Univerzita Pardubice
Fakulta ekonomicko-správní fakulty
Ústav systémového inženýrství a informatiky

Návrh modelu na bázi Soft Case-based Reasoning

Disertační práce

Autor: Ing. Filip Mezera
Školitel: doc. Ing. Jiří Křupka, PhD.

Pardubice 2018

University of Pardubice
Faculty of Economic and Administration
Institute of System Engineering and Informatics

Design of model on the basis of Soft Case-based Reasoning

Thesis

Author: Filip Mezera
Supervisor: Jiří Křupka

Pardubice 2018

Prohlašuji:

Tuto práci jsem vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Beru na vědomí, že v souladu s § 47b zákona č.111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů, a směrnicí Univerzity Pardubice č. 9/2012, bude práce zveřejněna v Univerzitní knihovně a prostřednictvím Digitální knihovny Univerzity Pardubice.

Souhlasím se zpřístupněním své práce v Univerzitní knihovně.

V Pardubicích dne 17. 4. 2018

Filip Mezera

PODĚKOVÁNÍ

Děkuji svému školiteli Jiřímu Křupkovi za spolupráci, podporu, cenné rady a připomínky nejen při psaní disertační práce, ale i během celého studia. Dále chci poděkovat mé rodině, bez jejíž pomoci a dobré vůle bych studium nedokončil.

ANOTACE

Případové usuzování (Case-based Reasoning) je jedním z přístupů k rozhodování. Vychází z anglosaského pojetí práva, které je založeno na precedentech, využívá porovnání nového případu (Case) se stávajícím, starým případem, podle kterého bylo již v minulosti rozhodnuto. Jednou z možných cest, jak zvýšit kvalitu rozhodování u případového usuzování, je využití metod výpočetní inteligence, které pomohou v případě, kdy je v datech zastoupena nepřesnost či neurčitost.

Cílem práce je představit skupiny dvou klasifikačních modelů, k jejichž řešení bylo využito metod případového usuzování a výpočetní inteligence. První skupina se věnuje modelování kvality ovzduší a druhá se zabývá klasifikací nového klienta v nebankovní finanční instituci. Navržený Huntův model případového usuzování pro hodnocení kvality ovzduší pracuje s daty z meteorologických a imisních stanic. Pro optimalizaci modelu jsou využity teorie rough i fuzzy množin. Model klasifikace nového klienta byl navržen na bázi 4R modelu případového usuzování. Následnou úpravou 4R modelu došlo k vytvoření hybridního inteligentního systému, který při hledání podobnosti případů využívá neuronovou síť, teorii rough množin a kombinuje různé metriky. Výsledky navržených modelů jsou porovnány s dalšími metodami, jako jsou například rozhodovací stromy nebo logistická regrese. Modely byly vytvořeny na reálných datech a jejich výsledky lze aplikovat v praxi.

KLÍČOVÁ SLOVA

Případové usuzování, výpočetní inteligence, měření vzdálenosti, kvalita ovzduší, nebankovní finanční instituce, management, klasifikace

TITLE

Design of model on the basis of Soft Case-based Reasoning

ANNOTATION

Case-based Reasoning is one of the approaches to decision making. Based on the Anglo-Saxon concept of law, which is based on precedents, it uses a comparison of a new case with an existing, old case, which was resolved in the past. One possible way to increase the quality of decision-making in Case-based Reasoning is to use computational intelligence methods to help with inaccuracies or uncertainties in the data.

The aim of this thesis is to introduce groups of two classification models, whose solutions were used Case-based Reasoning methods and Computational Intelligence. The first group

deals with air quality modelling and the second deals with the classification of a new client in a non-banking financial institution. The proposed Hunt model of Case-based Reasoning for air quality assessment works with data from meteorological and air pollution stations. The theory of both rough and fuzzy sets is used to optimize the model. The new client's classification model was designed based on the 4R model of Case-based Reasoning. Subsequent modification of the 4R model has created a hybrid intelligent system that uses a neural network, the rough sets theory and combines various metrics to find similarities. The results of the proposed models are compared with other methods, such as decision trees or logistic regression. The models were created on real data sets and their results can be applied in practice.

KEYWORDS

Case-based reasoning, Soft computing, Distance measurement, Air quality, Non-banking financial institution, Management, Classification

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

SEZNAM ZKRATEK

BI – Business Intelligence

CBR – Case-based Reasoning – Případové usuzování

CHMI – Český hydrometeorologický ústav

CI – Computational Intelligence – Výpočetní inteligence

CM – Confusion Matrix – Matice záměn

CRM – Customer Relationship Management – Řízení vztahů se zákazníkem

CZK – Czech Crown – Česká koruna (měna)

ČB – čištění báze případů

ČNB – Česká národní banka

ENh – Elastic Neighbourhood – Elastické (dynamicky se měnící) okolí případu

FIS – Fuzzy Inference System – Fuzzy inferenční systém

FO – Fyzická osoba – nepodnikající

FOP – fyzická osoba – podnikající

FST – Fuzzy Sets Theory – Teorie fuzzy (neostrých) množin

HC – Hill Climbing – Gradientní optimalizační algoritmus prohledávání stavového prostoru

LR – Logistic Regression – Logistická regrese

k-NNbs – k-Nearest Neighbors – k-nejbližších sousedů – metoda CBR porovnávající nový případ s k-nejbližšími případy z báze případů, z důvodu možné záměny se zkratkou neuronových sítí (NNs) byla použita tato zkratka oproti konvenční *k*-NNs

MLP – Multi Layer Perceptron – typ neuronové sítě

NN / NNs – Neural Network/s – neuronová síť / neuronové sítě

PI – Payment Institution (Electronic Money Institution) – Platební instituce (Nebankovní finanční instituce)

PM_{10} – Particulate Matter – Prachové částice o velikosti do 10 mikrometrů

PO – právnická osoba

RBF – Radial Basis Function – typ neuronové sítě

RFM analýza – Recency, Frequency, Monetary Analysis – analýza klientů pomocí zařazení do shluků dle doby od posledního obchodu, četnosti a velikosti obchodů

RdCBR – CBR s využitím hrubé vzdálenosti

RFA – Rough-Fuzzy Approach – přístup kombinující RST a FST

RSES – Rough Sets Exploration System – program

RST – Rough Sets Theory – Teorie rough množin

SNh – Static Neighbourhood – Staticky určené okolí případu

TM – Testovací množina

SCBR – Soft CBR – CBR s využitím metod výpočetní inteligence

SV – Stárnutí vzorů v bázi případů

TDIDT / TDIDTs – Top Down Induction Decision Tree / s – Rozhodovací strom / stromy

VM – Validační množina

VBA – Visual Basic for Applications – programovací jazyk používaný v balíčku Microsoft Office

ÚVOD

Tato práce představuje klasifikační modely založené na případovém usuzování (Case-based Reasoning, dále jen CBR) s využitím metod výpočetní inteligence (Computational Intelligence, dále jen CI). Rozhodování je velmi často modelováno pomocí řetězení generalizovaných pravidel. Tímto přístupem se tak snaží postihnout skutečnost. S rostoucí složitostí modelu však nastává velký nárůst počtu pravidel a v určité chvíli již není možné takový systém realizovat a model se musí zjednodušit, čímž dochází k jeho posunu od modelované skutečnosti. CBR se na rozhodování dívá z diametrálně odlišného úhlu. Zdrojem znalostí nejsou generalizovaná pravidla, ale získání řešení pomocí nejrelevantnějších případů uchovávaných v paměti a jejich přizpůsobením nové situaci [42]. CBR je tedy založeno na zapamatování si minulých případů – jejich podmínek, způsobu a výsledku řešení [58].

Vycházejí tak z řešení situací v reálném světě, kde je obvyklé, že podobné problémy mají podobná řešení. Z toho důvodu je vyřešení dřívějšího případu dobrým odrazovým můstkem pro další situace [58]. Druhým příznakem je, že jednotlivé typy rozhodovacích situací mají tendenci se opakovat. Budoucí situace tak budou podobné současným problémům. Pokud tyto dvě teze platí pro vybraný okruh problémů, pak je CBR velmi efektivním nástrojem na řešení problémů [42].

V případě, že podobnost případů není tak velká, musí řešitel vyvinout větší kreativitu. Ta je podstatná nejen pro úspěšné vyřešení konkrétního problému. V této fázi, totiž dochází k odchylkám, které mohou řešení posunout na vyšší úroveň. Tím se zároveň zlepší celkové výsledky, protože budoucí případy již budou porovnávány s tímto kvalitním řešení. Zde je opět vidět rozdíl oproti pravidlovým (rule-based) systémům, které tendují k určité expertem stanovené strategii řešení problému [42].

Primární využití CI v CBR je její schopnost se vypořádat s neurčitostí a generovat nové návrhy řešení. Jak CI, tak CBR jsou samostatně velmi dobře popsány. Stejně tak je možné najít množství primárně teoretických článků, které propojení CBR a CI nabízejí. Cílem této práce je nejenom vytvořit modely, ale taktéž popsat jejich možné praktické využití, výhody a nevýhody využití CI v CBR. Popsané modely jsou validovány pomocí dalších metod jako rozhodovací stromy (Top Down Induction Decision Trees – dále TDIDTs), neuronové sítě (Neural Networks – dále NNs, více viz [31]), shlukování příp. logistická regrese (Logistic Regresion – dále LR, v [34]).

V rámci mého výzkumu byly vytvořeny dvě skupiny modelů. První z nich se týká kvality ovzduší a řešení smogových situací (zvýšený výskyt prachových částic v ovzduší). Ty se na

problémech s ovzduším podílejí až 96 %. Současný stav je takový, že se omezení daná smogovou situací vyhlašují až s 36 hodinovým zpožděním. Cílem modelu bylo dosáhnout klasifikace tak, aby bylo možné takovou situaci ohlásit co nejdříve po zjištění prvního překročení imisních limitů. Zpracování a ověření výsledků je reálné v řádu tří hodin [38].

Základem tohoto modelu bylo sestavení modelu za pomocí hrubých (rough) množin (Rough Sets Theory – dále jen RST). Díky tomu došlo k redukci parametrů modelu a standardního CBR. Výsledky CBR byly následně konfrontovány s TDIDTs a NNs. Model založený na RST (v [38]) byl dále rozvíjen, resp. posloužil jako základ pravidel pro fuzzy inferenční systém (Fuzzy Inference System – dále FIS). Výsledky dosažené pomocí teorie Fuzzy množin (Fuzzy Sets Theory – dále FST) byly dále porovnávány s výsledky CBR. Na základě tohoto porovnání vznikla citlivostní analýza, která může být použita k rozvoji jak CBR, tak hybridního rough-fuzzy přístupu (Rough-Fuzzy Approach – dále RFA) v [37].

Druhý model se zabývá klasifikací klientů, kdy z obchodních dat získaných v krátkém období (v našem případě jsou to tři měsíce od podpisu smlouvy) zkoušíme zjistit dlouhodobý přínos klienta pro společnost. Správná klasifikace pak může znamenat snížení nákladů jak z hlediska vyplácených odměn obchodní síti za akvizici, tak z pohledu přiřazení odpovídajícího sazebníku a marže jednotlivým klientům tak, aby společnost do jednoho roku od začátku obchodního vztahu začala na klientovi generovat zisk. Zároveň je dobrým vodítkem pro marketingové kampaně a celkový přístup ke klientovi v rámci tzv. Customer Relationship Management (CRM). Rozdílem oproti standardní segmentaci klientů, jejichž typickým příkladem je RFM analýza (Recency, Frequency, Monetary Analysis – např. v [59]), je zaměření se nikoliv na stávající, ale na nové klienty. Model tak tyto segmentační metody nenahrazuje, ale doplňuje je [12].

Opět byl nejdříve vypracován model CBR. Tentokrát však byla báze rostoucí. Postupně se testovaly další algoritmy zkvalitňující rozhodování (časový rozdíl mezi vzorem a porovnávaným případem – tzv. „stárnutí“ případů v bázi – a také čištění báze od odchýlených případů). Zároveň bylo nutné porovnávat nastavení a vzájemné působení těchto algoritmů. Proto se využil nejdříve expertní odhad a následně optimalizační metoda Hill Climbing (HC). Vše bylo dále porovnáváno s metodami TDIDTs, NNs a LR [40]. Nakonec byly provedeny další experimentální úpravy využívající buď upravený princip k nejbližších sousedů (*k*-Nearest Neighbors, dále jen *k*-NNbs) nebo výpočet vzdálenosti na základě předpokladu neurčitosti. Předpokladem využití těchto metod byla především jejich robustnost, která zvyšovala kvalitu klasifikace především v případě nižšího počtu případů v bázi případů. Výsledkem je model s vysokou kvalitou

klasifikace na úrovni NNs, zároveň je robustní a velmi dobře uplatnitelný i v případě využití na částečně odlišných datech [42].

Data pro obě skupiny modelů byla předzpracována, normalizována a standardizována (transformace dle [34] s. 70 až 72). Pro výpočty a vytváření modelů bylo využito běžné programové vybavení. Výpočty NNs, TDIDTs a LR proběhly v programu IBM SPSS Modeler, RST v programu Rough Set Exploration System (RSES – Lehmanův algoritmus, dostupné v [52], popis v [20]), FST v MATLABu (FIS typ Mandani). Samotné CBR, včetně doplňujících algoritmů a výpočtu jejich nastavení pomocí metody HC (dle [48]) bylo zpracováno v programu Microsoft Excel, resp. naprogramováno v jeho doplňku programovacím jazykem Visual Basic for Application (VBA). Nastavení hodnot jednotlivých parametrů je provedeno pomocí metody HC v rámci trénovací množiny. U dynamického systému, kde nemůžeme přímo mluvit o testovací množině, se nastavení provádí na množině případů, která svou velikostí a charakterem odpovídá trénovací množině. V případě, že dojde k odchýlení od tohoto postupu, je to explicitně v textu zmíněno. Samotná metoda je realizována tak, jak je popsána v [48] a není porovnávána s jinými metodami.

1 CÍLE PRÁCE

Cílem práce je navrhnout klasifikační modely s využitím metody Soft CBR (SCBR) a ověřit jejich schopnost pracovat s neurčitostí v datech. V prvním případě se jedná o statický klasifikační model, v druhém případě o dynamický, přičemž budou testovány různé přístupy v rámci jednotlivých fází CBR. Dosažené výsledky budou porovnány s výstupy, které byly dosaženy pomocí dalších metod (například NNs, TDIDTs, RST).

Hlavní cíl je možné rozdělit na následující subcíle:

- stanovení vhodných metod omezujících počet vstupních atributů
- cílené využití metod definujících strukturu báze případů
- porovnání různých přístupů k získání nejpodobnějšího případu / případů a jejich vliv na výslednou klasifikaci
- vytvoření dynamického modelu CBR
- zjištění možnosti integrace výsledků jednotlivých klasifikačních metod včetně CBR, tedy vytvoření hybridního inteligentního systému, jak ho chápe Larry Medsker, např. v [33].

Uvedené dílčí cíle lze specifikovat následovně:

Pro stanovení vhodných metod omezujících počet vstupních atributů jsou kromě expertního odhadu použité metody, které jsou schopny určit důležitost jednotlivých atributů. Patří mezi ně RST, která generuje pravidla, na jejichž základě lze určit váhu atributů, popř. je možné je využít přímo k redukci „nadbytečných“ atributů. Podobně lze na základě výstupů z TDIDTs stanovit důležitost jednotlivých atributů. Poslední zkoumanou možností bude využití citlivostní analýzy.

V rámci statického modelu CBR bude zkoumána různá struktura báze případů, kdy báze nebude obsahovat všechny případy, ale pouze vybrané zástupce jednotlivých tříd. Třídy budou stanoveny dle výsledné imisní situace (3 třídy) a případy v rámci nich budou rozděleny do shluků. Následně bude každý shluk reprezentován jedním případem. Testováno bude využití buď případu, který je nejblíže středu shluku, nebo vytvoření umělého případu, jehož charakteristiky budou odpovídat středním hodnotám atributů případů v daném shluku (těžiště objektů ve shluku).

