, 2015 [odoslané].

Adaptívne metódy sa delia na tzv. slepé a informované. Slepé adaptívne metódy sa prispôsobujú vždy na príchode nových dát. Informované metódy využívajú mechanizmy na detekciu zmien v prichádzajúcich dátach a adaptujú sa iba ak je detegovaná zmena. Druhý prístup je v prostredí spracovania veľkých objemov dát, kde je obmedzený čas aj pamäť spracovania, vítanejší. Informované metódy využívajú na detekciu zmien rôzne štatisticky založené algoritmy, ktoré sledujú metriky polohy a rozptylu prichádzajúcich dát [114]. Alternatívou sú skupiny kontextovo založených modelov a pri detekcii zmeny aktuálneho kontextu (napr. typ dňa, počasie) sa vyberie vhodný vopred natrénovaný predikčný model [29].

Zaoberali sme sa krátkodobou predikciou spotreby elektriny na najbližších 24 hodín. Náš algoritmus založený na všeobecnom procese predikcie prúdových dát. Je zobrazený na obrázku 34.



Obr. 34. Predikčná metóda sa skladá z troch krokov: predpoveď (čierna), diagnostika (modrá) a aktualizácia (oranžová) (y_t je skutočná hodnota v čase t, $\hat{y}_t$ je predpovedaná hodnota na čas t, e_t je chyba predikcie v čase t).

Skladá sa z troch krokov:

1. *predpoveď* – pomocou metódy Holt-Wintersovho exponenciálneho vyrovnávania s dvojitou sezónnosťou sa predpovedá budúca hodnota časového radu (spotreba elektriny budúcich 15 min.).

2. *diagnóza* – predpovedaná hodnota sa porovná po príchode nových dát so skutočnou hodnotou. Chyba predpovede sa sleduje mechanizmom na detekciu zmien.

3. *aktualizácia* – Ak je detegovaná zmena v dátach, pretrénuje sa vyrovnávacia konštanta exponenciálneho vyrovnávania na dátach z posledných dvoch týždňov, aby bol model presnejší.

Predpoveď

Holt-Wintersovo exponenciálne vyrovnávanie s dvojitou sezónnosťou je definované formálne pomocou nasledujúcich rovníc [125]:

---

[49] VRABLECOVÁ, P., ROZINAJOVÁ, V., EZZEDDINE, A.B.: Data Stream Mining in the Power Engineering Domain. In: *Proceedings of Data a Znalosti 2015*, 2015.

Hladina $$S_t = [\alpha \frac{(y]_t}{D_{t-s_1} W_{t-s_2}}) + (1-\alpha)(S_{t-1} + T_{t-1})$$

Trend $$T_t = \gamma(S_t - S_{t-1}) + (1-\gamma)T_{t-1}$$

Sezónnosť 1 $$D_t = [\delta \frac{(y]_t}{S_t W_{t-s_2}}) + (1-\delta)D_{t-s_1}$$

Sezónnosť 2 $$W_t = \omega\left(\frac{y_t}{S_t D_{t-s_1}}\right) + (1-\omega)W_{t-s_2}$$

$$\hat{y}_{t+k} = (S_t + kT_t)D_{t-s_1+k}W_{t-s_2+k} + \phi^k(y_t - (S_{t-1} + T_{t-1}) * D_{t-s_1}W_{t-s_2})$$

V tejto definícii $\hat{y}_{t+k}$ je predpovedaná hodnota pre *k*-ty horizont. S_t je hladina vyrovnávania, ktorá je určená vyrovnávacou konštantou α . T_t je trend časového radu a je vyrovnávaný konštantou γ . D_t aW_t odzrkadľujú dvojitú (dennú a týždennú) sezónnosť. Sezónnosť je vyrovnávaná konštantami δ a ω . Pri časovom rade s 15-minútovou frekvenciou dát indexy s_1 a s_2 sú 96 (jeden deň) a 672 (jeden týždeň). Vyrovnávacie konštanty α , γ , δ a ω sú z intervalul (0,1). Taylor [125] ďalej vylepšil presnosť predikcie zavedením autoregresného modelu AR(1) rezíduí časového radu, ktorý má parameter ϕ . Ten je odhadovaný spolu s vyrovnávacími konštantami minimalizáciou štvorcovej chyby.

Diagnóza

Zmena v dátach bola detegovaná vždy, keď chyba predpovede za posledný deň presiahla 5 %. Táto hodnota bola stanovená podľa odporúčaní expertov z domény energetiky.

$$pe_t = \frac{|e_{t-95}| + \cdots + |e_{t-1}| + |e_t|}{y_{t-95} + \cdots + y_{t-1} + y_t}$$

Aktualizácia

Predpokladáme, že sezónne vyrovnávacie konštanty (δ a ω) sa s časom signifikantne nemenia. Preto zmenám v dátach je možné sa prispôsobovať úpravou vyrovnávacej konštanty hladiny α a parametru autoregresného modelu rezíduí ϕ . Ak je detegovaná zmena v dátach, ich hodnoty sú znovu odhadnuté z posledných dvoch týždňov údajov minimalizáciou štvorcovej chyby.

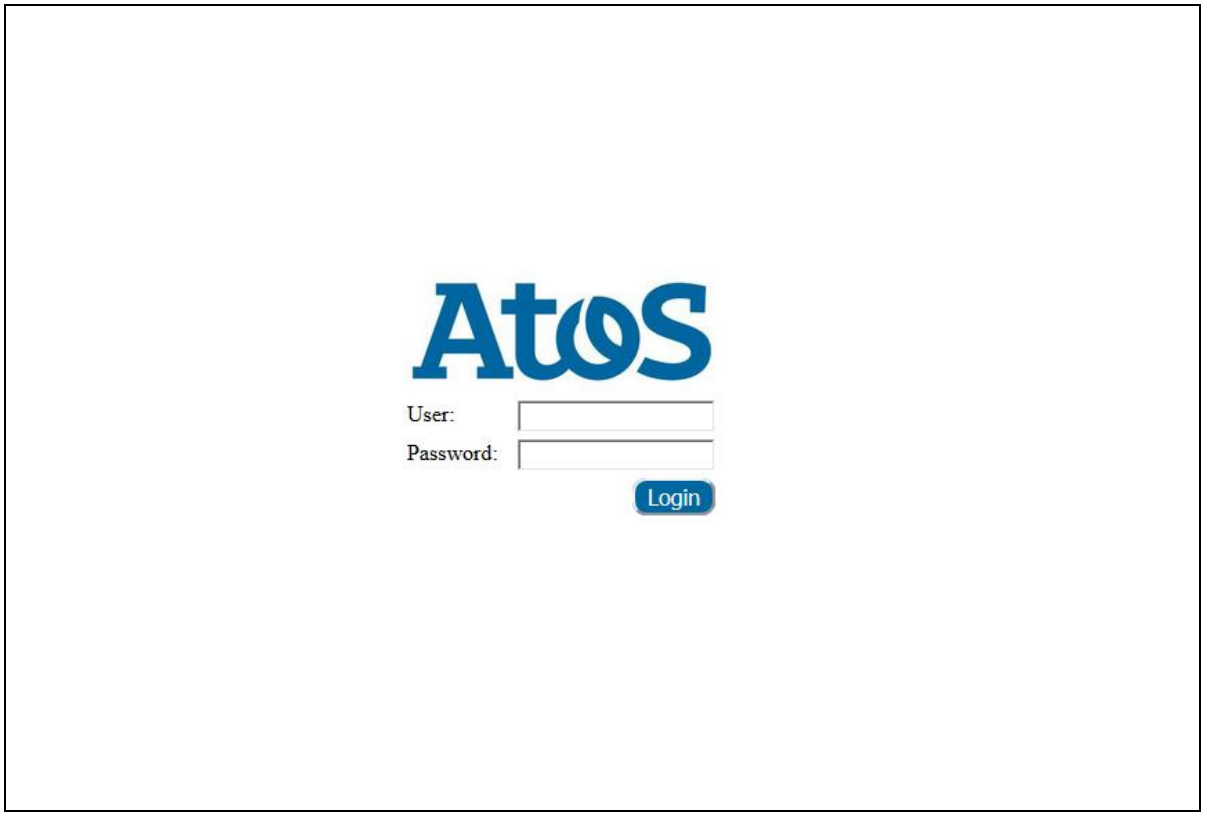
Počas overenia nás zaujímali odpovede na otázky ako ‘Do akej miery zlepšuje detegovanie zmien v dátach predikciu?’, ‘Môže byť informované prispôsobovanie tak presné ako slepé prispôsobovanie sa? Vyžaduje viac pamäťových/časových zdrojov?’ V našich experimentoch sme použili slovenské dáta. Trénovacia množina prvého modelu bola 8 týždňov. Testovacia množina bola 4 týždne. Vyhodnocovali sme chybu predikcie počas tohto obdobia a počet potrebných aktualizácií modelu. Experimenty sme vykonávali na dátach, v ktorých boli

zmeny a na dátach bez zmien. Na konci sme náš algoritmus porovnali s tradičných offline učením (slepým prispôsobovaním).

Zistili sme, že náš algoritmus vylepšil predikciu z hľadiska potrebných pamäťových a výpočtových zdrojov (o 55,35 % menej ako slepé prispôsobovanie). Táto vlastnosť je užitočná v prostredí veľkých objemov dát. Presnosť predikcií signifikantne neklesla (v priemere iba o 0,36 %). Dokázali sme udržať dennú odchýlku pod 5 %. Priemerná absolútna percentuálna chyba bola 4,40 % (4,04 % slepé prispôsobovanie). Na dátach bez zmien to bolo 3,40 % (informovaný) verzus 3,18 % (slepý).

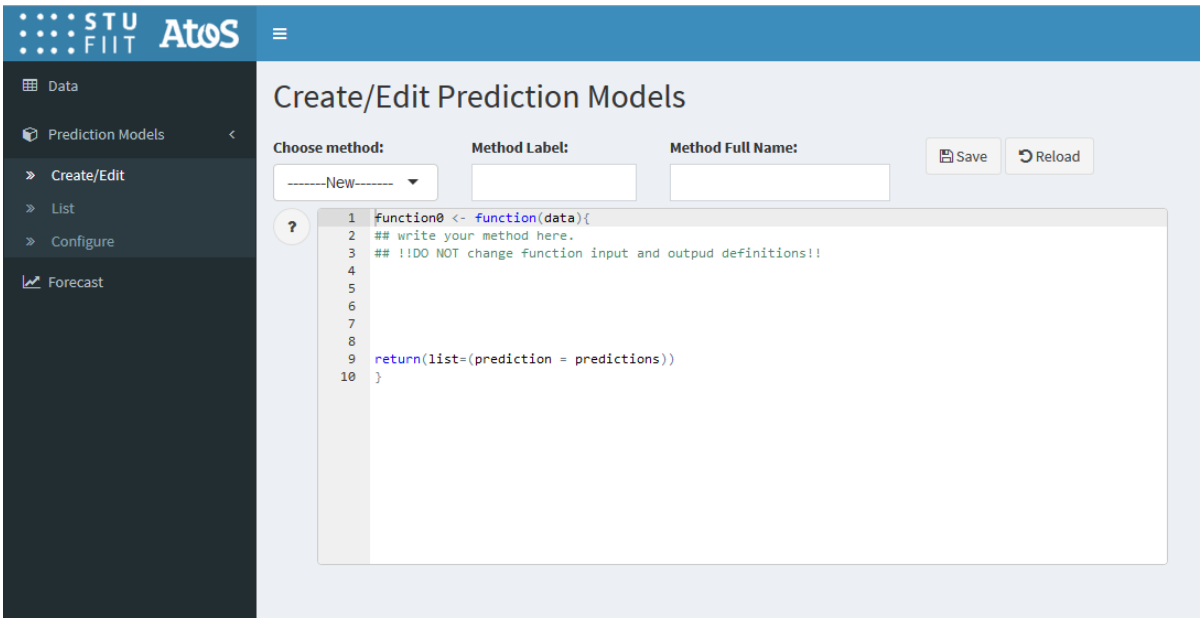
7.6 Testovacie prostredie pre overenie navrhnutých metód

Z povahy projektu počas jeho riešenia vznikla požiadavka vytvoriť jednotné testovacie prostredie v ktorom by mohli všetci výskumníci testovať svoje predikčné metódy a porovnávať ich. Bolo vhodné, aby prostredie poskytovalo možnosť testovať navrhnuté metódy za rovnakých podmienok na tých istých dátach s použitím rovnakých metrík ohodnocovania kvality predikcie. Výskumník môže do prostredia vkladať nové metódy a ladiť ich parametre. Ďalšou vlastnosťou prostredia je možnosť priamo porovnávať existujúce metódy s novými navrhnutými metódami. Výsledky predikcií je možné vyhodnocovať na základe vypočítaných hodnotách kvality, ako aj vizualizovať predikcie formou grafov. Ďalej opíšeme testovacie prostredie podrobnejšie.