Ve fázi získání nejpodobnějšího případu / případů (dále jako fáze získání) budou testovány různé metody výpočtu vzdálenosti (Euklidovská, Manhattan, Čebyševova metrika) a různé přístupy k výběru více podobných případů. Mezi ně kromě standardních metod *k*-NNbs (testováno bude CBR pro 3 a 5 nejbližších sousedů) patří také metody prohledávání okolí případu. Vedle

staticky určeného okolí, je to i dynamicky se měnící okolí v závislosti na vzdálenosti nového případu a nejbližšího případu z báze případů, nebo okolí měnící se dle zaplněnosti báze případů.

U dynamického modelu bude nutné vyřešit především praktické řešení problémů spojených s transformací dat (především standardizací). Dále pak dynamické řízení zachycení případů do báze případů. Zde je předpoklad využití expertem určených pravidel. Otázkou v takovémto systému je, jak zachytit trendy, které se v datech vyskytují, a následně dle toho upravit fázi získání. Stejně tak je důležité zajistit schopnost báze případů vypořádat se s odchýlenými či jinak nevalidními případy. Ve statickém systému se dají identifikovat a odstranit předem. V dynamickém je nutné každý případ podrobit automatizovanému zkoumání.

Posledním je integrace různých metod do dynamického klasifikačního modelu tak, aby takový model byl robustní, schopný si poradit s odchýlenými případy a reflektoval trendy v datech. Navíc rozdíl v kvalitě klasifikace musí být natolik průkazný, aby ospravedlnil zvýšené nároky na výpočetní prostředky.

2 POPIS SOUDOBÉHO STAVU ŘEŠENÉ PROBLEMATIKY

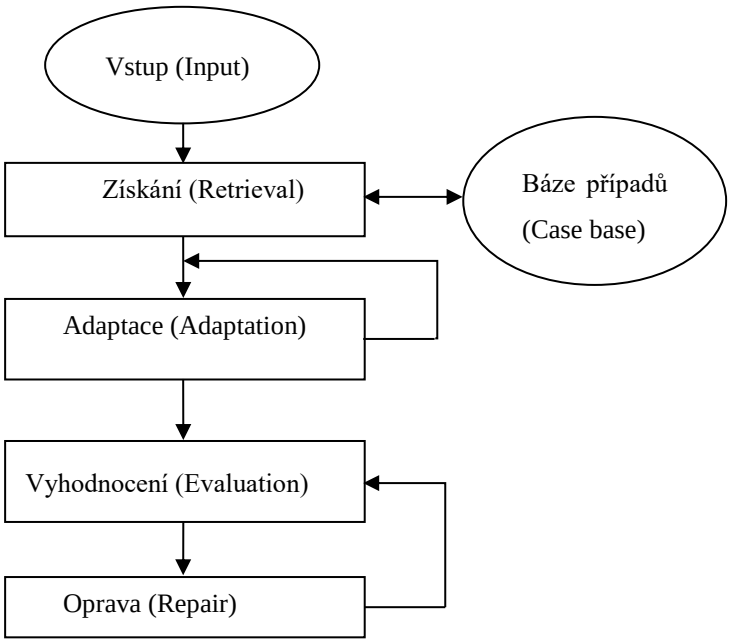
V rámci teoretických konceptů se vyvinulo několik modelů CBR, které rozdělovaly jednotlivé fáze CBR. Cílem rozdělení je zjednodušení řešení a zároveň lepší optimalizace těchto fází. Cíle CBR se dají rozdělit na dvě části: interpretaci a řešení problémů. Interpretační CBR využívá předcházející případy jako odrazový můstek pro klasifikaci nebo charakterizaci nové situace [4]. Formuje tak názor na ní. Naopak CBR řešící problémy využívá předchozí případy k navrhnutí řešení, které je možné aplikovat na nové podmínky. Proto oba dva typy potřebují rozdílný rozhodovací mechanismus ve chvíli, kdy jsou získány vhodné předchozí případy (vzory). V interpretačním CBR pak nastupuje fáze posouzení (justication) a v řešícím CBR následuje fáze adaptace (adaptation) [38]. Na tyto procesy navazuje přezkoumání (revision) předešlých řešení tak, aby vyhovovaly nové situaci. Záporné vyhodnocení nabízeného řešení pak spustí další adaptaci před tím, než je řešení aplikováno [57].

2.1 Huntův model CBR

Jeden z prvních modelů CBR vytvořili Janet Kolodner v [30] a David Leake v [32], kteří chápou CBR jako proces zapamatování a úpravy (v originále: „remember and adapt“) nebo zapamatování a srovnání (v originále: „remember and compare“). Oproti tomu Huntův model [54], který pracuje již s vytvořenou bází případů. Byl proto vybrán jako vhodný ke klasifikaci ovzduší, protože předpokládáme nejprve vytvoření báze na základě historických dat a teprve poté bude testována samotná klasifikace.

Základní struktura procesu v tomto modelu je na Obrázku 1. Jakmile je jednou získána báze případů, prvním krokem CBR systému je analyzovat vstupy do systému. Musí dojít k vyhledání důležitých informací. Následuje vyhledávání a porovnání podle těchto informací s bází případů. Takto jsou získány relevantní (nejvíce shodné) případy. Tím je uzavřena fáze získávání. Následuje fáze adaptace, která se snaží predikovat výsledek stávajícího případu podle toho, jak se liší současná situace od předchozích případů. Následuje fáze evaluace (vyhodnocení) navrhovaného řešení. Pokud je posouzeno jako přijatelné, je vloženo do báze případů, kde poslouží při posouzení budoucích případů [17].

Pokud nějaký aspekt současného problému není vyřešen, nastává fáze opravy, která do procesu rozhodování zahrne všechny části případu. Postup je následující: Nejdříve je třeba identifikovat, proč se nepodařilo případ vyřešit a následně použít tyto informace v opravném procesu.



Obrázek 1: Huntův model. Zdroj: [17, s. 60].

V modelu kvality ovzduší je fáze opravy mimo samotný model a určuje ji zvnějšku expert pomocí vyhodnocení dle expertního odhadu, popř. dle jiných klasifikačních metod.

2.2 Model 4R

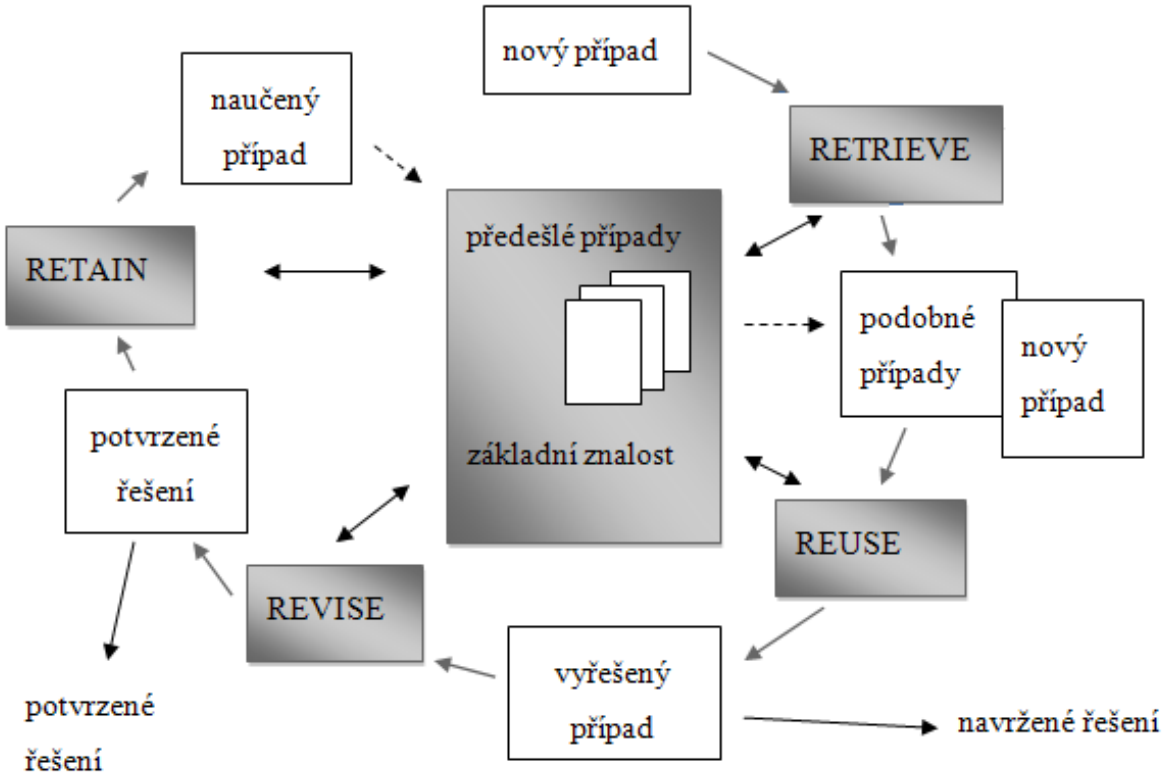
Aamodt a Plaza [1] navrhují 4R strukturu (cyklus) zobrazenou na Obrázku 2, se kterou se ztotožňuje nejvíce autorů. Vychází ze čtyř základních částí:

1. Retrieve (získání) – získá z báze předešlých případů ty, které vyhodnotí jako nejvíce podobné
2. Reuse (znovu užití) – použije výsledky podobných příkladů, které integruje
3. Revise (upravení) – adaptuje doporučené výsledné řešení tak, aby jím mohl být problém vyřešen
4. Retain (zachycení) – zachytí (vloží do báze) nové řešení problému, jakmile je potvrzeno a zkontrolováno

V praktických aplikacích je mnohdy obtížné od sebe fáze Reuse a Revise odlišit a někteří autoři používají jednu fázi – fázi adaptace, která výše zmíněné kombinuje a nahrazuje, což představuje

určitý návrat k dříve uvedeným modelům. Dělení na čtyři fáze přesto přetrvává, a to z důvodu jasnějšího naznačení komplikovaných činností v rámci práce s bází případů [42].

Na obrázku lze vidět, že k předchozím případům je přidána určitá základní znalost. Ta je závislá na typu úlohy a doplňuje bázi předešlých případů. Závislost může být jak velmi silná, tak slabá. Tato znalost může být definována pomocí IF-THEN pravidel nebo v podmínkách pro použití předchozích případů.



Obrázek 2: Model 4R. Zdroj: [42, s. 6].

Finnieho a Sunův model navazuje na předchozí model 4R. Přidává navíc další část: Repartition. Základním předpokladem je, že ne všechny aplikace mohou ihned zahájit fázi Retrieve. Báze případů je tak rozdělena podle možných stavů problémů a jejich řešení. Rozdělení je vytvořeno na základě vazeb, které vznikají mezi stavy s určitou pravděpodobností jejich míry podobnosti. CBR se tak transformuje v pravděpodobnostně založené usuzování. Dochází tak k dělení množiny možných stavů na podmnožiny podobných případů. Ty se dále dělí, až je podmnožinou myšlen konkrétní případ. Následně je tak použita podmnožina, která nejvíce odpovídá současnému problému. Čím jemnější odlišení použijeme, tím se ve fázi Retrive dostaneme ke konkrétnějším údajům (více v [17] a [54]).

Tento model bude využit pro tvorbu dynamického CBR, kde se budou případy postupně do báze přidávat. Případ nemůže být do báze přidán ihned po klasifikaci, protože se musí čekat do

konečného vyhodnocení, čímž se zjistí jeho skutečné zařazení. V této době je tedy umístěn v bufferu, v rámci něhož se pravidelně ověřuje podmínka, dle níž ho lze finálně zařadit. V našem případě, kdy jsou data z delšího období, je nutné zachytit trendy, které se v datech vyskytují.

2.3 Zkoumané fáze CBR

Zaznamenané případy v bázi případů mohou zobrazovat různé typy znalostí, které mohou být uloženy v různých formátech. Zamýšlený cíl konkrétního CBR pak bude významně ovlivňovat to, co je v bázi uloženo. Zřejmě se budou lišit systémy, které mají pomoci vytvořit nový návrh nebo plán (predikční modely) od těch, které problém diagnostikují (klasifikační modely). Proto v každém typu CBR systému mohou případy zobrazovat něco odlišného (lidi, objekty, situace, diagnózy, návrhy, plány, pravidla atd. [3]). Obvykle tak báze tvoří strukturu dvou množin. V první jsou uchovány atributy jednotlivých případů a v druhé množině pak jejich řešení. Tyto množiny obvykle nejsou strukturovány. Právě otázka, co přesně zaznamenat, patří mezi nejobtížnější rozhodnutí při tvorbě CBR, protože množina možných atributů je obvykle „bezrozměrná“.

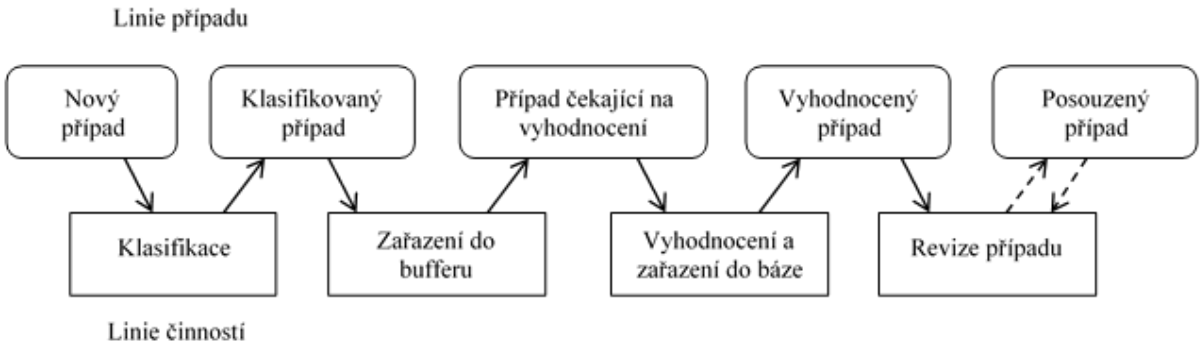
Bez ohledu na to, o jaký typ případů jde, musí být jeho vlastnosti uvedeny v nějakém formátu. Jednou z výhod CBR je flexibilita v přístupu, který nabízí vzhledem k reprezentaci vlastností. Záleží na typu atributů, které mají být reprezentovány a na používané platformě. Tato implementace se pohybuje od numerických, textových dat a časových řad až po vztahy mezi daty. Takto mohou být zachyceny databázové formáty, objekty nebo sémantické sítě. Nezáleží tedy přímo na tom, jak jsou data uložena nebo v jakém datovém formátu. Především jde o relevantnost informací vzhledem k cíli systému. Zároveň se ověřuje schopnost CBR zjistit nejvhodnější minulý případ k vyřešení současné situace.

Pokud CBR pracuje nad velkým množstvím dat, pak již nemusí být všechny nové situace zaznamenávány. K posouzení zařazení do báze dat nebo naopak vymazání pak slouží určitá kritéria (pravidla). K vymazání dochází především v případě velké podobnosti mezi jednotlivými případy. Dále je možné vytvořit jeden umělý případ, který integruje dva a více reálných případů [42]. Tento přístup bude zvolen u modelu zabývajícího se imisní situací, kdy budou otestovány možnosti využít buď případ, který je nejvíce středu shluku případů, které se mají integrovat, nebo vytvořit umělý případ, který bude tento shluk reprezentovat pomocí průměrných hodnot jednotlivých atributů.

2.3.1 Indexace případů

Indexace případů je důležitá především s ohledem na budoucí využití a porovnání případů. Volba vhodného indexu pak pomáhá získat nejpodobnější případ v co nejkratším čase. Indexace pak určuje nejdůležitější ukazatele (atributy), které rozhodují o vhodnosti případů. Z tohoto důvodu má samotná indexace určitou prediktivní schopnost [30].

Indexy jsou tedy určitým souhrnem, který zahrnuje potřebné okolnosti využité ve fázi retrieve. V případě, že by jich zahrnovaly příliš mnoho, pak by buď navracely velké množství historických případů, nebo by zpracování trvalo velmi dlouho. Ačkoliv určení indexů stále spočívá na rozhodnutí expertů, je snaha tuto činnost automatizovat. Například Bonzano a kolektiv [9] použili induktivní techniky ke zjištění lokálních vah atributů pomocí porovnání podobných případů v bázi případů. Tato metoda pomáhá určit, které vlastnosti jsou více důležité při predikci výstupu a zvyšují tak kvalitu fáze Retrive. Bruninghaus a Ashley [11] vyvinuli agenta, který spravuje víceúrovňovou znalost tak, že ji převádí na normativní údaje. Tímto způsobem učící se program klasifikuje texty a připravuje je k dalšímu využití. Díky tomu dojde k odfiltrování nepodstatných informací. Index je pak součásti hierarchického stromu, který představuje různé stavy důležité pro uživatele. Další techniky mohou využívat například indexování pomocí vlastností a jejich množin predikovaných napříč celou doménou problému [2], adaptací vedeného indexování a vrácení případu [53] až po na vysvětlování založených technikách.



Obrázek 3: Linie případu a linie činnosti v dynamickém modelu CRM. Zdroj: autor.

V případě CBR klasifikace kvality ovzduší je báze případů malá. Je však nutné v rámci ní odlišit případy zastupující jednotlivé shluky od těch zbylých, které nejsou ke klasifikaci dále využívány. Index případu tak v první řadě rozlišuje tento atribut a dále zahrnuje informaci o tom, do jakého shluku daný případ patří, a taktéž jeho jednoznačnou identifikaci. Pokud by se model nasadil do běžného používání, pak by se musel popasovat s dynamicky se rozšiřující bází. Řešení by odpovídalo druhému modelu klasifikace klientů. Jednotlivé fáze, kterými prochází případ, a činnosti s tím spojené jsou popsány na Obrázku 3.

Indexace se týká každé činnosti, která je s případem vykonána. V dynamickém systému klasifikace klientů dostává každý nový případ jedinečné ID, které ho pak provází celým jeho cyklem v CBR. Během klasifikace se doplní výsledná hodnota klasifikace a případ je zařazen do bufferu, kde čeká na finální vyhodnocení. Pokud jsou případy v bufferu technicky zařazeny do báze případů (s příznakem zatím nevalidního případu), je vhodné tento atribut do indexace taktéž zahrnout. Po vyhodnocení pak dostane případ další příznak, zda se shoduje klasifikace s výsledným hodnocením. Poslední zásadní příznak může dostat ve chvíli, kdy se zjistí, že byl buď špatně klasifikován vzhledem k jeho konečnému vyhodnocení, nebo byl na základě nově vstupujících případů do báze uznán jako nevalidní případ. Obě možnosti znamenají, že jsme ho identifikovali jako odchýlenou hodnotu a z báze ho buď musíme odstranit, nebo ho alespoň označit jako případ, který je nevalidní (ať už trvale či dočasně).

2.3.2 Fáze získání

Získání případu (fáze Retrieve – viz Obrázek 4) je procesem hledání a nalezení nejvhodnějších případů v bázi případů. Efektivní získání záleží především na zvolených kritériích, podle kterých hodnotíme podobnost, a zároveň na mechanismu prohledávání báze případů. Kritéria jsou potřebná k rozhodnutí, který případ je nejlepší získat pro porovnání dle podobnosti mezi současným případem a historickými případy (případy v bázi případů). Částečně při jejich výběru zohledňujeme i způsob prohledávání báze. Nejčastěji se projde celá báze. Pak zvolené atributy nehrají příliš velkou úlohu. U některých specializovaných technik prohledávání ale dochází k silné závislosti mezi touto volbou a výsledkem CBR. Jedná se především o případy, kdy se báze neskládá z dostatečného množství případů, a řešení je tvořeno syntézou několika případů. Nebo jsou historické případy natolik odlišné od současné situace, že není možné získat jasné řešení pomocí nejbližšího případu [42].

Procesy ve fázi Retrieve závisí především na modelu paměti a indexování. Vyhledávání pak probíhá za použití množství velmi odlišných metod. Od jednoduchého vyhledávání pomocí statistické metody nejbližšího souseda až po užití inteligentních agentů. Tato fáze je jednou z hlavních výzkumných oblastí CBR. Nejčastěji zkoumané techniky dle [42] jsou metoda nejbližšího souseda, rozhodovací stromy a pravidlové systémy:

1. Získání nejbližšího souseda: V tomto případě je získán takový případ, jehož součet vah atributů v porovnání se současným případem je větší než u ostatních případů v bázi. Pokud budou všechny atributy ohodnoceny stejně, pak je vybrán takový případ, který se shoduje v n atributech místo případu, kde je shoda v k atributech ($n > k$). Vlastnosti, které mají větší

váhu v procesu rozhodování, pak musí být lépe ohodnoceny i v procesu porovnání případů. Následně jsou možné úpravy tohoto mechanismu buď cestou k-nejbližších případů, kde pro vyhodnocení případu využijeme právě *k* nejbližších případů bez ohledu na to, jak jsou vzdálené, nebo prohledávání okolí případu v bázi, kdy pro vyhodnocení případu využijeme n případů, které splňují podmínku celkové vzdálenosti.

2. Induktivní přístup: Pokud je použit induktivní přístup ke sledování struktury báze, pak je výsledkem hierarchická struktura, která redukuje prostor prohledávání pro case retriever. Zároveň rozhoduje o relativní důležitosti atributů. To má za následek kratší čas pro prohledávání, ale také možné nepřesnosti vyhledávání.

3. Znalostně řízený přístup: Tento přístup získání je založen na doménové znalosti, která rozhoduje o důležitosti jednotlivých atributů v rámci porovnání budoucích případů. V některých situacích odlišné vlastnosti atributů budou mít odlišný stupeň důležitosti k úspěšnému přiřazení případů. Vzniká tím opět hierarchická struktura, která usnadňuje prohledávání báze.

4. Potvrzované získání: Existuje velké množství pokusů ke zlepšení fáze Retrieve. Jako příklad můžeme uvést ten, který navrhuje Simoudis v [51]. Skládá se ze dvou fází: První z nich vyhledá v bázi všechny případy, které mohou být nějak relevantní pro řešení daného problému. Ve druhé fázi se odvozuje, který z těchto případů je pro řešení daného problému nejvíce relevantní. Největší výhodou potvrzovaného získávání je, že se v první fázi nad celou bází mohou použít nenáročné výpočetní metody a ty náročné se použijí až v druhé fázi nad relativně malou podmnožinou případů z báze.

Je mnoho faktorů, které ovlivňují výběr metody pro fázi Retrieve:

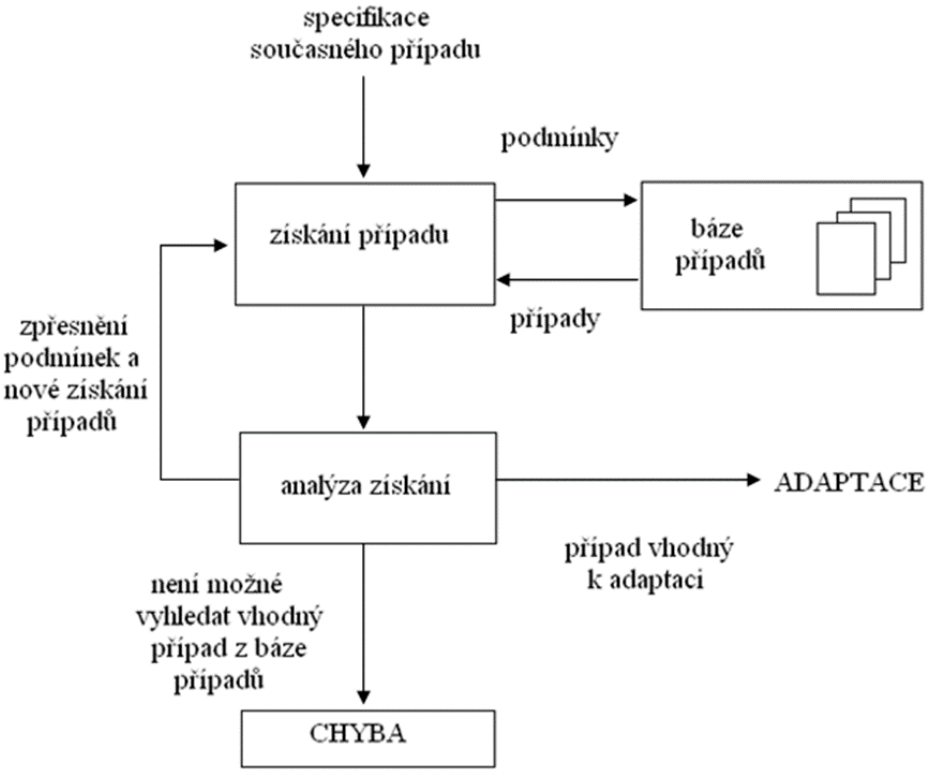
- množství případů v bázi případů,
- množství znalostí, které o případech máme (počet atributů),
- nutnost a složitost váhování individuálních proměnných,
- zda všechny případy mohou být stejně indexovány nebo zda je nutné pracovat s množstvím atributů, které řeší kvalitu výsledku či jeho relevanci za specifických podmínek.

Jakmile je vybrán nejpodobnější případ, je vhodné udělat analýzu, zda je vybraný případ z báze znalostí skutečně podobný (v případě řešení pomocí vzdálenosti, zda je dostatečně blízko) a je možné ho použít, nebo zda je nutné modifikovat parametry hledání a spustit prohledávání báze

znovu. Jakékoliv nesprávné rozhodnutí se projeví ve významném nárůstu času, který je potřebný pro zpracování úlohy, protože úprava nepodobného (vzdáleného) případu je obvykle významně delší než opětovné projití báze případů. Rozhoduje:

- čas a zdroje vyžadované pro adaptaci případu,
- čas a zdroje vyžadované pro prohledávání báze,
- počet případů a již prozkoumaných případů v bázi – tedy jaká je pravděpodobnost, že nalezneme vhodnější případ – vzor.

Proces je možné popsat pomocí následujícího Obrázku 4.



Obrázek 4: Fáze získání. Zdroj: [42, s. 17].

2.3.3 Adaptace případu

Adaptace případu je proces, který transformuje získané řešení problému v řešení, které je adekvátní současnému případu. Adaptace je jedním z nejdůležitějších kroků v CBR, který jinak relativně jednoduchému vyhledávání vzorů přidává prvky inteligence [57].

Nejčastějšími přístupy k adaptaci případů jsou:

- získané řešení může být použito k řešení současného případu bez modifikace nebo s modifikací, přestože řešení není zcela adekvátní stávající problému,

- kroky nebo procesy, které vedou k získání dřívějšího řešení, mohou být opakovány bez modifikací či s modifikacemi, které však nemusí být zcela vyhovující stávající situaci,
- pokud je vybráno více podobných případů jako vzory, řešení může být odvozeno z více případů, popř. může být prezentováno více možných řešení.

Adaptace využívá množství rozdílných technik například pravidel, nového spuštění CBR na přesněji specifikovaných částech problému. Při výběru možností adaptace je vhodné zvážit následující:

- jak moc se od sebe v průměru liší současný problém a získané řešení z báze,
- kolik charakteristik se v rámci případů porovnává, popř. kolik se jich obvykle mezi současným problémem a navrhovaným řešením liší,
- zda existují znalosti, které by mohly pomoci při tvorbě pravidel využitelných v rámci fáze adaptace.

Jakmile je fáze adaptace ukončena, je vhodné porovnat současný případ a případ získaný z báze případů. Především je nutné zjistit velikost rozdílů mezi jednotlivými atributy případů a posoudit jejich dopad na adaptaci řešení původního případu [58]. Pokud takto adaptované řešení zřejmě nepovede k vyřešení stávajícího případu, je nutné rozhodnout o dalších krocích. Může se jednat buď o odchýlený případ, změnu v hodnotách atributů v nových případech (změna dlouhodobého trendu), nebo o zcela nový typ případu. V rámci takovéhoto zařazení je pak nutné nejenom rozhodnout o řešení takovéhoto případu, ale i o jeho zpracování a vložení do báze případů [42].

2.4 Fuzzy množiny

Uplatnění fuzzy modelování [42], [60] je účelné ve všech případech, kdy se řeší problém spojený s neurčitostí, s nepřesností, případně pokud je problém silně ovlivněn subjektivním přístupem řešitele. FST se snaží pokrýt realitu v její nepřesnosti a neurčitosti. Fuzzy modelování slouží pro popis jevů, které lze jen obtížně popisovat klasicky (příliš složité či neurčité problémy buď exaktní řešení přímo vylučují, nebo je činí nepoužitelnými).

Základním předpokladem je převod verbálních prvků (vágních pojmů, např. vysoký člověk, nízká rychlost) do stupnice, kde se dají pojmy alespoň obecně kvantifikovat. Idea fuzzy množiny je velmi přirozená. Nejsme-li schopni stanovit přesné hranice množiny (v klasickém slova

smyslu) určené vágním pojmem, nahradíme rozhodnutí o náležení nebo nenáležení konkrétního prvku do dané množiny určitou mírou vybranou z předem definovaného intervalu.

V klasické teorii množin prvek do množiny buď patří (úplné členství v množině) nebo nepatří (žádné členství v množině). Fuzzy množina je množina, která kromě úplného nebo žádného členství umožňuje i částečné členství. Toto částečné členství je vyjádřeno prostřednictvím stupně příslušnosti. Funkce, která každému prvku *x* universa *U* přiřadí stupeň příslušnosti, se nazývá funkce příslušnosti μ_A fuzzy množiny *A* [60]. Vzhledem ke klasické teorii množin nabývá stupeň příslušnosti hodnot z intervalu $\langle 0;1 \rangle$. To znamená, že každému prvku *x* přiřazuje hodnotu funkce příslušnosti $\mu_A(x) \in \langle 0;1 \rangle$, který se nazývá stupněm příslušnosti prvku *x* do fuzzy množiny *A*. Je-li $\mu_A(x) = 0$, pak *x* nepatří do fuzzy množiny *A*, je-li $\mu_A(x) = 1$, pak *x* patří do fuzzy množiny *A*. Je-li $0 < \mu_A(x) < 1$, pak *x* částečně patří do fuzzy množiny *A* s danou hodnotou (stupněm) příslušnosti [60].

Formální zápis fuzzy množiny je pak

$$A = \left(\mu_A(x_i) \Big/ x_i \right) \text{ pro } \forall x_i \qquad (2.1)$$

Základní operace nad fuzzy množinami je možné nalézt v [61]. Každá proměnná může mít různý počet fuzzy množin, které odpovídají například různým lingvistickým hodnotám.

Klasické modely systémů jsou postaveny na základě vztahu mezi vstupem a výstupem systému. Fuzzy modely toto klasické pojetí transformují do skupiny pravidel.

Lingvisticky popsaný model je vyjádřen jako skupina pravidel IF – THEN s neurčitými tvrzeními. Je tedy založen na znalostech [60], [61].

Proces sestavení fuzzy inferenčního modelu lze popsat ve třech bodech:

- fuzzifikace – převedení ostrých (crisp) vstupních hodnot na neurčité (fuzzy) hodnoty,
- inference (odvozování) – na základě fuzzy inferenčních pravidel jsou ze vstupních neurčitých hodnot určeny výstupní, neurčité hodnoty (v rámci FIS),
- defuzzifikace – výstupní neurčité hodnoty jsou převedeny na výstupní ostré hodnoty.

První krok zahrnuje definici rozmezí hodnot, pro které jsou funkce příslušnosti dané fuzzy množiny platné. Toto rozmezí hodnot může být následně i pozměněno tak, aby přesněji vyjádřilo příslušnost hodnoty pro danou fuzzy množinu [60].

Druhý krok zahrnuje definici pravidel, jež popisují vztah mezi vstupními a výstupními proměnnými vyjádřenými pomocí funkcí příslušnosti. Jde vlastně o rozložení problému na množství rozhodnutí.

Defuzzifikace představuje proces, při kterém se z výsledné fuzzy množiny výstupní veličiny určuje například jedna konkrétní hodnota výstupu. Existuje několik metod defuzzifikace (metoda centroidů, střední hodnota součtů, …).

V případě případového usuzování na bázi CI je z fuzzy modelu možné využít případně jen fuzzifikaci. Následné výpočty se poté provádějí pomocí tzv. fuzzy podobnosti. Příklad výpočtu (několik kroků) je podrobněji rozepsán v [60].