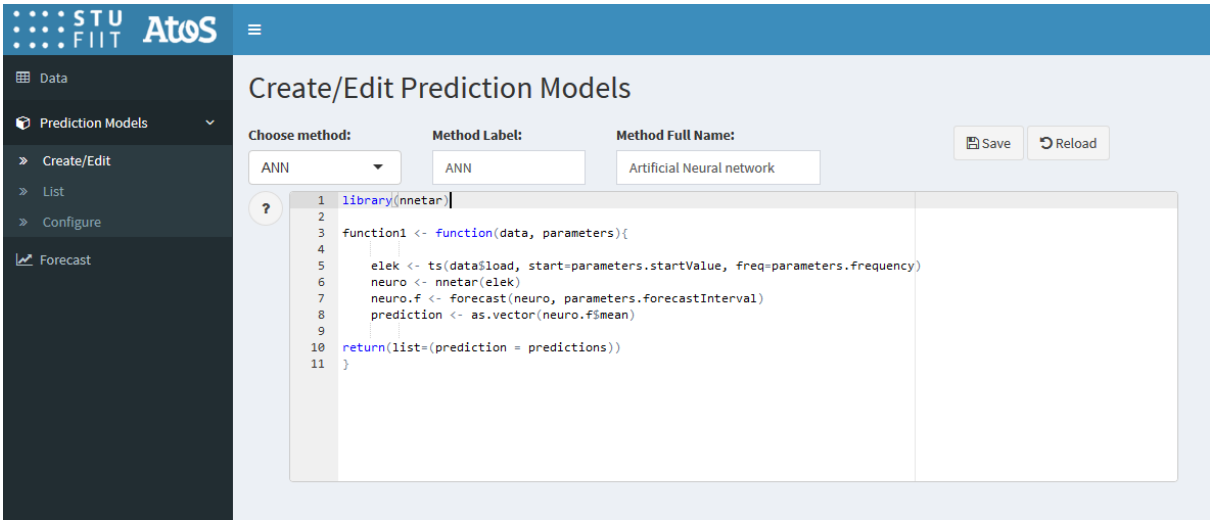
Obr. 35. Prihlasovacia obrazovka systému.

Prihlasovacia obrazovka systému je zobrazená na obrázku 35. Registrovaný používateľ musí zadať svoje prihlasovacie meno a heslo, potom môže pracovať so systémom.



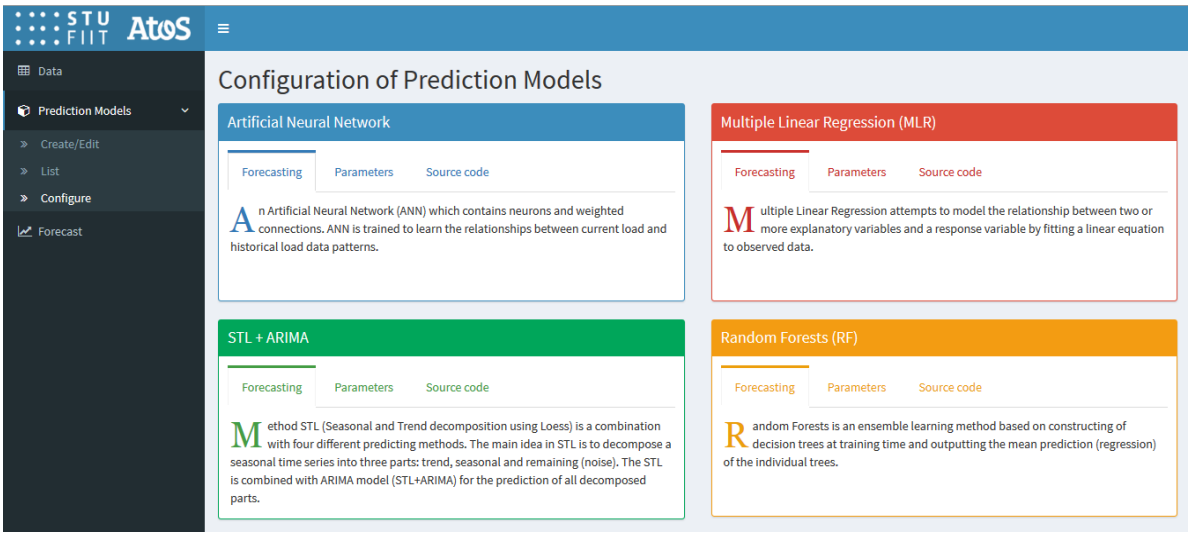
Obr. 36. Vytváranie predikčného modelu.

Funkcia na pridávanie nových predikčných metód (Obr. 36). Po výbere možnosti --------New-------- v záložke Choose method, systém vygeneruje obaľovaciu funkciu s definovanými vstupmi a výstupmi pre novo vytváranú predikčnú metódu. Používateľ musí pred uložením zadať názov metódy do poľa Method Full Name a akronym (skratku), pod ktorou bude metóda zobrazovaná v zoznamoch a grafoch. Názov metódy môže byť aj viac slovný. Akronym by mal byť čo možno najkratší (niekoľko znakov).



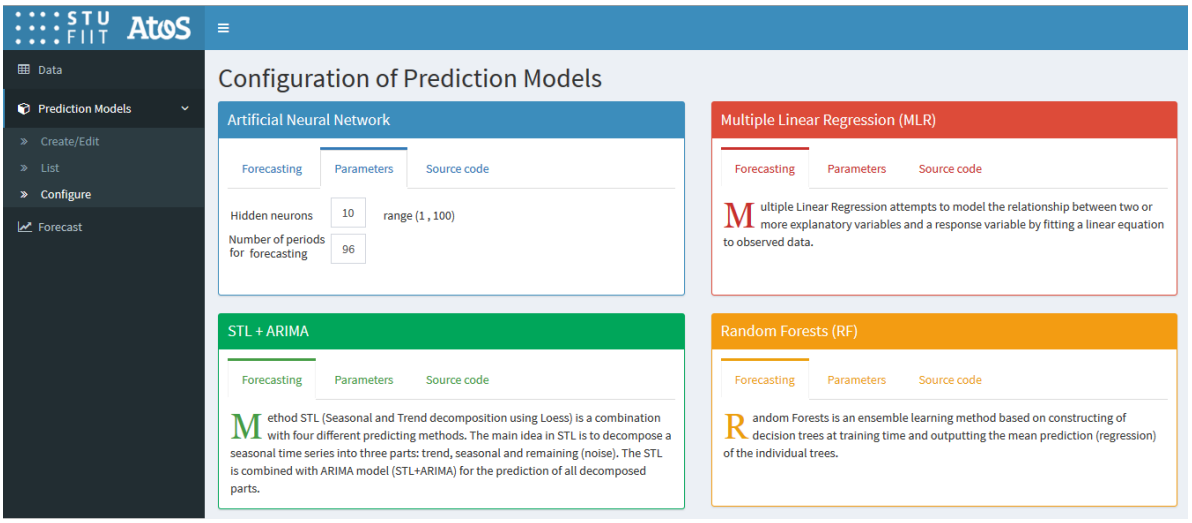
Obr. 37. Implementácia predikčnej metódy.

Na obrázku 37 je ukážka implementovanej predikčnej metódy. Súčasťou zdrojového kódu je aj volanie potrebnej knižnice (library(nnetar)), ktorá je potrebná na výpočet predikcie pomocou umelej neurónovej siete. Používateľ môže upravovať zdrojový kód metódy, ako aj jej názov a akronym. Vykonané zmeny je potrebné uložiť tlačidlom Save. V prípade, keď si používateľ nepraje uložiť vykonané zmeny, stlačí tlačidlo Reload a systém obnoví pôvodný zdrojový kód.



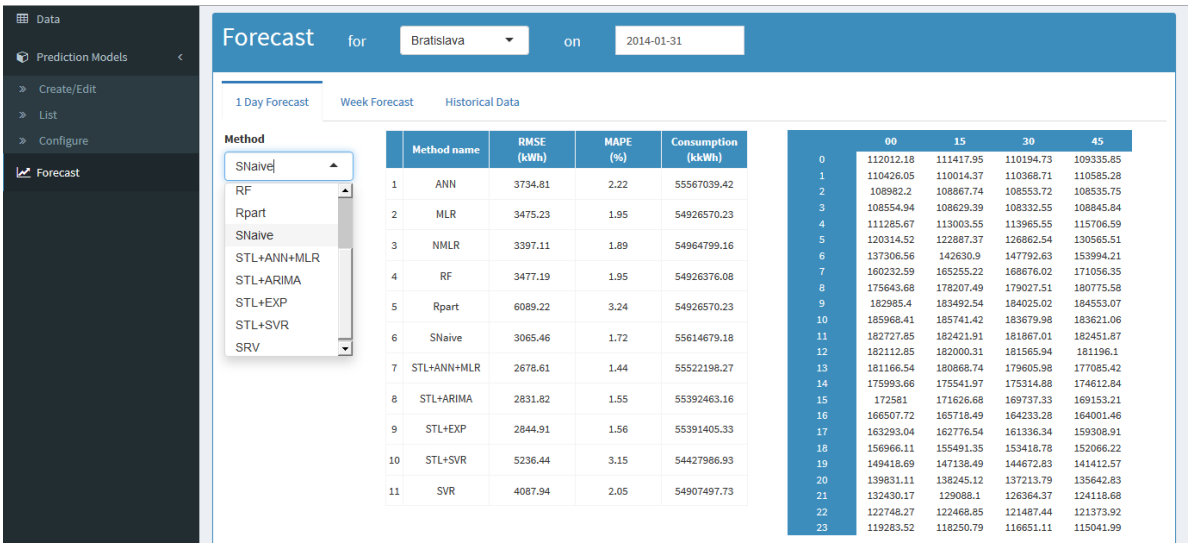
Obr. 38. Zoznam predikčných metód v testovacom systéme.

Ukážka zoznamu dostupných predikčných metód v systéme je zobrazená na obrázku 38. Každá metóda obsahuje krátky opis, ktorý informuje o vlastnostiach danej metódy.



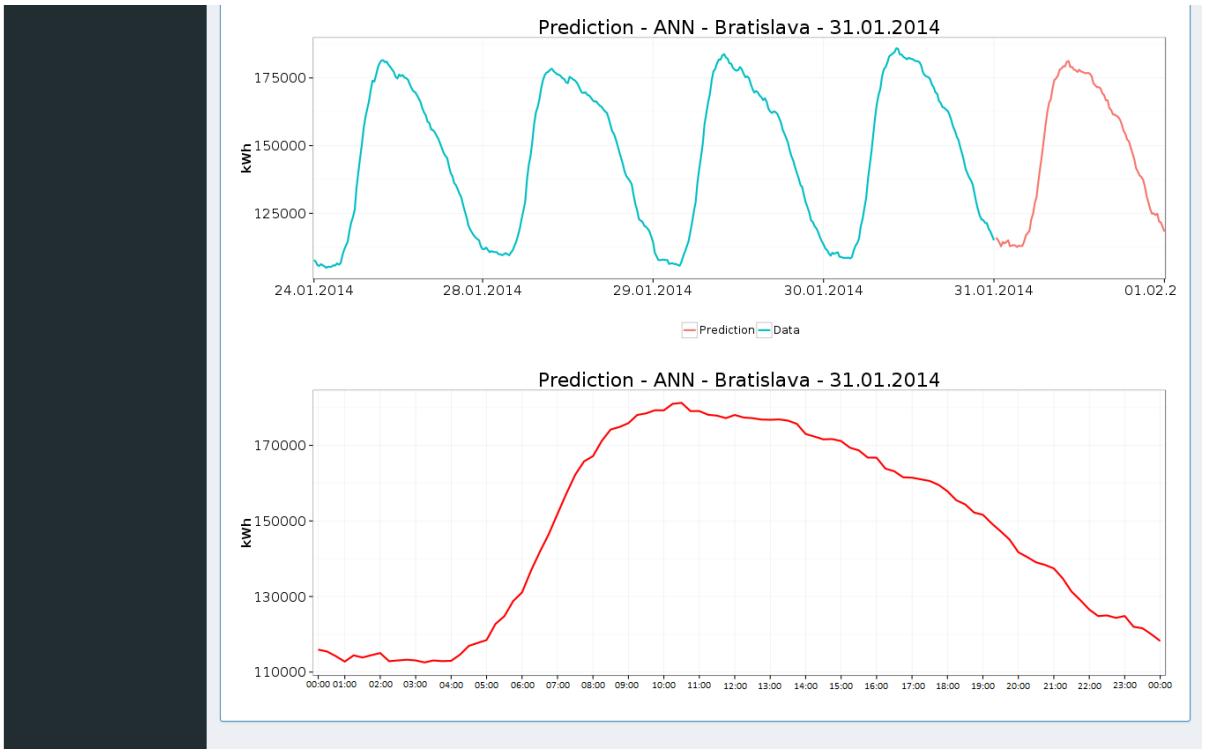
Obr. 39. Úprava parametrov metódy testovacom systéme.