Samotné využití fuzzy množin v procesu klasifikace je vhodné právě z důvodu schopnosti zpracovávat neurčitost, která je v datech zahrnuta. V rámci modelu imisní situace bude nejprve klasifikace zpracována pomocí RST. Výstupy budou následně využity pro CBR (eliminace nadbytečných atributů), tak pro FIS, kdy se podle pravidel vytvořených pomocí RST, vytvoří IF-THEN pravidla pro FIS. Zároveň dojde k analýze citlivosti dle počtu atributů a pravidel, které do FIS vstupují.

2.5 Neuronové sítě

Umělé NNs jsou inspirovány biologickým nervovým systémem – mozkem, který se skládá z velkého počtu (přibližně 10^{11}) vysoce spojených prvků – neuronů (přibližně 10^4 spojení na neuron). Mozek uchovává a zpracovává informace úpravou propojení mezi neurony. NNs jsou systémy pro zpracování signálu, které se snaží napodobit chování a způsoby zpracování informací v biologickém nervovém systému tím, že poskytují matematický model kombinace neuronů a jejich propojení v síti [42].

V umělé NN jsou umělé neurony navzájem propojeny prostřednictvím spojení. Každému spojení je přiřazena váha, která řídí tok informací mezi neurony. Když informace vstoupí do neuronu přes spojení, nejprve se zpracuje a pak projde transformací aktivační funkcí f (x; w). Výstupy této aktivační funkce budou zaslány jiným neuronům nebo zpátky k sobě ve formě nového vstupu. V NN se vstupní informace zpracovávají v neuronech paralelně. To zlepšuje rychlost zpracování a spolehlivost NN.

Některé výhody NN jsou shrnuty níže:

1. Adaptivita – Síť může upravit své váhy spojení pomocí některých tréninkových algoritmů nebo pravidel pro učení. Aktualizací váhy může NN optimalizovat své spojení, aby se přizpůsobila změnám v prostředích. To je nejdůležitější charakteristika NNs.

2. Paralelní zpracování – Když jsou informace vloženy do NN, tak jsou distribuovány do různých neuronů pro zpracování. Neurony mohou pracovat paralelně a synergicky, pokud jsou aktivovány vstupy. V tomto uspořádání je výpočetní síla NNs plně využita a doba zpracování je snížena.

3. Robustnost – Pokud dojde k selhání jednoho z neuronů, mohou být váhy spojů upraveny tak, aby se zachovala výkonnost NN. Pracovní neurony vytvoří silnější spojení mezi sebou, zatímco propojení s neúspěšným neuronem bude oslabeno [31]. Díky tomu se zlepšuje spolehlivost NN [42].

Architektura NNs se dá rozdělit na dvě základní kategorie v závislosti na spojeních a topologii neuronů:

1. Dopředné NNs - Signál se šíří pouze jedním směrem, a to od vstupní vrstvy k výstupní vrstvě, žádná zpětná vazba není možná. NN neobsahuje žádné smyčky nebo zpětné vazby ve vrstvách (ve skryté vrstvě). Tím se liší od rekurentních NNs, které právě smyčky či zpětné vazby obsahují.

2. Rekurentní (zpětnovazební) NNs - Vstupní signál se může šířit v obou směrech, což je možné právě díky smyčkám a zpětnovazebním propojení. Takovou síť je nutné podrobit trénování. V této fázi se váhy upravují pomocí některých gradientních algoritmů nebo předdefinovaných pravidel pro učení. Až po této fázi může být taková NNs využita k řešení problémů [42].

V práci budou využity tři typy NNs. Prvním jsou Kohonenovy mapy – NNs s učením bez učitele, které mají dopřednou architekturu. Budou využity jako podpůrná metoda pro odhad počtu zástupných případů do báze případů u modelu kvality ovzduší. Tento odhad bude pro správnou funkčnost báze kritický.

Dalšími typy NNs, které budou využity, jsou síť Multi Layer Perceptron (MLP) a Radial Basis Function (RBF). V obou případech se jedná o NNs s učením s učitelem. Rozdíl je primárně v různé vnitřní funkci neuronu. Sítě budou využity jak pro ověření kvality klasifikace CBR oproti jiným metodám, tak jako možná součást hybridního modelu klasifikace.

Základním pilířem NNs je učení se / trénování NN. Existují v zásadě tři způsoby trénování NNs: učení bez učitele, učení s učitelem a posílené učení. Učení s učitelem probíhá tak, že váhy se nastaví tak, aby minimalizovaly rozdíl mezi výstupem a předpokládaným výstupem. U učení bez učitele je váha modifikována na základě vstupního signálu. V tomto případě je využito předdefinovaných pravidel, která určují, jak se váhy nastavují či mění. Například jednoduché

Hebbovo pravidlo učení uvádí, že jestliže dva sousední neurony vykazují podobné výstupy, jejich spojení bude posíleno.

$$\Delta w_{ij} = \lambda y_i y_j \qquad (2.2)$$

Kde Δw_{ij} je modifikace váhy a y_i a y_j jsou výstupy *i*-tého a *j*-tého uzlu a λ je délka kroku, což je malé pozitivní číslo, které řídí míru učení. V rovnici (2.2) je změna hmotnosti mezi uzly *i* a *j* úměrná výsledku jejich výstupů. Pokud jsou oba výstupy y_i a y_j kladné nebo záporné, zvyšuje se váha a vazba je posílena. Systém upravuje své váhy, dokud není dosaženo stability vah nebo se nedostane do oscilačního stavu.

Posílené učení je podobné učení s učitelem kromě toho, že trénovací případy jsou získávány v rámci využití výstupů z NNs. Pokud zpětná vazba označí výstup jako úspěšný, je nastavení vstupu a výstupu označeno tréninkový případ a neprovedou se žádné modifikace na vektoru vah NN. Pokud je výstup nesprávný, jsou váhy změněny na základě doménové znalosti, která je nejčastěji vyjádřena pomocí rozhodovací tabulky či IF-THEN pravidel. Opravená dvojice vstup-výstup je uložena a použita jako tréninkový příklad pro modifikaci vah.

Aplikace NNs, které využívají učení s učitelem (více viz [31]), se využívají pro modelování vstupně-výstupních vztahů některých složitých systémů, kde je těžké vytvořit explicitní matematický model. NNs využívající učení bez učitele jsou vhodné pro klasifikaci dat či vytvoření struktury dat. NNs využívající posílené učení jsou vhodné především tam, kde je velmi nízký počet trénovacích případů [42].

V případě dynamického systému je pak otázkou, jakým způsobem vytvořit učící se NN, která nebude extrémně náročná na výpočetní prostředky. Je možné nalézt kód „open source“ [5], který umožňuje adaptaci a činnost NN. Problém je s nárůstem počtu vstupních případů, kdy výrazně klesá rychlost naprogramovaného řešení ve VBA. Programy jako IBM SPSS Modeler naopak nepočítají s častým přepočítáváním NN. Při použití NNs je problematická práce s odchýlenými (atypickými, nestandardními) „novými“ případy.

2.6 Rough množiny

Teorie rough množin [44] je dalším přístupem k neurčitosti. Podobně jako u FST není alternativou ke klasické teorii množin, ale je zakotvena v ní. RST lze považovat za specifickou implementaci Fregeovy představy o nejasnosti, tj. nepřesnost v tomto přístupu je vyjádřena hraniční oblastí množiny, a nikoliv částečným členstvím, jako v FST.

Koncept RST lze definovat zcela obecně pomocí topologických operací, vnitřních a uzavřených, nazvaných aproximace. Přesnější popis je následující. Předpokládejme, že dostaneme

množinu objektů *U* nazvaných univerzem a nerozlišitelným vztahem $R \subseteq U \times U$, což představuje náš nedostatek znalostí o prvcích *U*. Pro jednoduchost předpokládáme, že *R* je rovnocenný vztah. Nechť *X* je podmnožina *U*. Chceme charakterizovat množinu *X* vzhledem k *R*. Pro tento účel budeme potřebovat základní pojmy RST uvedené níže.

Spodní aproximace množiny *X* s ohledem na *R* je množina všech objektů, které mohou být klasifikovány jako *X* s ohledem na *R* (jistě náleží do množiny *X* vzhledem k hodnotám *R*). Horní aproximace množiny *X* vzhledem k *R* je množina všech objektů, které nemohou být klasifikovány jako *X* vzhledem k *R* (nemohou náležet do *X* vzhledem k hodnotám *R*). Okrajovou oblastí množiny *X* vzhledem k *R* je množina všech objektů, které nelze klasifikovat jako *X*, ani jako nikoliv *X* vzhledem k *R* (dle [44], [45]).

Definice RST je následující:

Množina *X* je ostrá (vzhledem k *R*), pokud je okrajová oblast *X* prázdná. Množina *X* je hrubá (nejasná, pokud jde o *R*) v případě, že hraniční region množiny *X* je neprázdný.

Soubor je tedy hrubý (nepřesný), pokud má nepropustnou hraniční oblast. Jinak je sada ostrá (přesná). To je přesně myšlenka neurčitosti, kterou navrhl Frege.

Aproximacemi a hraniční oblast může být definována přesněji. K tomu potřebujeme další poznámku: Třída ekvivalence *R* určená prvkem *x* bude označena *R* (*x*). Vztah nerozlišitelnosti v jistém smyslu popisuje náš nedostatek znalostí o vesmíru. [44]

Ekvivalence třídy vztahu nerozlišitelnosti, nazývané granule vytvořené *R*, představují základní část poznatků, jsme schopni vnímat kvůli *R*. Proto s ohledem na vztah nerozlišitelnosti obecně jsme schopni pozorovat jednotlivé objekty, ale jsme nuceni rozumět pouze dostupným granulím znalostí. Formální definice aproximací a hraniční oblasti jsou následující:

$$R_*(x) = \bigcup_{x \in U}\{R(x): R(x) \subseteq X\} \quad \text{pro } R \text{ – dolní aproximaci } X \qquad (2.3)$$

$$R^*(x) = \bigcup^{x \in U}\{R(x): R(x) \cap X \neq \emptyset\} \quad \text{pro } R \text{ – horní aproximaci} \qquad (2.4)$$

$$RN_R(X) = R^*(X) - R_*(X) \quad \text{pro } R \text{ – hraniční region } X \qquad (2.5)$$

Jak lze vidět z definic, jsou vyjádřeny z hlediska granulí znalostí. Dolní aproximace množiny (2.3) je spojení všech granulí, které jsou zcela zahrnuty v sadě; Horní aproximace – spojení všech granulí, které mají s prázdnou soustavou průsečík (2.4); Hraniční oblast množiny je rozdíl mezi horní a dolní aproximací (2.5).

Je zajímavé porovnat definice klasických množin, FST a RST. Klasická množina je primitivní pojem a je definována intuitivně nebo axiomaticky. Fuzzy množiny jsou definovány použitím funkce fuzzy členství, která zahrnuje pokročilé matematické struktury, čísla a funkce.

RST jsou definovány aproximací. Tato definice tak také vyžaduje pokročilé matematické pojmy.

Aproximace mají následující vlastnosti:

1) $R_*(X) \subseteq X \subseteq R^*(X)$ (2.6)
2) $R_*(\emptyset) = R^*(\emptyset) = \emptyset;\ R_*(U) = R^*(U) = U$ (2.7)
3) $R^*(X \cup Y) = R^*(X) \cup R^*(Y)$ (2.8)
4) $R_*(X \cap Y) = R_*(X) \cap R_*(Y)$ (2.9)
5) $R_*(X \cup Y) \supseteq R_*(X) \cup R_*(Y)$ (2.10)
6) $R^*(X \cap Y) \subseteq R^*(X) \cap R^*(Y)$ (2.11)
7) $X \subseteq Y \rightarrow R_*(X) \subseteq R_*(Y)\ ;\ R^*(X) \subseteq R^*(Y)$ (2.12)
8) $R_*(-X) = -R^*(X)$ (2.13)
9) $R^*(-X) = -R_*(X)$ (2.14)
10) $R_*R_*(X) = R^*R_*(X) = R_*(X)$ (2.15)

RST jsou v modelu využity k samotné klasifikaci a k porovnání s ostatními metodami. Zároveň jejich výstup v podobě pravidel lze využít jako podklad pro stanovení vstupních atributů do dalších klasifikačních metod, nebo tvorbu pravidel u metod, které s pravidly pracují (FST – FIS). Poslední možností je v případě hybridního systému řídit výstupy pomocí pravidel, která RST vygenerují.

3 KLASIFIKAČNÍ MODEL KVALITY OVZDUŠÍ

Rizika spojená s nekvalitním ovzduším jsou jednou z hlavních environmentálních hrozeb [21], se kterou se nevypořádají pouze jednotlivé regiony a státy, ale i mezinárodní organizace. Tento model se soustředí na prachové částice PM_{10} (dále jen jako PM_{10}), které sebou nesou rizika respiračních onemocnění. Dle [56] především malé děti mohou být postiženy astmatem nebo chronickým zánětem horních cest dýchacích. PM_{10} sebou také nesou i karcinogenní látky značně zvyšující riziko rakoviny. Prachové částice se navíc podílejí 96% na překročeních imisních limitů v ČR. Jejich vliv je tedy zcela zásadní, jak se uvádí [8], [10], [41].

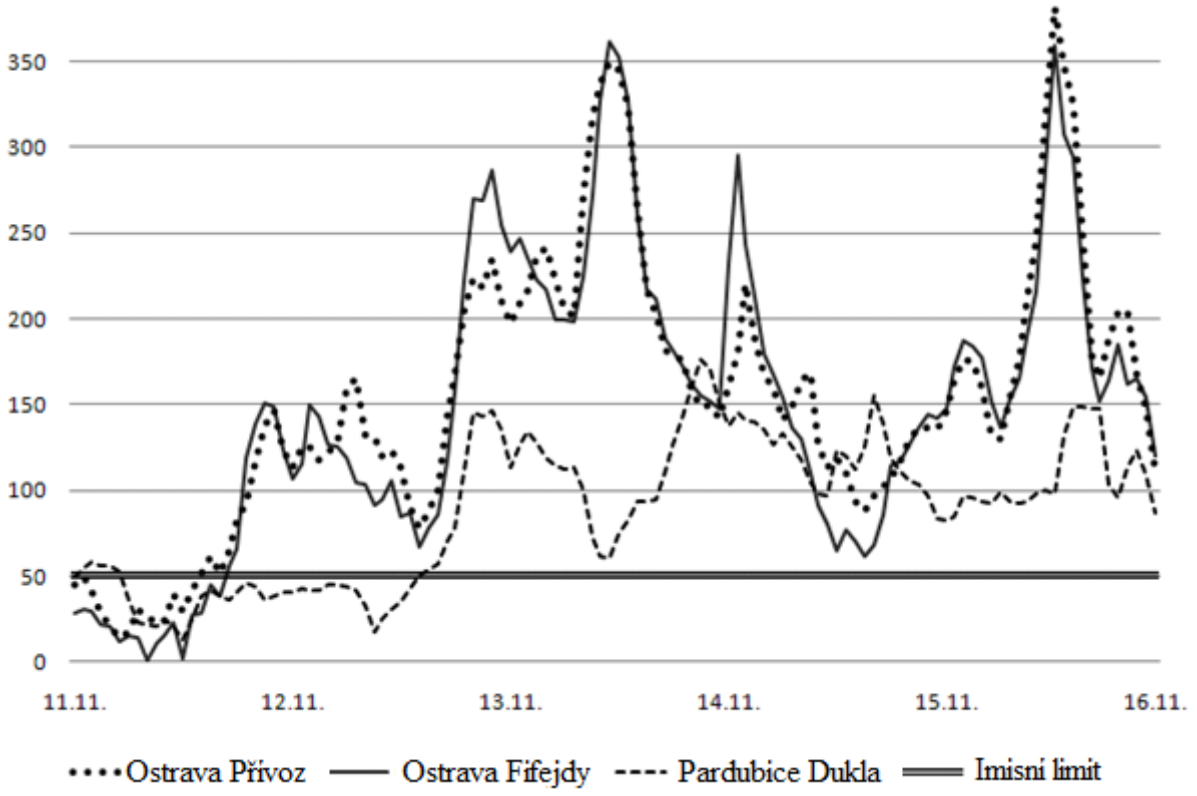
Situace v jednotlivých místech je závislá na lokálních podmínkách. Příkladem tohoto stavu je Moravskoslezský region, který je ojedinělý v koncentraci velkých stacionárních zdrojů znečištění (zdroje zahrnuté v databázích REZZO 1 a 2). Jejichž podíl na znečištění nepodléhá krátkodobým výkyvům. Vytvářejí tak lokální hladinu pozadí, díky které jsou stavy znečištění výrazně lépe identifikovatelné. Dalšími zdroji jsou doprava, kde PM_{10} vznikají při nedostatečném spalování paliva v motorech aut. Dalším zdrojem je lokální vytápění domů. Vzhledem k častému výskytu smogových situací je tak Ostravsko vhodné pro vytvoření obecného modelu, který bude s úpravami možné převést i na jiné regiony ČR.

3.1 Formulace řešeného problému kvality ovzduší

V letech 2006, 2008, 2009 a 2010 byly několikrát vyhlášeny stavy regulace pro Moravskoslezský kraj. 15. listopadu 2011 se s tímto problémem potýká velká část území České republiky, včetně Pardubicka. Smogovou situaci definuje ustanovení § 8 odst. 1 zákona [62] jako „*stav mimořádného znečištění ovzduší, kdy úroveň znečištění ovzduší znečisťující látkou překročí zvláštní imisní limit stanovený prováděcím právním předpisem*“ s tím, že „*zvláštním imisním limitem podle odstavce 1 se rozumí taková úroveň znečištění ovzduší, při jejímž překročení hrozí již při krátké expozici riziko poškození lidského zdraví nebo poškození ekosystému*“ (ustanovení § 8 odst. 2 zákona). Imisní limit je v případě PM_{10} 50 μgm^{-3} (v [62], [63]).

Jak je vidět na Obrázku 5, mírné překročení imisních limitů v Pardubicích bylo zaznamenáno 11. 11. 2011. To však nevedlo k signálu upozornění. Ten je vyhlašován, pokud koncentrace suspendovaných částic PM_{10} průměrně za posledních 24 hodin překročí limit 100 μgm^{-3}. Na delší dobu byly překročeny limity až 13. 11., kdy začaly dlouhodobě přesahovat 100 μgm^{-3}. Na Ostravsku hodnoty rostly již od 12. 11. a během 13. 11. dosahovaly hodnoty 200 a více μgm^{-3}.

Varování obyvatel pak probíhalo v Pardubickém kraji následovně. Hodnoty vedoucí k vyhlášení signálu upozornění byly dosaženy 13. 11. v 15:00, kdy průměr za posledních 24 hodin byl 101,52 μgm^{-3}. Naměřené hodnoty jsou dostupné se 3 hodinovým zpožděním na portálu Českého hydrometeorologického ústavu[1] (dále CHMI). Data ale nejsou verifikována. Zpoždění při verifikaci má za následek i zpoždění při informování veřejnosti. Pardubická radnice tiskové prohlášení o smogové situaci (viz. [43]) vydala až následující den, tedy 14. 11. 2011. Dalším zdrojem informací pro občany je regionální rozhlas, především veřejnoprávní. O této situaci informoval Český rozhlas Pardubice 14. 11. v 10:36 (viz [16]). Tento stav neodpovídá potřebám obyvatelstva, především pro seniory a malé děti je vystavení hodnotám vyšším jak 100 μgm^{-3} (dvojnásobek povolených norem) velkým rizikem. Při současném systému výstrahy může zpoždění dosáhnout až dvou dnů, přičemž v exponovaných místech a časech je možné, aby imise dlouhodobě přesahovaly 150 μgm^{-3}.



Obrázek 5: Množství imisí PM_{10} v [μgm^{-3}] (osa *y*) ve dnech 11. až 16. 11. listopadu 2011 (osa *x*) na měřících stanicích v Ostravě a Pardubicích. Zdroj: vypracováno dle [15].

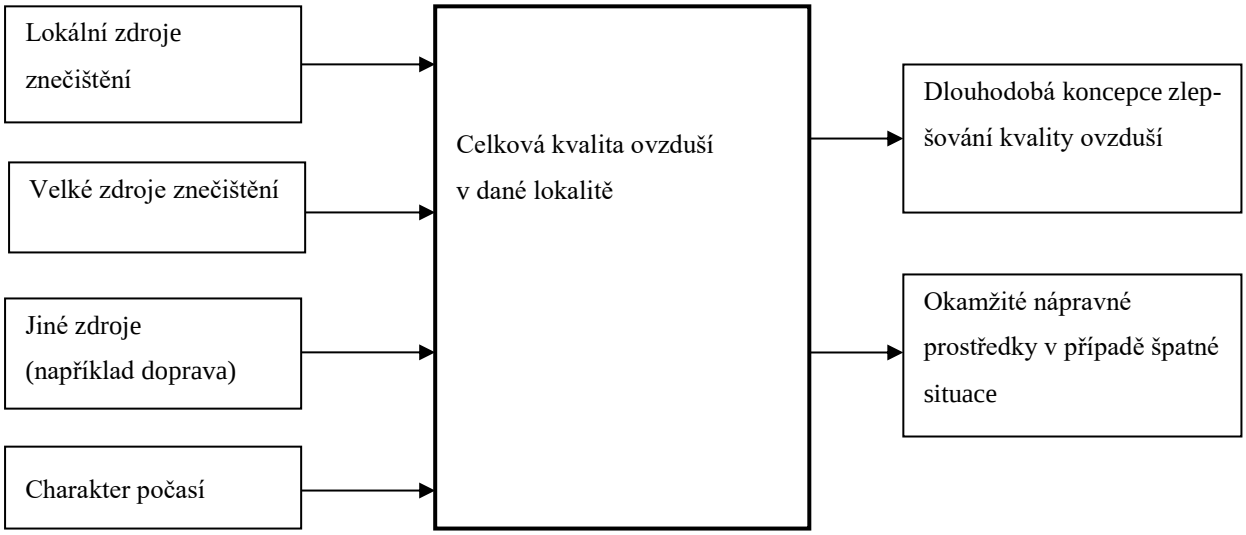
[1] Adresa měření: http://pr-asu.chmi.cz:8080/IskoAimDataView/faces/viewChart.xhtml

Jednou z možností zlepšení současného stavu je zpracování modelu kvality ovzduší, s jehož pomocí by bylo možné klasifikovat současný stav a tím i odhadnout následující vývoj. Je možné využít jak naměřené hodnoty PM_{10}, tak i charakter počasí včetně předpovědi. To má velký podíl na zhoršení kvality ovzduší. Pokud by daný model vykazoval vysokou míru přesnosti, bylo by možné ho využít pro informování obyvatel.

3.2 Model kvality ovzduší

Kvalitu ovzduší primárně ovlivňují dvě oblasti – zdroje znečištění a charakter počasí. Jelikož počasí není možné ovlivnit, tak je nutné se zaměřit při prevenci na zdroje znečištění. Jednoduchý model kvality ovzduší je navržen na Obrázku 6.

Velké zdroje znečištění vytváří vyšší normální hladinu pozadí. Jsou však dlouhodobě sledovány a v minulých dekádách obvykle prošly velkými investice i do nových ekologičtějších technologií ([23]). Navíc v případě extrémní situace mohou být regulovány (pokud jsou na území ČR), i přestože je taková regulace velice nákladná.



Obrázek 6: Model kvality ovzduší v dané lokalitě. Zdroj: autor.

Lokální zdroje (domácnosti, malí znečišťovatelé) mají výrazný podíl na situaci v nejbližším okolí a velmi těžce se regulují. Je zde předpoklad korelace mezi množství vypouštěných spalin a počasím (nejnižší noční teplotou). Na rozdíl od velkých zdrojů se zde zanedbala prevence. Technologické výměny přitom nejsou ve srovnání s technologiemi pro velké znečišťovatele příliš drahé. Zde bohužel stát nenaplnil regulační funkci.

Mezi dalšími zdroji znečištění je dominantní doprava. Tu lze regulovat dlouhodobě pomocí regulace překonaných technologií (dieselových motorů, zrušení přimíchávání biopaliv první

generace do benzínu ad.) a vývojem a podporou nových typů pohonu (elektrický, vodíkový atd.), a krátkodobě pak na příklad: jízdou pomocí MHD zdarma, omezení vjezdu do problémových lokalit. Více o znečištění je v [8], [11], [23].

Tabulka 1: Typy znečištění dle lokalit Ostravsko a Pardubice. Zdroj: [8], [11], [23].

Typ znečištění	Podíl znečištění - Ostrava v [%]	Podíl znečištění - Pardubice v [%]	Možnost řešení
Velcí znečišťovatelé – elektrárny, továrny	30 – 60	10 – 25	Krátkodobé: Omezení výrobních kapacit na postiženém území Dlouhodobé: Využití ekologičtějších technologií, optimalizace výroby
Malí znečišťovatelé – lokální topení	30 – 50	30 – 60	Krátkodobé: - Dlouhodobé: Dotační programy pro exponovaná místa – výměna kotlů, zateplení
Doprava	15 – 30	30 – 45	Krátkodobé: MHD zdarma, omezení vjezdu do exponovaných oblastí Dlouhodobé: omezení nekvalitních technologií, pohon na elektřinu, vodík nebo jiná ekologická paliva

Podíly jednotlivých znečištění v lokalitě Ostrava a lokalitě Pardubice jsou zobrazeny v Tabulce 1. Jedná se o relativní hodnoty, takže vysoký podíl neznamená automaticky velkého znečišťovatele, ale pouze jeho podíl na imisní situaci. Z Tabulky 1 a Obrázku 6 lze taktéž vysledovat základní rozdíl mezi Pardubicemi a Ostravou. Kromě vyšších „pozaďových“ hodnot znečištění, je rozdíl taktéž v hlavním zdroji celkového znečištění, který vede k překračování hodnot. Zatímco na Ostravsku to je lokální vytápění, které vede ke kulminacím v brzkých ranních hodinách, kdy je nejnižší denní teplota. V Pardubicích je to dopravní situace, která vede ke dvěma kulminačním bodům denně, a to okolo osmé hodiny ranní a 16 až 17 hodiny odpolední.

Role počasí v rámci krátkodobé kvality ovzduší je taktéž zcela zásadní. U velkých zdrojů předpokládáme jen částečnou závislost množství vypouštěných škodlivin na počasí. U malých zdrojů a dopravy však lze předpokládat vysokou korelaci s teplotními charakteristikami [8]. Dále bude na kvalitu ovzduší nepříznivě působit inverzní charakter počasí a nízká rychlost větru. Ty způsobují, že se zdraví škodlivé látky hromadí na exponovaných místech [23].

3.3 Popis dat a jejich předzpracování

Jako indikátory slouží hodnoty naměřené v určitých oblastech, Data o počasí jsou získána z meteorologických stanic na letištích Ostrava – Mošnov a Pardubice. Dále byly využity údaje ze stacionárních stanic měřících znečištění. Pro Ostravu to byly stanice: Bartovice, Českobratrská,

Fifejdy, Mariánské Hory, Poruba, Přívoz, Zábřeh. V Pardubicích se nalézá jen jedna stanice měřící PM_{10} a to je stanice Dukla. Vzhledem k nízkému počtu měřících stanic ve městech s nižší normální hladinou znečištění [15] jako jsou právě například Pardubice nebo Hradec Králové, musí být daný model s určitým stupněm zobecnění. V jednotlivých lokalitách je možné pozorovat určité rozdíly (více viz [29]). Modelováním lze zkoumat jednotlivé faktory a jejich dopad na konkrétní imisní situaci. V Pardubicích bylo provedeno šetření a výstupy jsou v [29]. Šetření bylo prováděno s přesnými daty odpovídajícími místním podmínkám. Díky těmto měřením lze nejdříve vytvořit model nad daty z Ostravy, kde jich je dostatečné množství a ten následně pomocí měření v konkrétních oblastech aplikovat i pro ostatní regiony. Model využívá 17 původních atributů a 3 atributy odvozené, které jsou popsány v následující tabulce.

Tabulka 2: vybrané proměnné modelu množství prachových částic PM_{10}. Zdroj: [37].

Název proměnné	Zkratka	Typ proměnné
Stanice	l_1	Množina – evidenční údaj
Typ stanice	l_2	Množina
Den v roce	c_1	Diskrétní proměnná
Den v týdnu	c_2	Diskrétní proměnná
Průměrná/Maximální/Minimální denní teplota ve městě	$m_1/m_2/m_3$	Spojitá proměnná
Průměrná/Maximální /Minimální denní teplota na stanici Lysá Hora	$m_4/m_5/m_6$	Spojitá proměnná
Průměrná/Maximální rychlost větru	m_7/m_8	Spojitá proměnná
Směr větru – v dopoledních / odpoledních hodinách	m_9/m_{10}	Diskrétní proměnná
Vlhkost vzduchu / Atmosférický tlak	m_{11}/m_{12}	Spojitá proměnná
Průměrný / maximální / minimální rozdíl v denní teplotě změřený na stanicích ve městě a na Lysé Hoře	$d_1/d_2/d_3$	Spojitá proměnná
PM_{10} - průměrná hodnota za posledních 24 hodin	d_4	Spojitá proměnná

Pro klasifikaci prachových částic se využívají atributy: l_1, l_2, m_1, m_2, m_3, m_4, m_5, m_6, m_7, m_8, m_9, m_{10}, m_{11}, m_{12}. Celkově bylo v letech 2006 až 2011 vybráno 451 pozorování. Trénovací a testovací množiny obsahovaly dohromady 366 pozorování z roku 2008. Validační množinu tvořily data z let 2006, 2009, 2010 a 2011, kdy docházelo k překračování imisních limitů a vyhlašování stavů regulace. Kompletní data z let 2007 a 2008 nebyla v době výzkumu k dispozici. Popis dat lze nalézt v Přílohách 1 až 5.

Při vytváření modelu se postupovalo vyhodnocením výsledků uvedených v [29]. Proto byly vybrány proměnné, které jsou pro množství prachových částic relevantní. Vyřazeny byly proměnné l_1, l_2, c_1, c_2, m_9 a m_{10}, které se ukázaly jako málo relevantní při použitém stupni zobecnění. Z parametrů m_1 – m_6 byly dále odvozeny 3 proměnné: rozdíl v průměrných denních teplotách mezi meteorologickou stanicí na letišti v Ostravě Mošnově nebo v Pardubicích proti teplotě na meteorologické stanici na Lysé Hoře přepočtený na sto metrů výšky d_1, rozdíl v minimálních denních teplotách mezi meteorologickou stanicí na letišti v Ostravě Mošnově nebo

v Pardubicích proti teplotě na meteorologické stanici na Lysé Hoře přepočtený na sto metrů výšky d_2 a rozdíl v maximálních denních teplotách mezi meteorologickou stanicí na letišti v Ostravě Mošnově nebo v Pardubicích proti teplotě na meteorologické stanici na Lysé Hoře přepočtený na sto metrů výšky d_3. Tyto odvozené proměnné zaznamenávají výskyt inverzního charakteru počasí. Dále byly využity hodnoty naměřené na jednotlivých stanicích, které měří množství imisí. Hodnotící stupnice kvality ovzduší uváděná CHMI [15] je v Tabulce 3. Tato stupnice byla využita nejenom pro kategorizaci výstupních hodnot v čase *t*, ale taktéž pro kategorizaci průměrných hodnot za posledních 24 hodin (čas *t*-1).