Niektoré metódy obsahujú nastavenia umožňujúce zadávať vstupné riadiace parametre (Obr. 39) napríklad dĺžku predikovaného obdobia alebo počet neurónov v skrytej vrstve umelej neurónovej siete.



Obr. 40. Ukážka výstupu predikcie v testovacom systéme.

Ukážka predikcie jednodňovej predikcie (Obr. 40). Používať si v nastaveniach zvolí odberné miesto, dátum predikcie a predikčnú metódu zo zoznamu dostupných metód. Prvá tabuľka obsahuje vypočítanú priemernú hodnotu chyby (RMSE aj MAPE) jednotlivých predikčných metód za posledné obdobie. V poslednom stĺpci a nachádza predikovaná kumulatívna hodnota spotreby el. energie pre vybraný dátum. V pravej tabuľke sa nachádza podrobnejšia predikcia obsahujúca predpokladané hodnoty spotreby v 15-minútových intervaloch. Riadky predstavujú jednotlivé hodiny v dni (0 - 23), stĺpce predstavujú jednotlivé 15-miutové intervaly v rámci hodiny (00, 15, 30, 45).



Obr. 41. Grafická reprezentácia výsledkov.

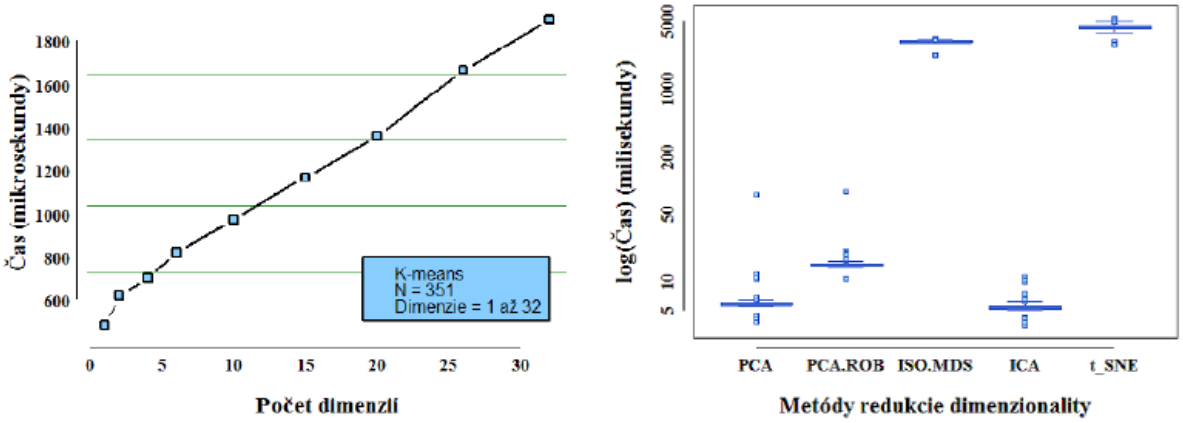
Súčasťou výsledku predikcie je aj grafická reprezentácia výsledkov. Horný graf (na obrázku 41) znázorňuje modrou farbou reálnu spotrebu el. energie z obdobia predchádzajúcich štyroch dní pred dňom, ktorý sa predikuje. Červenou farbou je znázornená samotná predikcia. Dolný graf poskytuje podrobnejší náhľad na vypočítanú predikciu.

7.7 Vplyv redukcie dimenzionality na zhlukovanie

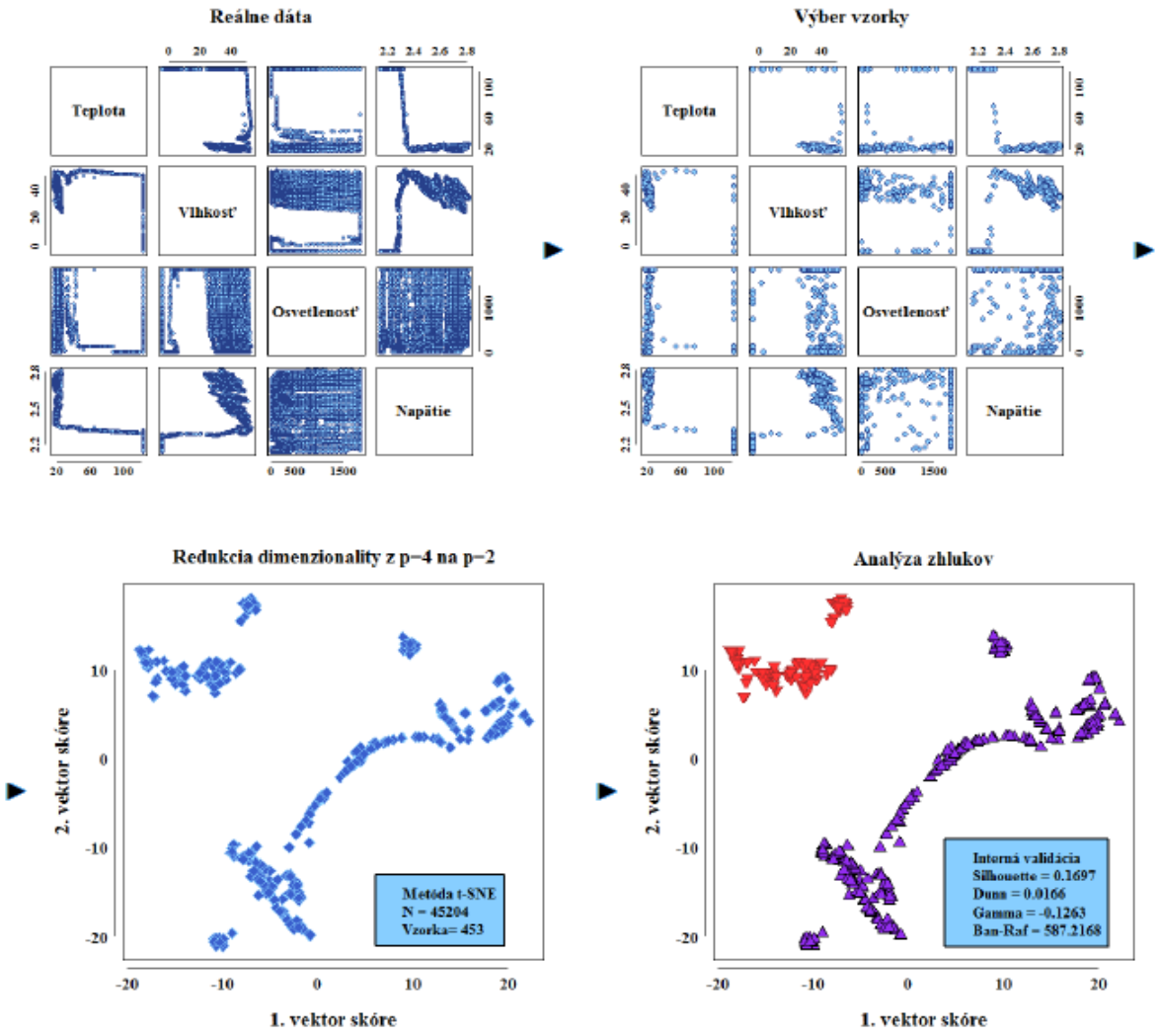
V príspevku *„Analýza vplyvu redukcie dimenzionality na zhlukovanie veľkých dátových množín“*[50] sme skúmali vplyv rôznych metód redukcie dimenzionality (PCA, robustná verzia PCA, neparametrické mnohorozmerné škálovanie, ICA, t-SNE a náhodná projekcia) na výsledky analýzy zhlukov za pomoci metód K-means, K-medoids a metódy analýzy zhlukov založenej na normálnom modeli. Cieľom je nájdenie takej metódy redukcie dimenzionality, ktorá transformuje pôvodné dáta do kompaktnejšieho podpriestoru umožňujúceho skvalitnenie výsledkov zhlukovania. Navrhujeme metodológiu, ktorou je vhodné vyhodnocovať vplyv redukcie dimenzionality na analýzu zhlukov (postup je zobrazený na obrázku 43). Analyzujeme kvalitatívne (interná validácia zhlukovania) aj kvantitatívne miery (časová náročnosť, zobrazená na obrázku 42) tohto vplyvu. Experimenty boli vykonávané na šiestich verejne dostupných reálnych dátových množinách. Výsledky ukazujú signifikantne priaznivý vplyv redukcie dimenzionality na analýzu zhlukov (po kvalitatívnej aj po

[50] LAURINEC, P., LUCKÁ, M.: Analýza vplyvu redukcie dimenzionality na zhlukovanie veľkých dátových množín. In: *Proceedings of Data a Znalosti 2015*, 2015, pp. 1–6.

kvantitatívnej stránke). Najlepšie výsledky internej validácie dosahuje náhodná projekcia a nelineárne metódy redukcie dimenzionality. Prísľubom do budúcnosti je použitie kombinácie inteligentného výberu dimenzií a metód redukcie dimenzionality.



Obr. 42. Porovnanie výpočtovej náročnosti.



Obr. 43. Postup zhlukovania s redukciou dimenzionality.

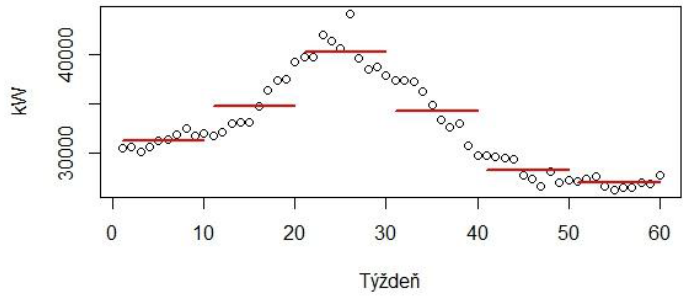
7.8 Symbolická reprezentácia časového radu pre prúdové spracovanie

Redukciou dimenzionality sme sa zaoberali aj v publikačných výstupoch[51,52]. Časové rady sú redukované na alternatívne reprezentácie, aby sa napríklad dali uchovávať ich väčšie úseky. Najznámejšie reprezentácie sú PAA, APCA, PLA.

Piecewise Aggregate Approximation (PAA) redukuje rad na sled rovnako dlhých úsekov. Každý úsek je reprezentovaný jedinou hodnotou – priemerom hodnôt daného úseku.

[51] SEVCECH, J., BIELIKOVA, M.: Repeating Patterns as Symbols for Long Time Series Representation. *Special Issue on Software Architectures and Systems for Real Time Data Stream Analytics, Journal of Systems and Software*, 2015.

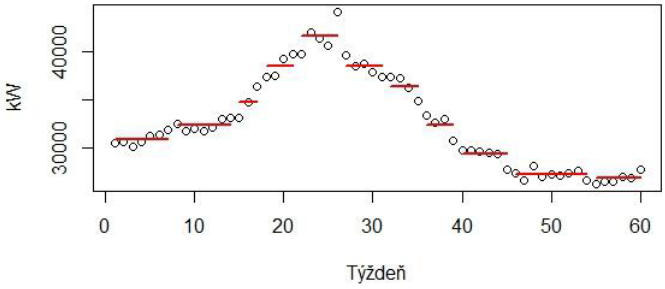
[52] SEVCECH, J., BIELIKOVA, M.: Symbolic Time Series Representation for Stream Data Processing. In: *The 1st IEEE International Workshop on Real Time Data Stream Analytics held in conjunction with IEEE BigDataSE-15*, 2015.