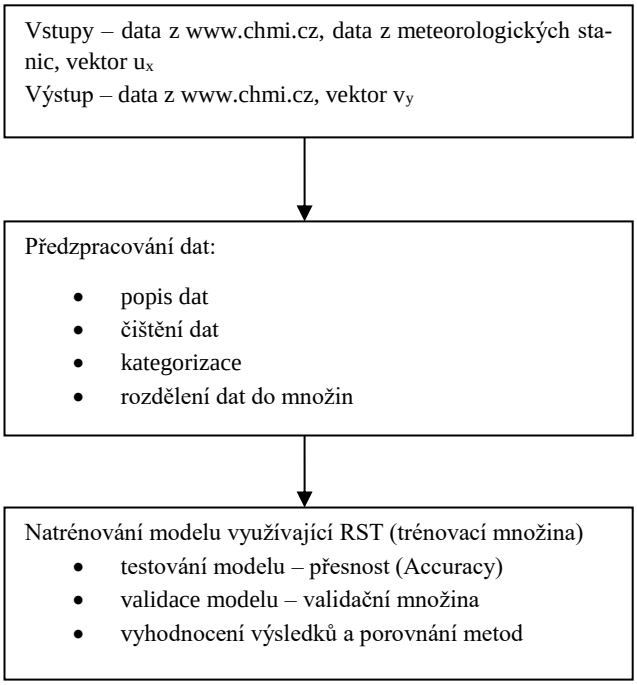
Tabulka 3: Stupnice kvality ovzduší. Zdroj: [15].

Index	**Kvalita ovzduší**	**PM_{10} v [μgm^{-3}] za 1 hodinu**
1	velmi dobrá	0 - 15
2	dobrá	> 15 - 30
3	uspokojivá	> 30 - 50
4	vyhovující	> 50 - 70
5	špatná	> 70 - 150
6	velmi špatná	> 150

Dále bylo nutné data předzpracovat. Bylo opraveno šest hodnot souboru teplot z meteorologické stanice na Lysé Hoře, kde při předchozím zpracování, popř. již při zaslání byla chyba v rozdílném znaménku u denních teplot. Tato úprava byla ověřena pomocí online záznamu a také pomocí teplot z jiných stanic. Pro jednotlivé metody pak došlo buď ke kategorizaci primárně pomocí ekvifrekvenčního škálování, nebo k normalizaci a standardizaci dat.

3.4 Modelování pomocí Rough množin

Postup modelování pomocí RST je zobrazen na Obrázku 7. RST popsaná [26], [45], [47], je založena na vyhledávání společných charakteristik dat. Zároveň pracují s neurčitostí, a to v rámci hranic (horní a dolní aproximace). Tyto hranice oddělují data, jež ukazují k danému jevu (výsledku) od dat, která neposkytují přesné určení jevu. Na Ústavu systémového inženýrství a informatiky byly již provedeny další výzkumy (v [26], [29]), které potvrdily možnost využití RST při modelování kvality ovzduší.



Obrázek 7: Postup modelování pomocí RST. Zdroj: [38].

Pro potřeby použití metody RST bylo potřeba vytvořit ze stávajících proměnných odvozené kategorizované proměnné k_1 – k_6 (viz. Tabulka 4) pomocí ekvifrekvenčního škálování kromě proměnné k_6. Jejich hodnoty a množství kategorií bylo stanoveno experimentálně, popřípadě vychází z popisu jevů podle [15]. Odhadovaný parametr v_y, průměrná hodnota PM_{10} v následujících 24 hodinách, má pouze dvě kategorie. Jako porovnávací metoda byly využity TDIDTs C5 a CRT [26]. Ty jako vstupní hodnoty využívají spojité proměnné. Z toho důvodu byly využity původní proměnné popsané v Tabulce 2 a taktéž v Příloze 1. Cílem všech těchto metod bylo zjistit, jaké parametry a jejich hodnoty vedou ke špatné kvalitě ovzduší.

Tabulka 4: Odvozené proměnné použité při výpočtu metodou RST. Zdroj: [38].

Atribut	**Kategorizace spojitých hodnot do kategorií 0 až 4**				
	0	**1**	**2**	**3**	**4**
k_1 – denní průměrná rychlost větru v [m/s]	< 9	9–13	>13–17	>17–20	> 20
k_2 – maximální rychlost větru v [m/s]	< 9	9–14	>14–23	> 23	
k_3 – inverzní charakter počasí ve [stupních Celsia] na každých 100 metrů rozdílu výšek	< 2	2 - 5	> 5		
k_4 – vlhkost v [%]	< 66	66 - 76	>76–86	> 86	
k_5 – tlak v [hpsc]	< 1005	1005–1014	>1014–1020	> 1020	
k_6 – průměrné množství PM_{10} v uplynulých 24 hodinách v [μgm^{-3}] za 1 hodinu	0–22	22–37	>37–70	> 70	
v_y – průměrné množství PM_{10} v následujících 24 hodinách v [μgm^{-3}] za 1 hodinu	< 70	> 70			

Pro určení parametru v_y bylo pomocí softwaru RSES [52] vygenerováno 21 pravidel. Pravidla se generovala za pomoci algoritmu LEM2 [20]. Získaná pravidla ukazují důležitost jednotlivých atributů. Nejdůležitějšími parametry jsou inverze, rychlost větru a průměrné množství PM_{10} v uplynulých 24 hodinách. Tyto parametry obsahuje všech devět pravidel určujících negativní hodnotu v_y. Nejsilnější negativní pravidlo je IF $k_1 = 0$ AND $k_3 = 1$ AND $k_6 = 2$ THEN $v_y = 1$. Z toho vyplývá i logicky odůvodnitelná úvaha, že v případě inverzního charakteru počasí a malé rychlosti větru zůstávají emise koncentrovány na místě a dále zhoršují kvalitu ovzduší v dané lokalitě. Úspěšnost predikce dosahovala 96,4 % (výsledná matice záměn – dále jako CM – viz Tabulka 5).

Tabulka 5: Matice záměn predikované hodnoty pomocí RST na testovací množině. Zdroj: [38].

		predikovaná hodnota	
		0	1
zjištěná hodnota	0	65	2
	1	1	15

Výsledek 0,964 byl porovnán s výsledky TDIDTs (vychází z práce [28]) a NNs. Tyto pracují jak se spojitými daty, tak s daty kategorizovanými. U kategorizovaných dat se použily proměnné popsané v Tabulce 5. Z vytvořených rozhodovacích stromů dva měly přesnost vyšší než 90 %, jeden využívající algoritmus $C5_{boost}$ a druhý algoritmus CRT_{boost}. U NNs byla data testována u dvou typů, a to MLP a RBF. Srovnání výsledků s RST zobrazuje Tabulka 6.

Tabulka 6: Porovnání výsledků RST s TDIDTs a NNs. Zdroj: [38].

Metoda	Přesnost (testovací množina) v [%]	Přesnost (validační množina) v [%]
RST	96,4	90,7
NN – RBF	95,2	90,8
NN – MLP	96,2	91,3
$C5_{boost}$	96,5	89,4
CRT_{boost}	93,8	84,7

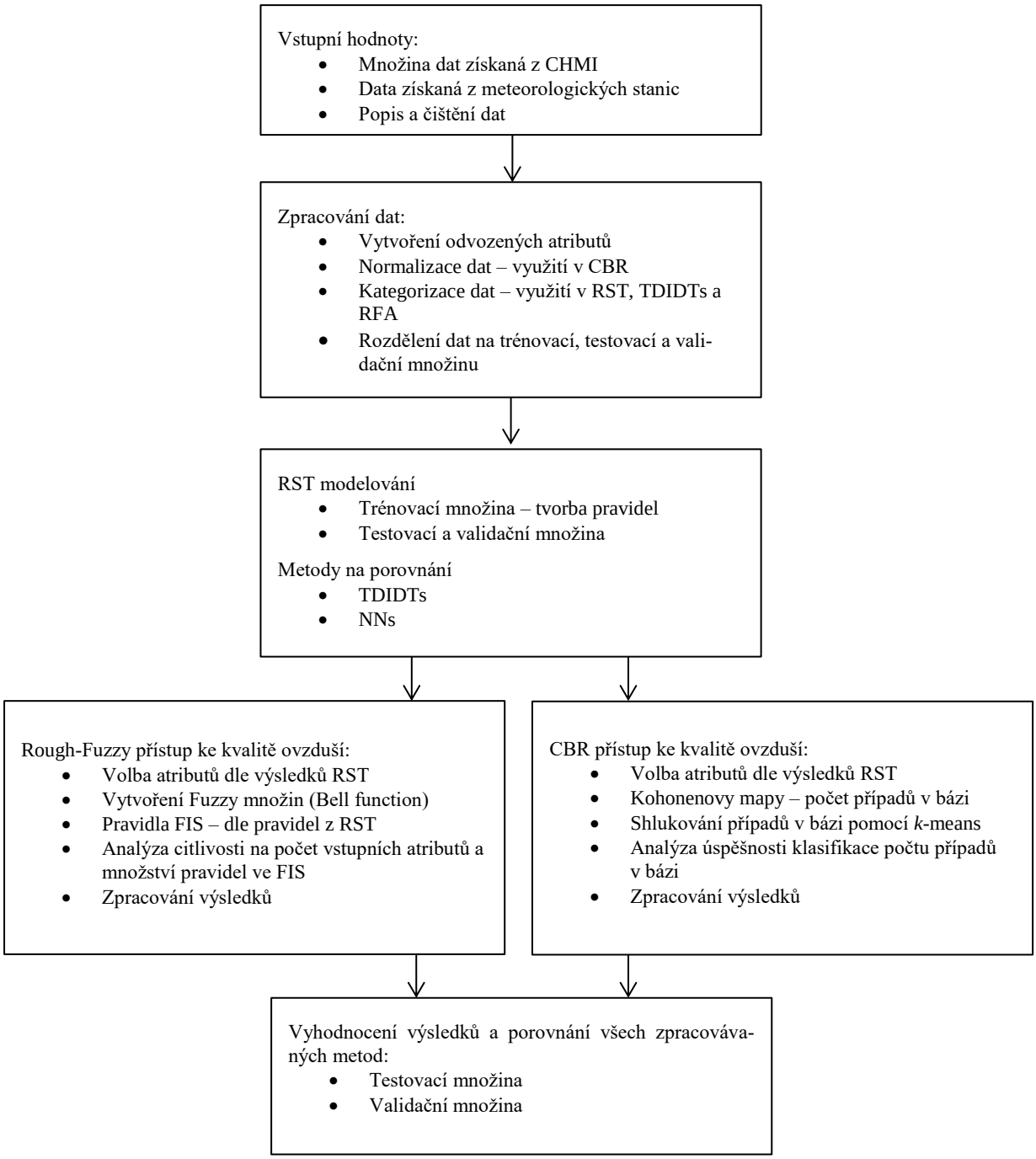
Zároveň bylo provedeno ověření výsledků na validační množině (viz Tabulka 6). Jejím specifikem byl výrazně vyšší poměr dní s negativní kvalitou ovzduší. To má zřejmě za následek nižší úspěšnost zkoumaných metod. Predikční schopnost RST a metody C5 je stále vysoká, přesto se RST jeví jako robustnější a vhodnější nástroj. Dále nebyly zjištěny výraznější rozdíly ve vlivu počasí na smogovou situaci Ostravy (přesnost 0,908) a Pardubic (přesnost 0,899). Stále zůstávají nejdůležitějšími parametry: inverzní charakter počasí a průměrná rychlost větru [38].

Kromě metody CRT, která měla na validační množině výrazně horší výsledky, jsou všechny výše uvedené metody vhodné pro modelování. Výhodou TDIDTs a RTS je dále to, že jejich výstupy v podobě pravidel lze využít pro modelování pomocí dalších metod. [38]

3.5 Modelování pomocí CBR a kombinace rough-fuzzy množin

Při využití CBR a kombinovaného RFA v rámci modelu ovzduší byly využity poznatky RST popsané v předchozí kapitole a výzkum uvedený v [27].

Návrh RST navazuje na předchozí práce [26], [27] a [28], které potvrdily možnost využití RST při modelování kvality ovzduší. Díky RST byl redukován počet parametrů modelů na 6 proměnných. Těchto 6 kategorizovaných proměnných $\{k_1, k_2, k_3, k_4, k_5, k_6\}$ bylo odvozeno z původních proměnných $\{m_7, m_8, d_1, m_{11}, m_{12}, d_4\}$ (viz Tabulka 4) pomocí ekvifrekvenčního škálování s výjimkou proměnné k_6. Počet kategorií v rámci jednotlivých atributů byl nastaven experimentálně na základě výsledků [15]. V rámci rozvoje tématu byl stanoven následující model zobrazený na Obrázku 8.



Obrázek 8: Návrh modelu využití CBR a RFA a jeho analýza. Zdroj: autor, upraveno dle [37].

3.5.1 Modelování na bázi rough-fuzzy přístupu

V rámci RFA byly kategorizované vstupní atributy namodelovány pomocí zvonové funkce příslušnosti (Bell function) a to tak, že v hraniční hodnotě měly obě funkce příslušnosti hodnotu 0,5 (zobrazení v programu MATLAB viz Přílohy 6 až 8). Takto byl vytvořen požadovaný počet funkcí příslušnosti jednotlivých atributů, který odpovídal počtu kategorií u daného atributu (viz

Tabulka 7). Celkový počet funkcí příslušnosti pro všechny vstupní atributy tak byl 24. Počet pravidel FIS tak byl 3840.

Výstupní atribut v_y, průměrná hodnota prachových částic PM_{10} v následujících 24 hodinách, má opět pouze dvě kategorie (Příloha 8), z nichž první odpovídá stavu, kdy množství imisí nepředstavuje riziko pro zdraví obyvatelstva, a není tedy důvodem pro vyhlášení regulatorních stavů a druhý množství imisí, kdy by měla být zahájena omezení. Byly opět vytvořeny funkce příslušnosti výstupního atributu totožně se vstupními atributy (zobrazení v programu MATLAB viz Přílohy 6 a 7). Výsledná přesnost 93,4 % na testovací množině a 91,2 % na validační množině je dobrá a na validační množině zpřesňuje původní metodu RST.

Tabulka 7: Výsledky RFA dle počtu vstupních atributů. Zdroj: autor.

Počet atributů	**Počet pravidel**	**Přesnost kvalifikace v [%] – testovací množina**	**Přesnost kvalifikace v [%] – validační množina**
6	3840	93,4	91,2
5	960	93,1	90,7
4	240	92,7	90,3
3	60	91,4	88,6

Následně došlo k analýze toho, jak odebrání atributu a snížení počtu pravidel ovlivní přesnost klasifikace. Nejprve byla určena váha jednotlivých atributů pomocí pravidel vytvořených RST. Na základě počtu případů zahrnutých do daného pravidla se stanovila váha pravidla, přičemž součet vah pravidel, ve kterých byl daný atribut využit, byl následně brán jako váha daného atributu. Atributy k_1, k_3 a k_6 (průměrná denní rychlost větru, inverzní charakter počasí, průměrné množství PM_{10} v uplynulých 24 hodinách), které vykázaly podobnou váhu, se dle jejich váhy v RST zdály v modelu nezastupitelné. Nejmenší váhu měl atribut k_5 (atmosférický tlak), následuje k_2 (maximální denní rychlost větru) a k_4 (vlhkost vzduchu). Průměrné snížení klasifikace při odebrání jednoho atributu bylo na testovací množině 0,667 % a 0,867 % na validační množině, přičemž největší rozdíl na validační množině je po odebrání posledního ze tří méně významných atributů (pokles o 1,7 %). Z toho vyplývá, že pro kvalitní klasifikaci jsou v tomto případě podmínkou minimálně 4 atributy, přičemž odebrané atributy se získávají z totožných zdrojů jako 3 základní, tudíž jejich zpracování není náročné a v rámci modelu je vhodné s nimi dále počítat.

3.5.2 Klasifikace pomocí CBR

Výsledky RST byly rovněž použity k přípravě modelování pomocí CBR, kde se pomocí nich stanovily vstupní proměnné. Zde bylo experimentováno s množstvím případů v bázi pravidel a s jejich obměnou. Využito bylo vlastnosti CBR:

"*V případě, že v bázi případů je k dispozici jen malé množství případů, CBR může pracovat jen s těmito pár známými případy a postupně budovat znalost tak, jak jsou do báze přidávány další případy. Přidávání nových případů způsobí expanzi znalostí ve směrech, které jsou podmíněny nově řešenými případy*." [42, s. 10]

V rámci trénovací množiny byly vybrány případy, které nejvíce reprezentují skupiny ostatních případů v bázi. Dle rozdělení v bázi to bylo osm (z toho 2 byly negativní výstupní hodnotu), čtrnáct (4 měly negativní výstupní hodnotu) a dvacet šest případů (6 měly negativní výstupní hodnotu). Zbylé případy z roku 2008 byly brány jako testovací množina (dále TM). Validační množinu (VM) tvořilo vždy 84 případů z let 2006–2011. Výsledky jsou zobrazeny v následujících tabulkách – maticích záměn.