Obr. 44. Príklad PAA reprezentácie časového radu.

Táto reprezentácia je vhodná pre časové rady, ktorých priebeh je viac stály. Hlavnou nevýhoda tejto reprezentácie je, že všetky segmenty musia byť rovnako dlhé a v prípade časových radov, ktoré sú dlho stále, ale v krátkych úsekoch sú veľmi nepravidelné táto reprezentácia nedokáže tieto zmeny detailnejšie zachytiť [82].

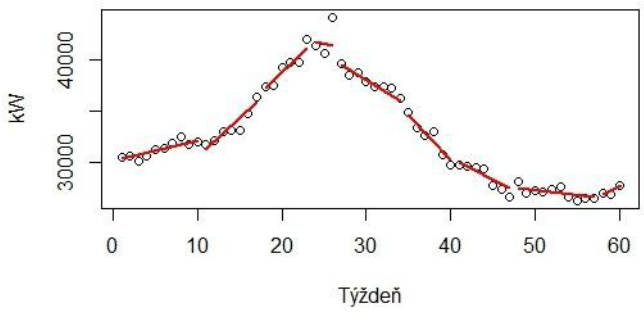
Adaptive Piecewise Constant Approximation (APCA) je založená na predchádzajúcej reprezentácii PAA. Líši sa v tom, že APCA môže obsahovať aj segmenty rôznej dĺžky [28].



Obr. 45. Príklad APCA reprezentácie časového radu.

Toto vylepšenie odstraňuje hlavnú nevýhodu PAA. Avšak APCA nie je taktiež vhodná pre reprezentáciu časových radov, ktoré obsahujú krátke úseky s veľkými zmenami, preto je vhodnejšie použiť reprezentáciu PLA.

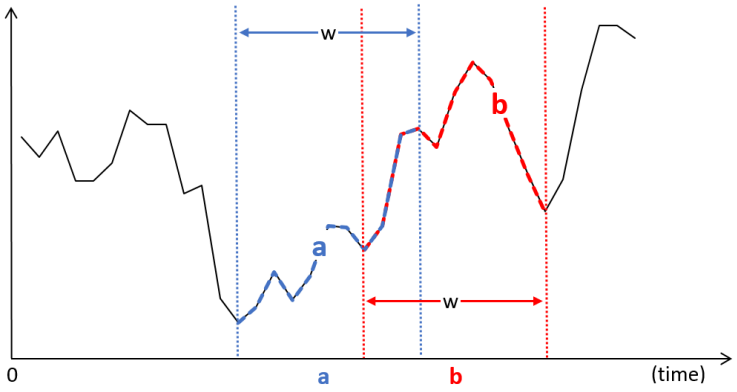
Piecewise Linear Approximation (PLA) je reprezentácia časového radu, v ktorej časový rad dĺžky reprezentujeme pomocou segmentov, ktoré, na rozdiel od PAA a APCA, nie sú definované konštantnou hodnotou, ale ľubovoľnou lineárnou funkciou. Dĺžky takýchto segmentov, podobne ako v APCA, môžu byť rôzne.



Obr. 46. Príklad PLA reprezentácie časového radu.

Výhodou oproti predchádzajúcim reprezentáciám je, že dokážeme jednoduchšie a efektívnejšie reprezentovať úseky, ktoré sa často menia. Nevýhodou oproti PAA je, že si potrebujeme pamätať dva body namiesto jednej hodnoty.

Vytváranie klasických reprezentácií časových radov je postavené na iteratívnom spracovaní, a preto nie sú vhodné pre prúdové spracovanie dát. Preto sme vymysleli novú reprezentáciu časového radu založenú na symboloch a spôsob transformácie časových radov do tejto reprezentácie. Symbolická reprezentácia odráža skutočný tvar dát. Časový rad je najskôr rozdelený na okná, okná sú normalizované, aby sa hodnoty radu nachádzali v rozpätí (0,1) a následne sú hodnoty zhlukované do symbolov. Symboly sú definované identifikátorom a násobkom a posunom voči normalizácii.



Obr. 47. Príklad symbolickej reprezentácie časového radu.

Riešenie bolo overené na viacerých dátových množinách a porovnané s existujúcimi reprezentáciami metrikou chyby rekonštrukcie.

7.9 Využitie paralelizácie pri spracovaní veľkých objemov dát

Počas práce na tomto pracovnom balíku sme sa zamýšľali aj nad využitím nástrojov na paralelné spracovanie veľkých objemov dát, a preto uvádzame ich prehľad. V konečnom dôsledku sa nepreukázala potreba ich využitia. Pre náročné výpočty pri experimentovaní počas vývoja predikčných modelov sme využívali jeden z typov distribuovaného počítania – dobrovoľnícky systém (angl. volunteer computing system) BOINC (Berkeley Open

Infrastructure for Network Computing). Pre urýchlenie výpočtového času bol navrhnutý nový algoritmus na rozvrhovanie úloh medzi výpočtovými uzlami[53].

Systém BOINC

Pre náročné výpočty počas experimentovania sme mali k dispozícii výpočtový systém uzlov založený na dobrovoľnom poskytnutí zdrojov každého počítača, ktorý sa doň zapojil. Jeho algoritmus pre rozvrhovanie výpočtových úloh je založený na princípe FCFS (angl. first-come-first-serve), teda výpočtová úloha je pridelená prvému dostupnému uzlu. Hardvérové konfigurácie výpočtových uzlov v dobrovoľníckych systémoch sa zvyčajne líšia, čo môže vnášať do výsledkov výpočtov chyby. Preto je výpočtová úloha vždy riešená na viacerých uzloch naraz a výsledok výpočtu je prijatý až po prijatí rovnakých výsledkov z predpísaného minimálneho počtu uzlov. Tento mechanizmus je jedným z hlavným rozdielov medzi dobrovoľníckym systémom a štandardnými distribuovanými výpočtovými systémami, ktoré používajú klastre.

V jednej z našich publikácii[53] bol navrhnutý algoritmus pre urýchlenie výpočtového času rozvrhnutím úloh uzlom podľa ich zatriedenia na základe ich hardvérovej konfigurácie. Nemala by teda nastať situácia, že náročná dlhá výpočtová úloha na konci výpočtového procesu bude pridelená uzlu, ktorý počíta pomaly. Algoritmus CBSA (angl. class-based scheduling algorithm) bol schopný zlepšiť výpočtový čas na konci procesu približne o 40 %.

8 Vývoj nových analytických algoritmov pre neštrukturalizované dáta

Okrem analýzy štruktúrovaných energodát sme sa v tomto balíku zaoberali aj analýzou neštruktúrovaných dát, najmä metódami pre spracovanie prirodzeného jazyka v kontexte mnohojazyčnej Európy. Zamerali sme sa na slovenský jazyk. V nasledujúcich častiach opisujeme základné metódy spracovania textu, špecifiká slovenčiny a opisujeme naše publikované výstupy, ktoré sa zameriavali na analýzu neštruktúrovaných dát (textových alebo obrazových) všeobecne ako aj analýzu veľkých objemov neštruktúrovaných dát.

8.1 Spracovanie prirodzeného jazyka z pohľadu mnohojazyčnej Európy a metódy automatického prekladania

Spracovanie prirodzeného jazyka predstavuje súbor metód a prístupov, ktorými je možné vytvoriť komunikačný prostriedok medzi človekom a počítačom, teda umožniť počítaču pochopiť text napísaný jazykom človeka. Práve preklenutie priepasti medzi svetom symbolov, z ktorých sa skladá text, a svetom ľudí, ktorí tieto symboly interpretujú prostredníctvom prirodzeného jazyka, je stále veľkou výzvou [25]. Keďže ide o obojsmernú komunikáciu (aj od človeka k stroju a aj od stroja k človeku), oblasť sa delí na dve základné

[53] UJHELYI, M., LACKO, P., PAULOVIČ, A.: Task Scheduling in Distributed Volunteer Computing Systems. In: *IEEE 12th International Symposium on Intelligent Systems and Informatics*, 2014, pp. 111–114.

časti: porozumenie prirodzeného jazyka (kde sa počítač snaží identifikovať text napísaný človekom) a generovanie prirodzeného jazyka (kde stroj vytvorí text, pomocou ktorého odovzdáva informácie človeku). Ide o rozsiahlu oblasť s množstvom nevyriešených problémov [26]. Využívajú sa tu prevažne metódy strojového učenia (napr. rozhodovacie stromy, asociačné pravidlá, štatistické modely, pravdepodobnostné rozhodovanie) a lingvistické metódy (napr. gramatiky jazykov).

Pri spracovaní prirodzeného jazyka je potrebné zohľadniť formát, v akom sú informácie dodané. Najčastejšie ide o čistý text (tzv. neštruktúrovaný text, ktorý možno označiť pojmom veľké dáta), ale môže ísť aj o štruktúrovaný text (napr. záznamy z databázových tabuliek), reč (napr. zvukový záznam, kde je potrebné použiť nástroj na prevod reči do textu) alebo obrázok (obrazový záznam, kde je potrebné použiť nástroj na rozpoznávanie znakov v prípade, že ide o text alebo nástroj na identifikovanie objektov, ak ide o grafické objekty) .

8.1.1 Fázy spracovania prirodzeného jazyka

Keďže je oblasť spracovania prirodzeného jazyka rozsiahla, realizuje sa vo viacerých krokoch. Pred samotným spracovaním jazyka je fáza predspracovania, v ktorej dochádza k základnej analýze textu. Zvyčajne ide o procesy, kedy sa vstupný text rozdelí na menšie časti (napríklad odseky, vety, slová), identifikujú sa špeciálne prvky textu (napríklad nadpisy, znaky ukončenia viet a pod.) a s takto spracovaným textom sa ďalej pracuje. Na tieto procesy existuje viacero nástrojov, ktoré sú však najmä pre rozšírené jazyky (napr. anglický jazyk je pomerne dobre pokrytý nástrojmi na spracovanie, zatiaľ čo v slovenčine je nástrojov veľmi málo a sú menej presné). Rovnako to platí aj pri nástrojoch, keď porovnáme WordNet pre anglický jazyk, EuroWordNet zameraný na najpopulárnejšie európske jazyky [77, 132, 133] a slovenský WodNet, ktorého Pilotná verzia[54] vznikla na Jazykovednom ústave Ľudovíta Štúra Slovenskej akadémie vied a v súčasnosti obsahuje okolo 25 tisíc synsetov s prepojením na anglický ekvivalent.

V podstate môžeme spracovanie jazyka rozdeliť do piatich základných fáz na základe úrovne analýzy [3, 104]:

1. morfologická analýza,
2. lexikálna analýza,
3. syntaktická analýza (parsovanie),
4. sémantická analýza,
5. pragmatická (kontextová) analýza.

Morfologická analýza textu predstavuje rozdelenie textu na morfologické jednotky – morfémy. Morféma je najmenšia časť slova s nejakým významom. Výstupom morfologickej analýzy je identifikácia gramatických kategórií jednotlivých slov. Lexikálna analýza

[54] http://korpus.juls.savba.sk/WordNet.html - WordNet - Slovenská verzia

identifikuje značky, ktoré sa nachádzajú v slovníku konkrétneho jazyka. Slovníkom môže byť ľubovoľný zoznam názvov/databáza (napríklad krstné mená, priezviská, rastliny, živočíchy), zoznam entít (populárne osobnosti, mestá, štáty) alebo slovník jazyka (lexikálna databáza slovných výrazov ako je WordNet alebo Slovenský národný korpus). Lexikálny analyzátor (označovaný tiež ako scanner) je konečný automat, ktorý rozdelí vstupnú postupnosť znakov na lexikálne jednotky (napr. čísla, kľúčové slová, identifikátory). Súčasťou môže byť normalizácia značiek (prevod na rovnaký formát), odstránenie stop slov, lematizácia (prevod slov na základný tvar) alebo stemming (prevod slov na koreň, čiže odstránenie predpôn a prípon). Výstupom lexikálnej analýzy je identifikácia slovných druhov. Syntaktická analýza (parsovanie, parsing) je proces analýzy sekvencie značiek (tokenov) na určenie ich gramatickej štruktúry s ohľadom na danú formálnu gramatiku (napr. skloňovanie). Výstupom syntaktickej analýzy je identifikácia syntaktických jednotiek: vetných členov a fráz. Sémantická analýza priraďuje jednotlivým pojmom v texte význam. Pragmatická analýza spája význam celej vety do kontextu.