Tabulka 8: Výsledková matice pro 8 případů v bázi. Zdroj: autor.

CM - 8 případů v bázi (358 TM a 84 VM)		**Testovací množina – 93,57 %**		**Validační množina – 83,33 %**	
		Naměřená hodnota			
		0	1	0	1
Predikovaná hodnota	0	311	8	33	8
	1	16	24	6	37

Tabulka 9: Výsledková matice pro 14 případů v bázi. Zdroj: autor.

CM – 14 případů v bázi (351 TM a 84 VM)		**Testovací množina – 95,16 %**		**Validační množina – 86,90 %**	
		Naměřená hodnota			
		0	1	0	1
Predikovaná hodnota	0	312	6	35	7
	1	11	22	4	38

Tabulka 10: Výsledková matice pro 26 případů v bázi. Zdroj: autor.

CM – 26 případů v bázi (339 TM a 84 VM)		**Testovací množina – 94,69 %**		**Validační množina – 91,67 %**	
		Naměřená hodnota			
		0	1	0	1
Predikovaná hodnota	0	301	4	36	4
	1	14	20	3	41

Dále byla testována možnost vytvořit umělé případy v bázi případů. K jejich tvorbě se data nejdříve rozdělila dle výstupní hodnoty a pak se vypočítaly shluky v rámci jednotlivých skupin.

Hodnoty se v rámci těchto shluků zprůměrovaly a byly tak vytvořeny umělé případy. Ty se pak staly trénovací množinou. Výsledky však ukázaly, že jejich vznik nezlepšuje, ale naopak zhoršuje klasifikaci. Výsledky je možné porovnat v Tabulce 11.

Tabulka 11: Porovnání úspěšnosti klasifikace CBR. Zdroj: autor.

Počet případů	Přesnost v [%] – testovací množina	Přesnost v [%] – validační množina	Přesnost v [%] – testovací množina – umělé případy v bázi	Přesnost v [%] – validační množina – umělé případy v bázi
8	93,57	83,33	93,29	80,95
14	95,16	86,90	94,30	85,71
26	94,69	91,67	94,40	90,48

Dalším krokem bylo testování různých metod výpočtu vzdálenosti mezi novým případem a případem v bázi. K Euklidovské metrice byly doplněny další metriky: Manhattan a Čebyševova. Kromě toho bylo testováno odebírání jednotlivých atributů a jejich vliv na přesnost modelu (citlivostní analýza). Přehled jednotlivých výsledků je v Příloze 10. Z celkových 42 pozorování byla Euklidovská metrika jako nejlepší vyhodnocena 22-krát, Manhattan 15-krát a Čebyševova 6-krát, přičemž se dá říci, že s množstvím odebraných atributů se srovnává poměr mezi Manhattan a Euklidovskou metrikou jako nejvhodnější metrikou pro výpočet vzdálenosti na tomto typu dat.

Nejlepší výsledky pro určitý počet odebraných atributů se jsou v Tabulce 12. Pokud je pod konkrétním atributem uvedena 1, znamená to, že je ve výpočtu vzdálenosti obsažen, pokud je zapsána 0, tak obsažen není.

Tabulka 12: Výsledky citlivostní analýzy CBR na testovací množině (více viz Příloha 2). Zdroj: autor.

Počet vstupů	Vstupní atributy						Přesnost modelu CBR v [%] dle využité metriky		
	k_1	k_2	k_3	k_4	k_5	k_6	Euklidovská metrika	Manhattan metrika	Čebyševova metrika
6	1	1	1	1	1	1	**94,69**	93,51	93,81
5	0	1	1	1	1	1	**94,69**	94,40	94,40
4	0	1	1	1	0	1	**94,40**	94,40	93,51
4	1	1	0	1	0	0	**94,40**	92,63	93,22
3	0	0	1	1	0	1	93,22	**93,81**	92,33
3	0	1	0	1	0	0	**93,81**	93,22	92,92

Z citlivostní analýzy vyplývá, že nejméně citlivý byl model na odebrání parametru k_1 (průměrná rychlost větru). Tento atribut byl také jako jediný vyhodnocen jako možná nadbytečný pro model CBR s dalšími pěti atributy (přestože ostatní modely s ním pracovaly jako s atributem důležitým). Následovaly na parametry k_5 (tlak), k_2 (maximální rychlost větru) a k_3 (inverzní charakter počasí), u kterých však relativně malé rozdíly a jsou tak pro model přibližně stejně

důležité. Další dva parametry se ukázaly jako klíčové, především pak parametr k_4 (vlhkost) je pro model důležitý. O kolik průměrně poklesla kvalita modelu při odebrání jednotlivých parametrů lze nalézt v Tabulce 13.

V tabulce 13 je zároveň porovnání citlivosti na odebrání atributů pro modely RFA a CBR. Citlivost u modelu RFA nebyla zkoumána natolik do hloubky jako u modelu CBR. Příčinou je především složitost vytvoření pravidel pro FIS. Navržený výpočet váhy atributů pomocí pravidel v RST tak nemohl být ověřen a jedná se pouze o experimentální metodu. Při porovnání jednotlivých metod je vidět, že RFA je díky fuzzy množinám odolnější vůči odebrání dalšího atributu, jelikož průměrný pokles na jeden atribut je 0,67 %.

U modelu CBR byly zaznamenány větší poklesy, ač přesné srovnání není možné, protože u modelu RFA nebylo z důvodu obtížného přepracování pravidel ve FIS testováno odebrání atributů, které byly považovány za klíčové. Největší rozdíl v hodnocení byl pak zaznamenán u atributů k_1 a především k_4, který se ve významnosti dostal na první místo.

Tabulka 13: Porovnání analýz citlivosti modelů RFA a CBR v [%]. Zdroj: autor.

Analýzy	**k_1**	**k_2**	**k_3**	**k_4**	**k_5**	**k_6**
Citlivostní analýza dle RST –pořadí důležitosti atributů	3	5	2	4	6	1
Snížení přesnosti modelu RFA odebráním atributu k	N/A	-0,4	N/A	-0,3	-1,3	N/A
Citlivostní analýza dle CBR – pořadí důležitosti atributů	6	4	3	1	5	2
Průměrné snížení přesnosti modelu CBR odebráním atributu k	-2,59	-2,84	-2,96	-3,68	-2,71	-3,30

Na konec byly finální výsledky FSA a CBR porovnány s výsledky RST (viz [38]) TDIDTs and NNs (viz [37]). Z výsledků dosažených pomocí TDIDTs byly vybrány dva ($C5_{boost}$ a CRT_{boost}), jejichž přesnost na testovací množině je vyšší jak 90 %. Dále NNs – MLP se šesti neurony ve skryté vrstvě a RBF s 10 neurony ve skryté vrstvě. Porovnání výsledků na testovací a validační množině je v Tabulce 14.

Výsledky na validační množině vykazují nižší přesnost. To je dáno tím, že ve validační množině je daleko větší průměrný počet dní s negativní kvalitou ovzduší. To vede k nižší přesnosti u všech zkoumaných metod. Přesto je přesnost klasifikace u RFA, CBR, RST, TDIDT $C5_{boost}$ a NNs velmi vysoká. Výběr vhodné metody závisí na cíli, k čemu má sloužit. Pokud je potřeba jednoduchá zpracovatelnost, pak je vhodné využít RST nebo TDIDTs. Naopak, pokud by mělo docházet ke generalizaci modelu nebo jeho použití v rozličných podmínkách, je vhodné využít CBR, RFA nebo NNs.

Nebyly nalezeny další odlišnosti mezi dopadem počasí na smogovou situaci mezi Ostravou (90,8 %) a Pardubicemi (89,9 %). Důležitost parametrů jednotlivých parametrů zůstává stejná pro obě oblasti.

Tabulka 14: Porovnání všech zkoumaných metod v rámci modelu kvality ovzduší. Zdroj: autor.

Metoda	Přesnost – testovací množina v [%]	Přesnost – validační množina v [%]
RST	96,4	90,7
RFA	93,4	91,2
CBR	94,7	91,7
$C5_{boost}$	96,5	89,4
CRT_{boost}	93,8	84,7
NN – MLP	96,2	91,3
NN – RBF	95,2	90,8

3.6 Shrnutí dosažených výsledků pro modely hodnocení kvality ovzduší

V rámci výzkumu byl vyvinut model kvality ovzduší zaměřený na prachové částice PM_{10}, který je možné využít k časnému varování obyvatelstva. Cílem modelu bylo klasifikovat aktuální situaci, abychom byli schopni odhadnout výhled na následujících 24 hodin. Odhad je možný pouze v lokalitách, kde jsou k dispozici historické hodnoty, a tudíž je možné podle nich připravit klasifikační model. Testovány byly dvě rozdílné lokality – Ostravsko a Pardubice.

Nejdříve byl zpracován teoretický model, který identifikoval vstupy a výstupy systému kvality ovzduší a jeho regulace. Jako hlavní vstupy byly určeny aktuální imisní situace, typ lokality a charakter počasí. Následovalo získání dat – meteorologická data z hydrometeorologických stanic na Lysé Hoře a z letišť v Pardubicích a Ostravě Mošnově a data ohledně imisní situace ze stránek CHMI.

Poté došlo na zpracování dat – čištění, kategorizace, popř. normalizace a standardizace. Celkové množství záznamů využitých k výpočtům bylo 451. Tato skupina dat byla rozdělena na trénovací, testovací a validační skupinu. Nejdříve byla zpracována klasifikace pomocí RTS (algoritmus LEM2). Výsledek byl porovnám s metodami klasifikace pomocí TDIDT a NNs. Zároveň se pomocí pravidel, která jsou výstupem této metody, identifikovaly klíčové atributy. Ty byly využity jako vstup při tvorbě pravidel pro FIS v rámci FST a jako atributy pro výpočet vzdálenosti mezi případy v CBR.

Následně byly porovnány výsledky všech výše popsaných metod. Výsledky, jejichž kvalita přesahuje 95 % na testovací množině a 91 % na trénovací množině, lze označit za dostatečně

kvalitní vzhledem k obecnému charakteru modelu. Pro vyšší přesnost by bylo vhodné mít dlouhodobě získávaná lokální data. Např. vzhledem k charakteru imisní situace v Pardubicích by bylo zřejmě vhodné dodat i atribut rozlišující všední a sváteční den právě z důvodu zvýšené dopravy, která má v této lokalitě velký vliv na znečištění.

S možností přidání dalších atributů dle lokality do tohoto modelu se nabízí i případná citlivostní analýza, popř. váhování vstupních atributů. Citlivostní analýzy byly zpracovány dvěma způsoby. Prvním bylo využití pravidel získaných v rámci RST, které posloužily k vypracování pořadí důležitosti jednotlivých atributů. Následně se atributy podle váhy od nejméně důležitého ubíraly z FIS. Jiné pořadí nebylo vyzkoušeno z důvodu dlouhého vytváření pravidel pro FIS.

Naopak v rámci modelu CBR byly vyzkoušeny všechny možné kombinace atributů až do minimálního počtu tří atributů. Zároveň s tím byly testovány kromě Euklidovské ještě další dvě metriky, a to Manhattan a Čebyševova. Pořadí důležitosti atributů vyšlo jinak, než tomu bylo dle výpočtu z pravidel RST. Jelikož se jedná o ověření na kombinacích jednotlivých vstupních atributů a celkových výsledcích, je možné takto získanou váhu atributů považovat za robustnější, než je tomu u váhy vypočtené dle vah pravidel RST. Dále citlivostní analýza ukázala jednu zajímavou věc. S počtem odebraných atributů se srovnávají výsledky vzdáleností získané pomocí Manhattan a Euklidovské metriky, která je při vyšším počtu atributů je vhodnější.

Samotný cíl, tedy vytvoření obecného modelu a porovnání jednotlivých metod klasifikace, byl splněn. Došlo k eliminaci atributů s nízkou váhou pro obecný model a zbylé atributy byly porovnány pomocí citlivostní analýzy. Výsledky, které byly publikované v [37] a [38], jsou využitelné pro prevenci a řešení případných imisních situací. Navrhované jsou především omezení pobytu dětí mateřských a základních škol ve venkovním prostoru a umožnění cestování městskou hromadnou dopravou zdarma. Tím se alespoň částečně zamezí negativnímu vlivu na zdraví dětí a mírně se sníží zatížení ovzduší dopravou.

Z tohoto důvodu měl výstupní atribut v_y – průměrná hodnota prachových částic PM_{10} v následujících 24 hodinách – pouze dvě kategorie, z nichž první odpovídá stavu, kdy množství imisí nepředstavuje riziko pro zdraví obyvatelstva, a není tedy důvodem pro vyhlášení regulatorních stavů a druhý množství imisí, kdy by měla být zahájena omezení. Navrhované jsou především omezení pobytu dětí mateřských a základních škol ve venkovním prostoru a umožnění cestování městskou hromadnou dopravou zdarma. Tím se alespoň částečně zamezí negativnímu vlivu na zdraví dětí a mírně se sníží zatížení ovzduší dopravou.

4 KLASIFIKAČNÍ MODELY V NEBANKOVNÍ FINANČNÍ INSTITUCI

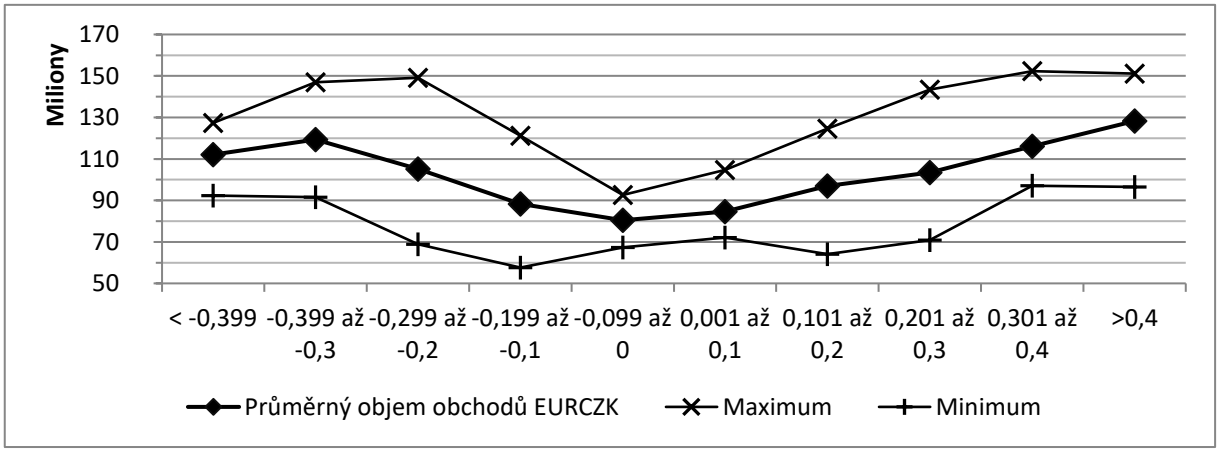
Řízení vztahů se zákazníky (CRM) je přístupem k řízení firemních interakcí se stávajícími či budoucími zákazníky. Tento přístup se snaží analyzovat historii klientů, komunikaci s nimi, popř. veřejně dostupné informace tak, aby firmě umožňoval vytvořit individualizovanou nabídku pro konkrétního klienta. Tím dojde ke zvýšení spokojenosti klienta a k růstu příjmů firmy. V současnosti se soustředí především na stálé klienty dle pravidla, že udržet stávajícího klienta je levnější než akvizice nového klienta [12]. Cenu akvizice zvyšuje především nutnost nastavení marže bez dostatečné historicky podložené znalosti klienta. Na počátku obchodního vztahu se nastavují všem novým klientům nadstandardní podmínky. Po určité době dochází k revizi těchto podmínek a klienti jsou klasifikováni do jednotlivých skupin dle významnosti pro společnost. Velcí klienti generují společnosti významný zisk. Na druhou stranu jsou operace s nimi náročnější na lidské i výrobně – skladovací kapacity. Menší profitabilní zákazníci pak vytváří pozadí, které je významné pro stabilitu firmy. Někteří z nich se postupem času stanou velkoobjemovými zákazníky.