V jednotlivých fázach spracovania textu vznikajú problémy súvisiace s nejednoznačnosťou významu. Jeden pojem (slovo alebo slovné spojenie) môže mať viacero rôznych významov a taktiež sa dá zapísať viacerými slovami (napr. pri použití synoným). V prirodzenom jazyku môže rovnako znejúce slovo vystupovať vo forme viacerých slovných druhov. Napríklad slovenské slovo rev môže byť vo význame podstatného mena alebo slovesa v rozkazovacom spôsobe. Tieto nejednoznačnosti sa vyskytujú aj v iných jazykoch (napríklad anglické slovo *patient* môže mať nielen rôzny význam, ale aj odlišný slovný druh: *patient* ako podstatné meno pacient, *patient* ako prídavné meno trpezlivý a pod.).

8.1.2 Spracovanie slovenského jazyka

Tak ako každý jazyk, aj slovenčina má svoje špecifické črty, a preto je potrebné pristupovať k spracovaniu jazyka špecifickejšie. V prvom rade ide o jazyk s bohatou morfologickou štruktúrou. Pri predspracovaní anglických textov sa morfologická analýza nahrádza procesom identifikácie slovných druhov, keďže angličtina má morfologický systém pomerne chudobný, avšak v slovenčine je proces identifikácie koreňa slova náročnejší. S tým súvisí aj to, že na slovenský jazyk nie je toľko nástrojov na spracovanie textu (nástroje na identifikáciu slovných druhov, zjednoznačňovanie slov a pod.), resp. nedosahujú takú presnosť, ako nástroje pre populárnejšie jazyky [74]. Nástroje sú zvyčajne vyvíjané na univerzitách. Na fakulte informatiky a informačných technológií vzniklo niekoľko projektov na spracovanie slovenského jazyka, ktoré je možné nájsť na webe http://text.fiit.stuba.sk. Ide o študentské projekty, ktoré sú použiteľné formou služieb. K dispozícii je nástroj na lematizáciu, nástroje na určovanie slovných druhov slov, nástroj na doplňovanie diakritiky alebo nástroj na extrakciu názvoslovných (pomenovaných) entít.

Proces lematizácie je v slovenčine pomerne náročný, keďže ide o jazyk, ktorom je výrazné zastúpenie skloňovaných slov v porovnaní s počtom slov, ktoré sa vyskytujú v základnom tvare. Keďže skloňovanie slov vychádza z viacerých pravidiel, môžu že stať, že v určitých

prípadoch je pri odvodzovaní nejaký znak navyše, alebo nejaký znak chýba. V slovenčine sa zmeny tvaru slov zvyčajne uskutočňujú na základe skupín pravidiel podľa vzoru konkrétneho slova. Napríklad mužské podstatné mená používajú vzory chlap, hrdina, dub a stroj. Taktiež majú svoje vzory podstatné mená ženského a stredného rodu, iné vzory sú pre prídavné mená a pod. A k tomu všetkému treba zohľadniť aj to, že existujú mnohé výnimky z pravidiel, keďže slovenčina preberá aj cudzie slová a ich skloňovanie je často špecifické. Keďže ide zvyčajne o rozdiely niekoľkých znakov (príkladom môže byť spodobovanie), tak je možné použiť algoritmy na výpočet podobnosti slov. K najznámejším algoritmom na určovanie podobnosti slov patria metóda N-gramov, Hammingova vzdialenosť, Levenshteinova vzdialenosť, Jaro vzdialenosť a Jaro-Winkler vzdialenosť.

Metóda podobnosti N-gramov porovnáva reťazce na základe počtu podobných n-tíc. Za číslo n môžeme zvoliť ľubovoľné kladné celé číslo (zvyčajne 2 alebo 3), slová rozdelíme po znakoch o dĺžke n a porovnáme počet zhodných reťazcov, ktoré takto vzniknú. Výhodou použitia n-gramov je, že mnohé databázové systémy majú zabudované algoritmy na ich indexovanie.

Levenshteinova vzdialenosť počíta počet úprav, ktoré je potrebné vykonať, aby sa dva reťazce zhodovali. Medzi tieto úpravy patrí vloženie znaku, odstránenie znaku a výmena znaku. Teda ak sa dve slová líšia v jednom znaku, ich Levenshteinova vzdialenosť je číslo jeden. Špeciálne v lingvistike ide o populárnu metódu. Rozšírením tejto metódy o operáciu výmeny dvoch znakov označujeme metódu s názvom Demerau-Levenshteinovu vzdialenosť.

Hammingova vzdialenosť počíta počet pozícii znakov, kde sa líšia porovnávané slová. Keďže ide o akúsi zjednodušenú verziu predchádzajúcich algoritmov, jej výsledky sú menej presné. Výhoda je pri použití na slová, ktoré majú rovnakú dĺžku a v efektívnosti výpočtu.

Jaro vzdialenosť je výhodná pre porovnávanie názvoslovných entít. Výhoda oproti Levenshteinovej vzdialenosti spočíva v tom, že je zohľadnená aj dĺžka reťazca (čiže zámena v dlhšom slove je menej významná v porovnaní so zámenou v kratšom slove). Výsledkom však nie je počet (číslo), ale percentuálna podobnosť.

Jaro-Winkler vzdialenosť je rozšírená metóda Jaro vzdialenosti, ktorá zvýhodňuje prvé znaky. Využíva sa pri porovnávaní mien názvoslovných entít, kde je oveľa väčšia pravdepodobnosť, že nastane chyba v neskorších znakoch v slove.

Všetky tieto metriky je možné využiť v procese lematizácie a zlepšiť tak pravdepodobnosť určenia lemy slova, čo je v prípade jazykov s bohatou morfológiou veľký problém.

V našej práci[55] sa využíva spracovanie prirodzeného jazyka pre automatickú extrakciu hierarchických vzťahov medzi kľúčovými slovami textu. Hierarchické vzťahy sú základom

[55] VRABLECOVÁ, P., ŠIMKO, M.: Supporting Semantic Annotation of Educational Content by Automatic Extraction of Hierarchical Domain Relationships. *IEEE Transactions on Learning Technologies*, 2015 [posudzované].

ontológií, ktoré slúžia aj pre pochopenie prirodzeného jazyka strojmi. Práca teda prispieva k riešeniu problému tvorby reprezentácie poznatkov o doméne, ktorá je opísaná pomocou množiny textov, s cieľom pokročilého spracovania informácií. Prezentuje generizovateľný prístup k extrakcii ontologických hierarchických vzťahov z rozsiahleho korpusu textov, ktorý je uplatniteľný aj pri spracovaní textových dát v doméne energetiky. Metóda je založená na štatistickom spracovaní. Kľúčové slová sú reprezentované vektormi slov, ktoré sa vyskytujú okolo nich v textových dokumentoch. Následne je vyhodnotená kosínusová podobnosť týchto vektorov a odvodené vzťahy podobnosti kľúčových slov. Z nich sú navrhnutým algoritmom extrahované hierarchické vzťahy podľa toho, ako sú slová priradené k stromovej štruktúre doménových dokumentov. Experimenty boli vykonané na slovenských výučbových materiáloch.

8.1.3 Strojový preklad

Jednou z popredných úloh spracovania prirodzeného jazyka je strojový preklad – teda automatický preklad textu z jedného jazyka do druhého. Klasické prístupy štandardne vychádzali zo slovníkových metód (preklad slov podľa slovníkov). Nakoľko je vo vetách poradie slov kľúčové a každý jazyk má zvyčajne špecifický slovosled a pravidlá na usporiadanie viet (vetná štruktúra), takýto spôsob jazykového prekladu nie je postačujú. Okrem toho aj tu nastáva problém s nejednoznačnosťou slov (napríklad spomínané slovo *patient*). Google Translate, známy prekladač firmy Google, je jeden z prvých nástrojov, ktorý využíva pri preklade veľké dáta. Autori navrhli novú metódu strojového učenia, pri ktorej sa prekladací nástroj učí prekladať z paralelných textov v rôznych jazykoch Mayer et al., 2013). Na začiatok autori do systému vložili obrovské sady textu (paralelné korpusy) a nechali, aby nástroj sám pomocou slovníku identifikoval, ako treba vety transformovať z jedného jazyka do druhého. V prípade jazykov, ktorý má k dispozícii veľké sady textu, je preklad kvalitný. Keďže slovenský jazyk nie je natoľko rozšírený, preklad napríklad medzi slovenčinou a angličtinou je slabší (lebo len jeden z dvojice: zdrojový jazyk a cieľový jazyk) sú z kategórie svetových jazykov. Keď si zoberiem ako oblasť Európu a preklad medzi európskymi jazykmi, problémom môže byť, ak budú obidva jazyky málo populárne. Ak by sme napríklad chceli použiť takýto strojový preklad na preklad zo slovenčiny do maďarčiny, tak by kvalita bola slabá, keďže slovensko-maďarské elektronické paralelné texty je náročné získať. V praxi sa zvykne používať metóda prechodu cez tretí jazyk, kedy sa najskôr preloží text z jedného menej rozšíreného jazyka do známeho jazyka (napr. zo slovenčiny do angličtiny) a v druhom kroku z tohto jazyka do cieľového jazyka (z angličtiny do maďarčiny). Tu však treba individuálne posúdiť, či dvojnásobný preklad nespôsobí ešte viac chýb, ako preklad medzi dvomi menej známymi jazykmi. Na záver možno zhrnúť, že automatizovaný strojový preklad je medzi populárnymi jazykmi relatívne kvalitný, ale v mnohojazyčnej Európe sa používa viacero jazykov, ktoré sú z hľadiska tejto úlohy slabo pokryté.

8.2 Spracovanie metadát o používateľoch

V článku[56] predstavujeme doménovo nezávislý model používateľa uvažujúci personalizované váhovanie metadát na rozličných úrovniach abstrakcie. Takto vieme záujmy používateľov modelovať a porovnávať na viacerých úrovniach. To znamená v prvom rade sledovať typ metadát, ktoré používateľ implicitne pokladá za dôležité a až následne sledovať preferenciu akú prikladá konkrétnym metadátam. Možnosť porovnávať preferenice používateľov na viacerých úrovniach predstavuje výhodu pri aplikovaní na veľkých dátových kolekciách a pri veľkom množstve používateľov, pretože prvotné porovnanie odfiltruje dostatočné množstvo používateľov ešte pred fázou poddrobnejšieho porovnania na nižšej úrvovni abstrakcie. Rovnako pri modelovaní preferencií pomocou navrhnutého prístupu dokážeme na základe viacerých úrovní záujmu vyjadriť používateľom preferované koncepty aj v rozsiahlych dátových kolekciách.

V článku[57] sa zaoberáme spracovaním prúdu dát a predikciou krátkodobých akcií používateľa. Súčasťou riešenia je vysporiadanie sa s meniacimi sa vlastnosťami dát v čase, nevyváženosťou dát, či veľkými objemami prichádzajúcich dát. Navrhované riešenie sa zaoberá aktuálnou otázkou straty používateľa, pričom sa jej dotýka z inovatívneho pohľadu krátkodobého správania používateľa.

8.3 Spracovanie obrázkov

V prípade spracovania médií, akými sú napríklad obrázky, je potrebné do procesu predspracovania pridať krok, ktorým sa tento formát prevedie na textový formát.

V prípade znakov vo forme obrázku máme k dispozícii nástroje na prevod do textu, tzv. optické rozoznávanie znakov (angl. Optical Character Recognition, OCR). Rozpoznávanie znakov je využité aj pri rozpoznávaní adries v automatickom triedení pošty.

Náročnejšia situácia je, keď potrebujeme identifikovať objekty, ktoré sú na obrázku. Prvým krokom pri analýze obrazu je jeho segmentácia. Segmentáciou môže byť odstránené pozadie obrazu a ponechané len dôležité časti. Príklad využitia obrazových materiálov je vylepšenie výsledkov vyhľadávania v internetových obchodoch, kde sa zákazník rozhoduje, či zakúpi výrobok najmä podľa jeho obrázkov. Zo segmentovaných obrázkov je možné vydolovať vlastnosti ako rozmery, jas, kontrast, farebnosť. Objektom na obrázkoch je potom možné priradiť farbu, textúru alebo tvar. Vyhľadávanie podľa obrázkov a vlastností ako farba, veľkosť je implementované aj v popredných webových vyhľadávačoch. Identifikácia objektov na obrázkoch je využiteľná v automatizovaných filtroch, napr. na filtrovanie nevhodného obsahu.

---

[56] KAŠŠÁK, O., KOMPAN, M., BIELIKOVÁ, M.: User Preference Modeling by Global and Individual Weights for Personalized Recommendation. *Acta Polytechnica Hungarica*, 2015.

[57] KAŠŠÁK, O., KOMPAN, M., BIELIKOVÁ, M.: Students' Behavior in a Web-Based Educational System: Exit Intent Prediction. *Engineering Applications of Artificial Intelligence*, 2015.

Existujú nástroje, ktorými je možné automatizovane anotovať obrázky v podobe kľúčových slov. Náš nástroj ANNOR[58] (Automatic image aNNOtation Retriever) využíva metódy strojového učenia pri rozpoznávaní objektov na obrázku. V ich prístupe vychádzajú z postupu, ako by anotoval obrázok používateľ – rozpoznávanie tvarov na základe okrajov a ich extrakcia z plátna. V prvom kroku teda identifikujú na obrázku objekty a rozdelia ich na samostatné časti. Následne sa snažia každej časti identifikovať kľúčové slová, ktoré by opísali, o aký objekt ide. Kvalita anotácie závisí od dostupnej vzorky anotovaných obrázkov, ktoré sa využívajú vo fáze učenia (objekty, ktoré sa nachádzajú častejšie v databáze trénovacích objektov, sa podarilo anotovať s lepšou presnosťou).

9 Záver

V rámci balíka „Architektúra veľkých dát , štruktúra dát a ich analýza“ sme sa zamerali na základný výskum v oblasti analýzy veľkých dát. V rámci pomerne krátkeho trvania projektu sme napísali 20 pôvodných vedeckých článkov, z ktorých 17 je publikovaných, resp. prijatých na publikovanie a 3 sú v etape posudzovania. Niekoľko z nich (6) je zaradených do publikačnej kategórie A.

V rámci výskumu sme navrhli viacero nových prístupov a metód riešenia zvolených problémov.

Na skvalitnenie predikčných schopností sme navrhli a overili heterogénny inkrementálny model učenia súboru metód na predikciu časových radov v situáciách, keď sa v časovom rade vyskytujú postupné alebo náhle zmeny. Využili sme na to vhodné biologicky inšpirované metaheuristiky, kde sa nám pri riešení optimalizačných úloh výpočtu optimálnych váh podarilo dosiahnuť veľmi dobré výsledky. Zamerali sme sa predovšetkým na rojovo orientované a ekologicky orientované algoritmy, pretože poskytujú mechanizmy na výmenu informácií medzi členmi populácie (agentmi) počas behu algoritmu, čo urýchľuje nájdenie optimálneho riešenia.

Navrhli sme adaptívny a inkrementálny algoritmus na predikciu spotreby elektrickej energie, ktorý je schopný detegovať zmeny (angl. concept drift) v prichádzajúcich dátach a prispôsobovať sa im.

Vytvorili a overili sme jednotné testovacie prostredie, umožňujúce testovanie, overovanie a porovnávanie zvolených predikčných metód. Prostredie poskytuje možnosť testovať navrhnuté metódy za rovnakých podmienok na tých istých dátach s použitím rovnakých metrík ohodnocovania kvality predikcie. Výsledky predikcií je možné vyhodnocovať na základe vypočítaných kritérií kvality, ako aj vizualizovať predikcie formou grafov.

[58] KURIC, E., BIELIKOVA, M.: ANNOR: Efficient image annotation based on combining local and global features. *Computers & Graphics*, vol. 47, 2015, pp. 1–15.

Navrhli sme nové metódy predikcie v energetike s akcentom na charakteristiky dát po implementovaní inteligentných meračov, kde sme sa zamerali na inkrementálne modely. Využili sme pritom algoritmy založené na podporných vektoroch, ktoré sa javia ako vhodné metódy pri využití paradigmy distribuovaného spracovania.

Navrhli sme nový algoritmus na rozvrhovanie úloh medzi výpočtovými uzlami, vedúci k zrýchleniu paralelných distribuovaných výpočtov.

Navrhli a overili sme inkrementálny a adaptívny prístup predikcie spotreby elektrickej energie s modifikáciou Holt-Wintersovho exponenciálneho vyrovnávania.

Skúmali sme vplyv rôznych metód redukcie dimenzionality na analýzu zhlukov a navrhli metodológiu ako viesť experimenty na veľkých dátových množinách.

Experimentovali sme s metódami dolovania vo veľkých objemoch dát so zameraním na predikciu prúdových numerických dát a zhlukovanie statických a prúdových dát.

Predstavili sme symbolickú reprezentáciu časových radov spolu s vzdialenostnou mierou na transformované dáta. Použiteľnosť je demonštrovaná na základe veľmi dlhých časových radov a potenciálne nekonečných prúdov dát, teda na údajoch o spotrebe elektrickej energie získaných počas takmer desaťročného obdobia.

Predstavili sme symbolickú reprezentáciu časových radov na základe reprezentácie skupín podobných podpostupností ako symboly. Navrhli sme ich reprezentáciu spolu s Euklidovou vzdialenostnou mierou a dolnou hranicou na pôvodné dáta.

Vytvorili sme nástroj ANNOR (Automatic image aNNOtation Retriever) na automatizované anotovanie obrázkov v podobe kľúčových slov, ktorý využíva metódy strojového učenia pri rozpoznávaní objektov na obrázku. Predstavili sme doménovo-nezávislý model používateľa uvažujúci personalizované váhovanie metadát na rozličných úrovniach abstrakcie, čím sme dokázali modelovať a porovnávať záujmy používateľov na viacerých úrovniach. Táto možnosť predstavuje výhodu pri aplikovaní na veľkých dátových kolekciách a pri veľkom množstve používateľov. Rovnako pri modelovaní preferencií pomocou navrhnutého prístupu dokážeme na základe viacerých úrovní záujmu vyjadriť používateľom preferované koncepty aj v rozsiahlych dátových kolekciách.

10 Použitá literatúra

[1] ADHIKARI, R., AGRAWAL, R.K.: A Novel Weighted Ensemble Technique for Time Series Forecasting. In: *Advances in Knowledge Discovery and Data Mining*, 2012, pp. 38–49.

[2] ADILAH, N. et al.: Electricity Load Demand Forecasting Using Exponential Smoothing Methods. *World Applied Sciences Journal*, vol. 22, 2013, no. 11, pp. 1540–1543.

[3] AHMAD, S.: *Tutorial on Natural Language Processing*. Cedar Falls, IA: , 2007.

[4] ALFARES, H.K., NAZEERUDDIN, M.: Electric load forecasting: Literature survey and classification of methods. *International Journal of Systems Science*, vol. 33, 2002, no. 1, pp. 23–34.

[5] ALMA, Ö.G.: Comparison of robust regression methods in linear regression. *International Journal of Contemporary Mathematical Sciences*, vol. 6, 2011, no. 9-12, pp. 409–421.

[6] ARMSTRONG, J.S.: *Long-range forecasting*. New York: Wiley New York, 1985. 687 p.

[7] ARORA, S., TAYLOR, J.W.: Short-Term Forecasting of Anomalous Load Using Rule-Based Triple Seasonal Methods. *IEEE Transactions on Power Systems*, vol. 28, 2013, no. 3, pp. 3235–3242.

[8] ASBER, D. et al.: Non-parametric short-term load forecasting. *International Journal of Electrical Power & Energy Systems*, vol. 29, 2007, no. 8, pp. 630–635.

[9] BA, A.: Adaptive Learning of Smoothing Functions : Application to Electricity Load Forecasting. In: *NIPS*, 2012.

[10] BARC INSTITUTE: 2013. Big Data Survey Europe [cit. 2015-02-09]. http://www.pmone.com/fileadmin/user_upload/doc/study/BARC_BIG_DATA_SURVEY_EN_final.pdf.

[11] BASHIR, Z. A., EL-HAWARY, M.E.: Applying Wavelets to Short-Term Load Forecasting Using PSO-Based Neural Networks. *IEEE Transactions on Power Systems*, vol. 24, 2009, no. 1, pp. 20–27.

[12] BEYER, M.A., LANEY, D.: *The Importance of "Big Data": A Definition*. 2012, 7 p.

[13] BIANCO, V., MANCA, O., NARDINI, S.: Linear Regression Models to Forecast Electricity Consumption in Italy. *Energy Sources, Part B: Economics, Planning, and Policy*, vol. 8, 2013, no. 1, pp. 86–93.

[14] BIEHN, N.: 2013. The missing V's in big data: viability and value [cit. 2015-02-09]. http://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/.

[15] BIFET, A.: Mining Big Data in Real Time. *Informatica*, vol. 37, 2013, no. 1, pp. 15–20.

[16] BIFET, A. et al.: MOA: Massive Online Analysis. *The Journal of Machine Learning Research*, vol. 11, 2010, pp. 1601–1604.

[17] BIG PROJECT: *D2.2.2 Final Version of Technical White Paper*. 2014, 155 p. http://big-project.eu/sites/default/files/BIG_D2_2_2.pdf.

[18] BOWERMAN, B.L., CONNELL, R.T.O., KOEHLER, A.B.: *Forecasting, time series, and regression: an applied approach*. [s.l.]: Thomson Brooks, 2005. 720 p.

[19] BOX, G.E.P., JENKINS, G.M.: *Time series analysis: Forecasting and control*. 3. Ed. New Jersey: Prentice Hall, 1994. 592 p.

[20] BREIMAN, L.: Bagging Predictors. *Machine Learning*, vol. 24, 1996, no. 2, pp. 123–140.

[21] BREIMAN, L.: Random Forests. *Machine Learning*, vol. 45, 2001, no. 1, pp. 5–32.

[22] BU, Y. et al.: HaLoop. *Proceedings of the VLDB Endowment*, vol. 3, 2010, no. 1-2, pp.

285–296.

[23] CAI, Y. et al.: An efficient approach for electric load forecasting using distributed ART (adaptive resonance theory) & HS-ARTMAP (Hyper-spherical ARTMAP network) neural network. *Energy*, vol. 36, 2011, no. 2, pp. 1340–1350.

[24] CANCELO, J.R., ESPASA, A., GRAFE, R.: Forecasting the electricity load from one day to one week ahead for the Spanish system operator. *International Journal of Forecasting*, vol. 24, 2008, no. 4, pp. 588–602.

[25] CIMIANO, P.: *Ontology learning and population from text: algorithms, evaluation and applications*. [s.l.]: Springer Verlag, 2006. ISBN 9780387306322.

[26] CLARK, A., FOX, C., LAPPIN, S.: *The Handbook of Computational Linguistics and Natural Language Processing.* [s.l.]: Wiley-Blackwell, 2010. 802 p.

[27] CONDIE, T. et al.: MapReduce online. In: *NSDI'10*, 2010, pp. 21–21.

[28] DAN, J. et al.: Piecewise Trend Approximation: A Ratio-Based Time Series Representation. *Abstract and Applied Analysis*, vol. 2013, 2013, pp. 1–7.

[29] DANNECKER, L. et al.: Context-Aware Parameter Estimation for Forecast Models in the Energy Domain. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6809 LNCS, 2011, pp. 491–508.

[30] DEAN, J., GHEMAWAT, S.: MapReduce. *Communications of the ACM*, vol. 51, 2008, no. 1, p. 107.

[31] ESTER, M. et al.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: *KDD-96 Proceedings*, 1996, pp. 226–231.

[32] FAN, J.Y., MCDONALD, J.D.: A real-time implementation of short-term load forecasting for distribution power systems. *IEEE Transactions on Power Systems*, vol. 9, 1994, no. 2, pp. 988–994.

[33] FAN, S., CHEN, L.: Short-Term Load Forecasting Based on an Adaptive Hybrid Method. *IEEE Transactions on Power Systems*, vol. 21, 2006, no. 1, pp. 392–401.

[34] FAN, W., BIFET, A.: Mining big data. *ACM SIGKDD Explorations Newsletter*, vol. 14, 2013, no. 2, p. 1.

[35] FISTER JR., I. et al.: A Brief Review of Nature-Inspired Algorithms for Optimization. *Elektrotehniski Vestnik*, vol. 80, 2013, no. 3, pp. 1–7.

[36] FRALEY, C., RAFTERY, A.E.: Model-based Methods of Classification: Using the mclust Software in Chemometrics. *Journal of Statistical Software*, vol. 18, 2007, no. 6, pp. 1–13.

[37] FREUND, Y., SCHAPIRE, R.E.: Experiments with a New Boosting Algorithm. In: *Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)*, 1996, pp. 148–156.

[38] GABER, M.M. et al.: Data stream mining in ubiquitous environments: state-of-the-art and current directions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, 2014, no. 2, pp. 116–138.

[39] GÁL, F.: *Možnosť a skutočnosť*. [s.l.]: Obzor, 1990. 192 p.

[40] GAMA, J. et al.: A survey on concept drift adaptation. *ACM Computing Surveys*, vol. 46, 2014, no. 4, pp. 1–37.

[41] GAMA, J.: Data Stream Mining: the Bounded Rationality. *Informatica*, vol. 37, 2013, no. 1, pp. 21–25.

[42] GAO, Y., ER, M.J.: NARMAX time series model prediction: feedforward and recurrent fuzzy neural network approaches. *Fuzzy Sets and Systems*, vol. 150, 2005, no. 2, pp. 331–350.

[43] GIRAUD-CARRIER, C.: A Note on the Utility of Incremental Learning. *AI Communications*, vol. 13, 2000, no. 4, pp. 215–223.

[44] GOULD, P.G. et al.: Forecasting time series with multiple seasonal patterns. *European Journal of Operational Research*, vol. 191, 2008, pp. 205–220.

[45] HAGAN, M.T., BEHR, S.M.: The Time Series Approach to Short Term Load Forecasting. *IEEE Transactions on Power Systems*, vol. 2, 1987, no. 3, pp. 785–791.

[46] HAHSLER, M., FORREST, J.: *Introduction to stream : An extensible Framework for Data Stream Clustering Research with R*. 2014.

[47] HALL, M. et al.: The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, vol. 11, 2009, no. 1, p. 10.

[48] HAN, J., KAMBER, M.: *Data mining concepts and techniques*. 2006. 743 p. ISBN 9781558609013.

[49] HAUPT, R.L., HAUPT, S.E.: *Practical Genetic Algorithms*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2003. ISBN 0471455652.

[50] HIPPERT, H.S., PEDREIRA, C.E., SOUZA, R.C.: Neural networks for short-term load forecasting: a review and evaluation. *IEEE Transactions on Power Systems*, vol. 16, 2001, no. 1, pp. 44–55.

[51] HOLLAND, J.: *Adaptation in Natural and Artificial Systems*. Cambridge, MA: MIT Press, 1992. 211 p.

[52] HONG, T.: Energy Forecasting : Past , Present and Future. *Foresight: The International Journal of Applied Forecasting*, 2014, no. 32, pp. 43–48.

[53] HONG, T.: *Short Term Electric Load Forecasting*. Raleigh, NC: North Carolina State University, 2010, 157 p.

[54] HONG, W.-C.: Electric load forecasting by seasonal recurrent SVR (support vector regression) with chaotic artificial bee colony algorithm. *Energy*, vol. 36, 2011, no. 9, pp. 5568–5578.

[55] HONG, W.-C.: Electric load forecasting by support vector model. *Applied Mathematical Modelling*, vol. 33, 2009, no. 5, pp. 2444–2454.

[56] HONG, W.-C.: *Intelligent Energy Demand Forecasting*. London: Springer London, 2013. ISBN 978-1-4471-4967-5.

[57] HOR, C.-L., WATSON, S.J., MAJITHIA, S.: Analyzing the Impact of Weather Variables on Monthly Electricity Demand. *IEEE Transactions on Power Systems*, vol. 20, 2005, no. 4,

pp. 2078–2085.

[58] HU, J.: *Data mining in practice: Key techniques and real world applications*. 2013. http://sites.ieee.org/scv-cs/archives/data-mining-in-practice-key-techniques-and-real-world-applications.

[59] HUDEC, M.: Radoslav Haluška a kol: Smart metering áno, ale rozumne. *energia.sk*, 2012. http://www.energia.sk/rozhovor/elektrina-a-elektromobilita/smart-metering-ano-ale-rozumne/7845/.

[60] HURWITZ, J. et al.: *Big Data for Dummies*. 1. Ed. [s.l.]: For Dummies, 2013. 336 p.

[61] HYDE, O., HODNETT, P.F.: An adaptable automated procedure for short-term electricity load forecasting. *IEEE Transactions on Power Systems*, vol. 12, 1997, no. 1, pp. 84–94.

[62] HYNDMAN, R.J. et al.: *Forecasting with exponential smoothing: The state space approach*. Berlin: Springer-Verlag, 2008. 362 p.

[63] HYVÄRINEN, A., OJA, E.: Independent component analysis: algorithms and applications. *Neural Networks*, vol. 13, 2000, no. 4-5, pp. 411–430.

[64] CHEN, J.-F., WANG, W.-M., HUANG, C.-M.: Analysis of an adaptive time-series autoregressive moving-average (ARMA) model for short-term load forecasting. *Electric Power Systems Research*, vol. 34, 1995, no. 3, pp. 187–196.

[65] CHENG, Q. et al.: Short-term load forecasting with weather component based on improved extreme learning machine. In: *2013 Chinese Automation Congress*, 2013, pp. 316–321.

[66] CHENTHUR PANDIAN, S. et al.: Fuzzy approach for short term load forecasting. *Electric Power Systems Research*, vol. 76, 2006, no. 6-7, pp. 541–548.

[67] CHIU, C.-C., KAO, L.-J., COOK, D.F.: Combining a neural network with a rule-based expert system approach for short-term power load forecasting in Taiwan. *Expert Systems with Applications*, vol. 13, 1997, no. 4, pp. 299–305.

[68] IBM: 2013. Secure and protect big data environments [cit. 2015-02-09]. http://www-01.ibm.com/software/data/guardium/secure-big-data/.

[69] JUBERIAS, G. et al.: A new ARIMA model for hourly load forecasting. In: *1999 IEEE Transmission and Distribution Conference (Cat. No. 99CH36333)*, 1999, pp. 314–319 vol.1.

[70] KALMAN, R.E.: A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME--Journal of Basic Engineering*, vol. 82, 1960, no. Series D, pp. 35–45.

[71] KARABOGA, D., BASTURK, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *Journal of Global Optimization*, vol. 39, 2007, no. 3, pp. 459–471.

[72] KENNEDY, J., EBERHART, R.: Particle swarm optimization. *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4, 1995.

[73] KUMAR, V. et al.: DEDUCE. In: *Proceedings of the 13th International Conference on Extending Database Technology - EDBT '10*, 2010, p. 657.

[74] LACLAVÍK, M., ŠELENG, M.: *Vyhľadávanie informácií*. Bratislava: Nakladateľstvo STU, 2012. 128 p.

[75] LANEY, D.: *3D data management: Controlling data Volume, Velocity and Variety*. Stanford, CT: META Group, 2001, 3 p.

[76] LAURINEC, P.: Application of genetic algorithm on model-based cluster analysis. In: *Proceedings of 11th Student Reasearch Conference in Informatics and Information Technologies*, 2015, pp. 115–122.

[77] LEACOCK, C., MILLER, G.A., CHODOROW, M.: Combining Local Context and WordNet Similarity for Word Sense Identification. In: *WordNet: An Electronic Lexical Database*, 1998, pp. 265–283.

[78] LEE, K.-H. et al.: Parallel data processing with MapReduce. *ACM SIGMOD Record*, vol. 40, 2012, no. 4, p. 11.

[79] LEGÉŇ, M.: Šanca na inteligentné šetrenie. *TREND*, 2013, no. 8. http://www.etrend.sk/trend-archiv/rok-2013/cislo-8/sanca-na-inteligentne-setrenie.html.

[80] LESKOVEC, J., RAJARAMAN, A., ULLMAN, J.D.: *Mining of Massive Datasets*. [s.l.]: Stanford University, 2014. .

[81] LI, Y., HE, X., ZHANG, W.: Applications of adaptive CKF algorithm in short-term load forecasting of smart grid. In: *Proceedings of the 33rd Chinese Control Conference*, 2014, pp. 8145–8149.

[82] LIN, J. et al.: Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, vol. 15, 2007, no. 2, pp. 107–144.

[83] LIU, W. et al.: Neural Network Based on Self-adaptive Differential Evolution for Ultra-Short-Term Power Load Forecasting. In: *Intelligent Computing in Bioinformatics*, 2014, pp. 403–412.

[84] LLOYD, S.: Least squares quantization in PCM. *IEEE Transactions on Information Theory*, vol. 28, 1982, no. 2, pp. 129–137.

[85] MA, S. et al.: The variable weight combination load forecasting based on grey model and semi-parametric Regression Model. In: *2013 IEEE International Conference of IEEE Region 10 (TENCON 2013)*, 2013, pp. 1–4.

[86] MAATEN, L. VAN DER, HINTON, G.: Visualizing Data using t-SNE. *Journal of Machine Learning Research*, vol. 9, 2008, no. Nov, pp. 2579–2605.

[87] MAIA, C. A., GONÇALVES, M.M.: A methodology for short-term electric load forecasting based on specialized recursive digital filters. *Computers & Industrial Engineering*, vol. 57, 2009, no. 3, pp. 724–731.

[88] MARZ, N.: *Big Data: Principles and best practices of scalable realtime data systems*. [s.l.]: Manning, 2014. 259 p. ISBN 9781617290343.

[89] MBAMALU, G.A.N., EL-HAWARY, M.E.: Load forecasting via suboptimal seasonal autoregressive models and iteratively reweighted least squares estimation. *IEEE Transactions on Power Systems*, vol. 8, 1993, no. 1, pp. 343–348.

[90] MCDONALD, J.R., LO, K.L., SHERWOOD, P.M.: The application of short-term adaptive

forecasting techniques in energy management for the control of electrical load. *Transactions of the Institute of Measurement and Control*, vol. 11, 1989, no. 2, pp. 79–91.

[91] MENDES-MOREIRA, J. et al.: Ensemble approaches for regression. *ACM Computing Surveys*, vol. 45, 2012, no. 1, pp. 1–40.

[92] MÉSZÁROS, A.: Akumulácia elektrickej energie v elektrickej sieti. *Elektroenergetika*, vol. 5, 2012, no. 2, pp. 34–37.

[93] MIKOLAJ, J., KĽUČKA, J., VANČO, B.: *Plánovanie a prognostika*. Žilina: FŠI ŽU, 2005. 242 p.

[94] MIRASGEDIS, S. et al.: Models for mid-term electricity demand forecasting incorporating weather influences. *Energy*, vol. 31, 2006, no. 2-3, pp. 208–227.

[95] MIRJALILI, S., MIRJALILI, S.M., LEWIS, A.: Grey Wolf Optimizer. *Advances in Engineering Software*, vol. 69, 2014, pp. 46–61.

[96] MOGHRAM, I., RAHMAN, S.: Analysis and evaluation of five short-term load forecasting techniques. *IEEE Transactions on Power Systems*, vol. 4, 1989, no. 4, pp. 1484–1491.

[97] MOHAMMED, O. et al.: Practical experiences with an adaptive neural network short-term load forecasting system. *IEEE Transactions on Power Systems*, vol. 10, 1995, no. 1, pp. 254–265.

[98] MORALES, G.D.F.: SAMOA : A Platform for Mining Big Data Streams. In: *WWW '13 Companion Proceedings of the 22nd international conference on World Wide Web companion*, 2013, pp. 777–778.

[99] NEUMEYER, L. et al.: S4: Distributed Stream Computing Platform. In: *2010 IEEE International Conference on Data Mining Workshops*, 2010, pp. 170–177.

[100] OPPENHEIM, A., VERGHESE, G.: 2010. Introduction to Communication, Control, and Signal Processing. Chapter: Estimation with Minimum Mean Square Error [cit. 2015-06-07]. http://ocw.mit.edu.

[101] PAARMANN, L.D., NAJAR, M.D.: Adaptive online load forecasting via time series modeling. *Electric Power Systems Research*, vol. 32, 1995, no. 3, pp. 219–225.

[102] PAI, P.-F.: Hybrid ellipsoidal fuzzy systems in forecasting regional electricity loads. *Energy Conversion and Management*, vol. 47, 2006, no. 15-16, pp. 2283–2289.

[103] PAPALEXOPOULOS, A.D., HESTERBERG, T.C.: A regression-based approach to short-term system load forecasting. *IEEE Transactions on Power Systems*, vol. 5, 1990, no. 4, pp. 1535–1547.

[104] PARALIČ, J.: *Dolovanie znalostí z textov*. Košice: Equilibria, s.r.o., 2010. 182 p.

[105] PARK, J.H., PARK, Y.M., LEE, K.Y.: Composite modeling for adaptive short-term load forecasting. *IEEE Transactions on Power Systems*, vol. 6, 1991, no. 2, pp. 450–457.

[106] PERRONE, M.P., COOPER, L.N.: When networks disagree: Ensemble methods for hybrid neural networks. In: *How We Learn; How We Remember: Toward an Understanding of Brain and Neural System*, 1995, pp. 342–358.

[107] PORTER, M.E.: *Competitive Advantage: Creating and Sustaining Superior Performance*. [s.l.]: Free Press, 1985. 592 p.

[108] POWER, D.J.: What is the “true story” about data mining, beer and diapers? *DSS News*, vol. 3, 2002, no. 23.

[109] RAHMAN, S., BHATNAGAR, R.: An expert system based algorithm for short term load forecast. *IEEE Transactions on Power Systems*, vol. 3, 1988, no. 2, pp. 392–399.

[110] RAHMAN, S., HAZIM, O.: A generalized knowledge-based short-term load-forecasting technique. *IEEE Transactions on Power Systems*, vol. 8, 1993, no. 2, pp. 508–514.

[111] REYNOLDS, A.P. et al.: Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *Journal of Mathematical Modelling and Algorithms*, vol. 5, 2006, no. 4, pp. 475–504.

[112] RICHARDSON, M., DOMINOWSKA, E., RAGNO, R.: Predicting clicks: estimating the click-through rate for new ads. In: *WWW ’07*, 2007, pp. 521–530.

[113] ROSENBLATT, F.: A probabilistic model for information storage and organization in the brain. *Psychological Review*, vol. 65, 1958, no. 6, pp. 386–408.

[114] ROSS, G.J., TASOULIS, D.K., ADAMS, N.M.: Nonparametric Monitoring of Data Streams for Changes in Location and Scale. *Technometrics*, vol. 53, 2011, no. 4, pp. 379–389.

[115] SAS: 2013. What is big data? [cit. 2015-02-09]. http://www.sas.com/en_us/insights/big-data/what-is-big-data.html.

[116] SED: *Ročenka 2013*. 2013. http://www.sepsas.sk/seps/DispSkladacka2013.asp?kod=575.

[117] SELAKOV, A. et al.: Hybrid PSO–SVM method for short-term load forecasting during periods with significant temperature variations in city of Burbank. *Applied Soft Computing*, vol. 16, 2014, pp. 80–88.

[118] SHIRKHORSHIDI, A.S. et al.: Big Data Clustering: A Review. In: 2014, pp. 707–720.

[119] SIMON, D.: Biogeography-Based Optimization. *IEEE Transactions on Evolutionary Computation*, vol. 12, 2008, no. 6, pp. 702–713.

[120] SOARES, L.J., MEDEIROS, M.C.: Modeling and forecasting short-term electricity load: A comparison of methods with an application to Brazilian data. *International Journal of Forecasting*, vol. 24, 2008, no. 4, pp. 630–644.

[121] SOLIMAN, S.A. et al.: Application of least absolute value parameter estimation technique based on linear programming to short-term load forecasting. In: *Proceedings of 1996 Canadian Conference on Electrical and Computer Engineering*, 1996, pp. 529–533.

[122] SOUZA, R.C., MIRANDA, C.V.C. DE: Short Term Load Forecasting Using Double Seasonal Exponential Smoothing and Interventions to Account for Holidays and Temperature Effects. 2007, pp. 1–13.

[123] TAYLOR, J.W., MCSHARRY, P.E.: Short-Term Load Forecasting Methods: An Evaluation Based on European Data. *IEEE Transactions on Power Systems*, vol. 22, 2007, no. 4, pp. 2213–2219.

[124] TAYLOR, J.W., SNYDER, R.D.: Forecasting intraday time series with multiple seasonal cycles using parsimonious seasonal exponential smoothing. *Omega*, vol. 40, 2012, no. 6, pp. 748–757.

[125] TAYLOR, J.W.: Short-Term Electricity Demand Forecasting Using Double Seasonal Exponential Smoothing. *Journal of Operational Research Society*, vol. 54, 2003, pp. 799–805.

[126] TAYLOR, J.W.: Smooth transition exponential smoothing. *Journal of Forecasting*, vol. 23, 2004, no. 6, pp. 385–404.

[127] TAYLOR, J.W.: Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research*, vol. 204, 2010, no. 1, pp. 139–152.

[128] TAYLOR, J.W.: Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research*, vol. 204, 2010, no. 1, pp. 139–152.

[129] THOMAS, J.J., COOK, K.A.: *Illuminating the path: The research and development agenda for visual analytics*. [s.l.]: IEEE Computer Society Press, 2005. 184 p.

[130] TRUDNOWSKI, D.J., MCREYNOLDS, W.L., JOHNSON, J.M.: Real-time very short-term load prediction for power-system automatic generation control. *IEEE Transactions on Control Systems Technology*, vol. 9, 2001, no. 2, pp. 254–260.

[131] TSEKOURAS, G.J. et al.: A non-linear multivariable regression model for midterm energy forecasting of power systems. *Electric Power Systems Research*, vol. 77, 2007, no. 12, pp. 1560–1568.

[132] VOSSEN, P.: *EuroWordNet: A multilingual database with lexical semantic networks*. Dordrecht: Springer Netherlands, 1998. ISBN 978-90-481-5120-2.

[133] VOSSEN, P.: WordNet, EuroWordNet and Global WordNet. *Revue Francaise de Linguistique Appliquée*, vol. VII, 2002, no. 1, pp. 27–38.

[134] WAI, R.-J. et al.: Performance comparisons of intelligent load forecasting structures and its application to energy-saving load regulation. *Soft Computing*, vol. 17, 2013, no. 10, pp. 1797–1815.

[135] WANG, B. et al.: A new ARMAX model based on evolutionary algorithm and particle swarm optimization for short-term load forecasting. *Electric Power Systems Research*, vol. 78, 2008, no. 10, pp. 1679–1685.

[136] YING, L.-C., PAN, M.-C.: Using adaptive network based fuzzy inference system to forecast regional electricity loads. *Energy Conversion and Management*, vol. 49, 2008, no. 2, pp. 205–211.

[137] YOO, H., PIMMEL, R.L.: Short term load forecasting using a self-supervised adaptive neural network. *IEEE Transactions on Power Systems*, vol. 14, 1999, no. 2, pp. 779–784.

[138] YU, Y., ZHOU, Z.-H., TING, K.M.: Cocktail Ensemble for Regression. In: *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 2007, pp. 721–726.

[139] ZDRUŽENIE DODÁVATEĽOV ELEKTRINY: *Posúdenie výhodnosti zavedenia inteligentných meračov elektriny v podmienkach SR*. 2012, 1-40 p.

11 Príloha: Zoznam výstupov projektu

PUBLIKOVANÉ

1. Rozinajová, V., Kosková, G., Ezzeddine, A. B., Lucká, M., Bieliková, M., & Návrat, P. (2014). Informácia o riešení témy Big Data v projekte „ Medzinárodné centrum excelentnosti pre výskum inteligentných a bezpečných informačno - komunikačných technológii a systémov “. In V. Svátek & O. Zamazal (Eds.), *Proceedings of 13th Annual Conference Znalosti 2014: Exhibition, Education and Expert finding* (pp. 78–81). Jasná pod Chopkom, Nízke Tatry, Slovakia: Oeconomica.
2. Kosková, G., Ezzeddine, A. B., Lucká, M., Rozinajová, V., & Laurinec, P. (2014). Perspektívy modelovania a predikovania veľkých dát v energetike. In L. Hluchý, M. Bieliková, & J. Paralič (Eds.), *Proceedings of 9th Workshop on Intelligent and Knowledge Oriented Technologies* (pp. 57–62). Smolenice, Slovakia: Nakladateľstvo STU.
3. Vrablecová, P. (2014). Inteligentná analýza veľkých objemov dát. In L. Hluchý, M. Bieliková, & J. Paralič (Eds.), *Proceedings of 9th Workshop on Intelligent and Knowledge Oriented Technologies* (pp. 126–129). Smolenice, Slovakia.
4. Vrablecová, P. (2015). Power demand forecasting from stream data with concept drifts. In M. Bieliková (Ed.), *Proceedings of 11th Student Reasearch Conference in Informatics and Information Technologies* (pp. 131–132). Bratislava, Slovakia: Vydavateľstvo STU.
5. Ujhelyi, M., Lacko, P., & Paulovič, A. (2014). Task Scheduling in Distributed Volunteer Computing Systems. In *IEEE 12th International Symposium on Intelligent Systems and Informatics* (pp. 111–114).
6. Kuric, E., & Bielikova, M. (2015). ANNOR: Efficient image annotation based on combining local and global features. *Computers & Graphics*, *47*, 1–15. http://doi.org/10.1016/j.cag.2014.09.035
7. Martoš, I., & Rozinajová, V. (2015). Software service recommendation using context. In *IEEE Fourth Eastern European Regional Conference on the Engineering of Computer Based Systems* (pp. 32–38).

PRIJATÉ NA PUBLIKOVANIE

8. Vrablecová, P., Rozinajová, V., & Ezzeddine, A. B. (2015). Data Stream Mining in the Power Engineering Domain. In *Proceedings of Data a Znalosti 2015*. Prague, Czech Republic.
9. Grmanová, G., Rozinajová, V., Ezzeddine, A. B., Lucká, M., Lacko, P., Lóderer, M., … Laurinec, P. (2015). Application of Biologically Inspired Methods to Improve

Adaptive Ensemble Learning. In *7th World Congress on Nature and Bilogically Inspired Computing*. [prijaté].

10. Rozinajová, V., Ezzeddine, A. B., Lucká, M., Lacko, P., Grmanová, G., Laurinec, P., & Vrablecová, P. (2015). Otvorené smery výskumu v oblasti dátovej analytiky. In *Proceedings of 10th Workshop on Intelligent and Knowledge oriented Technologies 2015*. Košice, Slovakia.
11. Kosková, G., Laurinec, P., Rozinajová, V., Ezzeddine, A. B., Lucká, M., Lacko, P., … Návrat, P. (2015). Incremental ensemble learning for electricity load forecasting. *Acta Polytechnica Hungarica.*
12. Lóderer, M., Rozinajová, V., & Ezzeddine, A. B. (2015). Predikcia spotreby elektrickej energie založená na kombinácii predikčných metód. In *Proceedings of Data a Znalosti 2015*. Prague, Czech Republic.
13. Laurinec, P., & Lucká, M. (2015). Analýza vplyvu redukcie dimenzionality na zhlukovanie veľkých dátových množín. In *Proceedings of Data a Znalosti 2015* (pp. 1–6).
14. Sevcech, J., & Bielikova, M. (2015a). Repeating Patterns as Symbols for Long Time Series Representation. *Special Issue on Software Architectures and Systems for Real Time Data Stream Analytics, Journal of Systems and Software*.
15. Sevcech, J., & Bielikova, M. (2015b). Symbolic Time Series Representation for Stream Data Processing. In *The 1st IEEE International Workshop on Real Time Data Stream Analytics held in conjunction with IEEE BigDataSE-15*. Helsinki, Finland: IEEE.
16. Kaššák, O., Kompan, M., & Bieliková, M. (2015b). User Preference Modeling by Global and Individual Weights for Personalized Recommendation. *Acta Polytechnica Hungarica.*
17. Farkaš, T., Kubán, P., Lucká, M. (2016). Effective Parallel Multicre-optimized K-mers Counting Algorithm, SOFSEM 2016: 42nd International Conference on Current Trends in Theory and Practice of Computer Science, Harrachov, Czech Republic, January 23–28, 201.

ODOSLANÉ/POSUDZOVANÉ

18. Vrablecová, P., Rozinajová, V., Ezzeddine, A. B., Grmanová, G., Lucká, M., & Lacko, P. (2015). Stream Data Analysis for Power Demand Forecasting. *International Transactions on Electrical Energy Systems.*
19. Vrablecová, P., & Šimko, M. (2015). Supporting Semantic Annotation of Educational Content by Automatic Extraction of Hierarchical Domain Relationships. *IEEE Transactions on Learning Technologies.*
20. Kaššák, O., Kompan, M., & Bieliková, M. (2015a). Students' Behavior in a Web-Based Educational System: Exit Intent Prediction. *Engineering Applications of Artificial Intelligence*.