Ze všech těchto důvodů je rychlá klasifikace klienta pro budoucí vývoj obchodních vztahů klíčová. Bohužel ji ovlivňuje velké množství proměnných, které se nedají dopředu predikovat. Proto je nutné klasifikaci neustále podrobovat revizi, zda odpovídá současnému stavu a zda ji pozitivně či negativně neovlivňují faktory, které nejsou do klasifikace zahrnuté. Cílem této části práce je vytvořit teoretický model CBR s využitím metod CI pro platební instituci (dále jen PI), která se pohybuje ve vysoce konkurenčním prostředí. Model respektuje současné nastavení procesů v PI a snaží se o optimalizaci klasifikace v rámci těchto procesů.

Stejně jako v CRM je možné využít klasifikační model i ve veřejné správě. Týká se to především například nově založených firem a klasifikace jejich chování v rámci povinností vůči finanční správě, což je v současnosti velmi diskutované téma. Součástí problému je povinnost finanční správy hájit zájmy státu v oblasti daní, dále to, že ta samá finanční správa nemá dostatek lidí (popř. kvalifikovaných lidí) na kontrolu všech subjektů, složitost daňových zákonů a dalších právních norem a velmi proměnlivé chování firem v dynamicky se měnícím tržním prostředí. Podobný klasifikační model by pak mohl pomoci alokovat omezené lidské zdroje finanční správy na ty případy, které vykazují vzorce chování firem, které zákon porušily.

4.1 Formulace problému klasifikace klienta

Základem rozhodování v PI je běžný reporting. Ten nejenom shrnuje základní ukazatele subjektu, ale je nápomocen i jejich hodnocení. Příkladem může být report denního profitu. To je jeden z klíčo1vých ukaza$_t$elů firmy a je součástí dalších činností, jakými jsou např. plánování nákladů. Podle tohoto indikátoru jsou dále hodnoceni vedoucí pracovníci, popř. pracovníci přímo zodpovědní za prodej. Tento indikátor však ovlivňuje velké množství dalších faktorů. V PI je jedním z takových faktorů pohyb kurzu jednotlivých měn. Po velmi dobrých výsledcích z několika následujících dní po intervenci České národní banky (dále ČNB), která je zavedla v listopadu 2013, se dostavil propad profitu na měnovém páru EURCZK. Proto byla vypracována Ad Hoc analýza, jejíž výsledky shrnuje Obrázek 9. Levá strana ukazuje objem směněných prostředků klientů za jeden den a dolní lišta denní změnu kurzu na měnovém páru euro – česká koruna v korunách [14]. Z této analýzy vyplynulo, že pokud se kurz nepohybuje, profit společnosti klesá o desítky procent.



Obrázek 9: Průměrný objem obchodů za den v milionech CZK (osa *y*) na měnovém páru EURCZK dle denní změny kurzu koruny vůči euru (osa *x*). Zdroj: [36].

Po tomto zkoumání došlo ke klasifikaci klientů dle jejich dlouhodobých charakteristik. Základními atributy, které byly využity při zkoumání PI, jsou objem směněných peněz a_1 a profit z této směny a_2 (dále jen objem a profit). Dále bylo počítáno s atributem počet transakcí a_3 a průměrný profit na transakci a_4. To vše je primárně sledováno z hledisek klientů a času – jednotlivý klient, region, stát, den, týden, měsíc, kvartál a rok. Data byla počítána od 1. 1. 2010. Obecnou charakteristiku klientů lze najít v [39].

Pro dané atributy byly zpracovány základní statistické parametry. Velký rozptyl je především u objemu a profitu. Hodnoty atributů mají logaritmicko-normální (dále log-normální) rozdělení pravděpodobnosti. Následně byly dle [55] vypočítány korelační koeficienty. Ty jsou

zobrazeny v Tabulce 15. Z nich vyplývá, že objem, profit a počet transakcí jsou velmi závislé veličiny. Nahradit tyto parametry pouze jedním atributem je možné při zkoumání jednoho klienta, kde objem a profit spojuje nastavená marže klienta, a tudíž se korelační koeficient blíží k hodnotě 1. U zkoumání větších skupin klientů již není možné nahrazení jedním atributem provést bez ztráty podstatných údajů. [39]

Tabulka 15: Korelační koeficienty základních atributů. Upraveno dle [39].

Atribut	**Atribut**		
	Celkový objem (a_1)	Celkový profit (a_2)	Počet transakcí (a_3)
Celkový objem (a_1)	1		
Celkový profit (a_2)	0,8267	1	
Počet transakcí (a_3)	0,7043	0,7387	1
Průměrný profit na transakci (a_4)	0,1877	0,2162	-0,0269

Prvním krokem k analýze klientů pomocí shlukování byla transformace (normalizace a standardizace) dat. Jak bylo zmíněno výše, data měla před normalizací log-normální rozdělení pravděpodobnosti. Normalizace se provedla pomocí logaritmických funkcí. Následovala standardizace dat. Poté již bylo možné data rozdělit do skupin. Počet těchto skupin nebylo možné odhadnout, proto se pro stanovení počtu využily Kohonenovy mapy (NN, která využívá učení bez učitele – více viz [31]). Pomocí nich se stanovil počet skupin na 10. Jako shlukovací metoda byla využita nehierarchická metoda *k*-means. Ta rozdělila data do deseti skupin. [24] Jejich charakteristika je v Příloze 12. Při bližším pohledu na jednotlivé skupiny došlo k rozdělení klientů do dvou velkých skupin na exportéry (prodejce cizí měny) a importéry (prodejce domácí měny). Rozdíl mezi těmito dvěma skupinami je dán charakterem a možnostmi podnikání. Importéři jsou obvykle obchodní firmy. Nakupují, když jim dochází zboží k prodeji. Směňovat tedy musí i v dobách, kdy je pro ně kurz méně výhodný. Dále mají výrazně vyšší počet transakcí než exportéři. Z těchto důvodů jsou citlivější na poplatky spojené se směnou a zasíláním peněz (tedy typ sazebníku). Snaží se zároveň držet nízkou marži. Jejich vyjednávací pozice však není z důvodů menších objemů, tak silná.

Naproti tomu exportéři prodávají zboží a inkasují cizí měnu. Pokud mají i prodej na domácím trhu, tak využívají tzv. přirozený hedging (zajištění) vůči kurzovým pohybům a to tak, že mzdy a daně platí domácí měně, zatímco bankovní úvěry, platby za energie a mezifiremní faktury platí v cizí měně (obvykle v euru). Díky tomu nesměňují tak často a vzhledem k velikostem firem, je směna relativně malá, a ne častější než jednou až dvakrát za kvartál. Jako příklad lze uvést firmu BRAVO Isolit z Jablonného nad Orlicí. I u firem, které mají více jak 90 % příjmů

ze zahraničí (např. výrobce klavírů Petrof), lze vysledovat, že mění devizy méně často než importéři. Zde to je tak jednou až dvakrát do měsíce. Vzhledem k tomu, že tyto firmy devizy neposílají, ale pouze inkasují, nejsou zatíženy poplatky spojenými s převodem, a tudíž na ně nejsou tak citlivé. Navíc mají prostor pro to, aby si počkaly na výhodnější kurz. Pokud je tedy trh volatilní, uvítají takové firmy sledování kurzu pomocí objednávky a v případě, že trh dosáhne stanovených hodnot, rychle ho potvrzují. Pokud je trh málo volatilní (stav po začátku intervencí), pak vzhledem k objemům je vyjednávací pozice exportérů silná a mohou si vynutit snížení marže (více viz [6], [7], [35]).

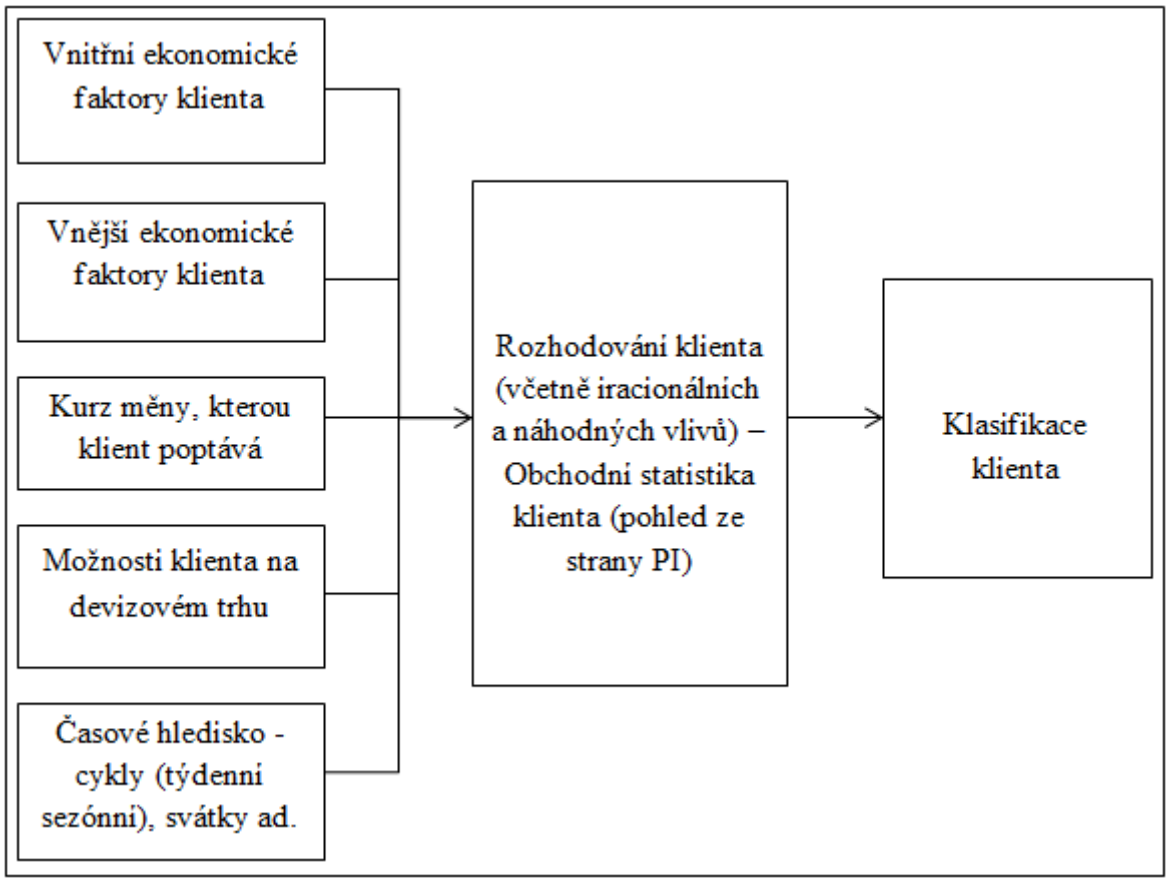
Takovéto rozdělení klientů do shluků podle jejich významu pomohlo PI správně nastavit sazebníky [36]. Dále má velký význam pro marketingové akce nebo nabídky směny či objednávek. Dlouho se také podle průměrného profitu na transakci klasifikovali noví klienti, u kterých chyběla delší historie. I z toho důvodu se prosadila právě shluková analýza proti taktéž testované RFM analýze. Popis obou metod a jejich využití je v [24], [59].

V polovině roku 2014 bylo patrné, že stávající stav zařazování nových klientů do shluků pouze pomocí průměrného profitu na transakci, není udržitelný z důvodu nárůstu chybovosti. Kurz se v té době stabilizoval díky intervenčním opatřením ČNB nad hranicí 27 Kč/Euro. Díky tomu bylo pro klienty jednodušší se v nabízených kurzech orientovat, což zvýšilo tlak na snížení marží i snížení profitu na transakci. Proto bylo nutné odhadnout po třech měsících spolupráce (dále d_2) od podpisu smlouvy (dále jako d_1), zda bude klient dlouhodobě profitabilní nebo neprofitabilní. Standardně se klienti vyhodnocují rok od podpisu smlouvy (dále d_3). Proto je odhad po třech měsících velmi rizikový. Do obchodního vztahu vstupuje velké množství faktorů, které nemohou být jednoduše kvantifikovány. Jedná se například o nabídky konkurenčních firem či bank, výpadky produkce klientů, kurz jdoucí proti potřebám klienta a podobně. Přesto je tento odhad nutný vzhledem k nastavení marže klienta, sazebníku, odměny pro obchodní zástupce za akvizici nových klientů atd.

I přes všechny tyto analýzy se nedařilo zvýšit profit, který byl ovlivněn dlouhodobou stagnací kurzu. Jeho vývoj nelze spolehlivě predikovat, proto se do plánu zahrnul pouze jeho dlouhodobý předpokládaný vývoj. Hodnocení profitu a případné úpravy plánu se však provádí se znalostí této veličiny. Spolu s dalšími faktory jako jsou sezónní a týdenní cyklus, svátky atd. je podstatnou částí reportingu (dle [19], [22], [36]).

Jak bylo uvedeno v úvodu, rozhodnutí o zařazení klasifikaci klienta je nutné udělat dříve, než lze jednoznačně prokázat, zda klient bude nebo nebude profitabilní. Termín takové klasifikace byl v PI stanoven na 3 měsíce od podpisu smlouvy s klientem (d_2). Vzhledem k různému

postavení firmy na trzích v jednotlivých státech, kde působí, a rozdílným podmínkám pro hodnocení profitability, se jednotlivé trhy zkoumají zvlášť, protože rozhodování klientů na jednotlivých trzích ovlivňují jiné, popřípadě jinak důležité vstupní parametry. Ty jsou, stejně jako celý klasifikační proces zobrazen na Obrázku 10.



Obrázek 10: Klasifikační proces v PI. Zdroj: autor

Bázi zkoumaných klientů nyní tvoří data klientů z ČR (2750 klientů). Data obsahují 34 atributů (popis dat lze najít v Příloze 12). Ty se rozdělují do skupin:

- evidenční (3 atributy) – ID klienta, právní subjektivita, trh klienta,
- datové údaje (3 atributy) – Datum podpisu RS (d_1), Datum klasifikace (d_2), Datum vyhodnocení bonity (d_3),
- údaje o bonitě (7 atributů):
    - základní ukazatele (6): objem za prvních 6 a 12 měsíců od podpisu RS, profit za prvních 6 a 12 měsíců od podpisu RS, počet transakcí za prvních 6 a 12 měsíců od podpisu RS
    - agregovaný výstupní atribut (1): Bonitní klient [Ano/Ne]

Vstupní klasifikační údaje (21 atributů):

- objemové ukazatele (7): celkový objem za první 3 měsíce od podpisu smlouvy, objem první až šesté transakce klienta,
- profitové ukazatele (7): celkový profit za první 3 měsíce od podpisu smlouvy, profit první až šesté transakce klienta,
- frekvenční ukazatele (7): počet transakcí za první 3 měsíce od podpisu smlouvy, doba mezi podpisem RS a prvním obchodem ve dnech, doba mezi následujícími šesti transakcemi ve dnech.

Nejdříve byly základní statistické analýzy zaměřeny na největší trh, tedy ČR. Během nich byly zjištěny následující skutečnosti. Čtyři klienti byli z dalších analýz vyjmuti, protože vykazovali výrazně odchýlené hodnoty. Dále byla odebrána skupina klientů, kteří v prvních třech měsících od podpisu smlouvy neudělali jediný obchod. Zkoumání těchto klientů není při daných parametrech modelu možné. Obecně jde říct, že jde valnou většinou (okolo 90 %) o neprofitabilní klienty. Celkově tedy model v rámci českého trhu pracuje s 2489 klienty. Jejich rozdělení dle právní subjektivity (právnické osoby – PO, fyzické osoby podnikající – FOP, fyzické osoby nepodnikající – FO) je v Tabulce 16.

Tabulka 16: Rozdělení CZ klientů do skupin dle právní subjektivity. Zdroj: autor.

Skupina	**Typ klienta**			
	PO	**FOP**	**FO**	**Celkem**
Počet klientů – trénovací a testovací skupina	1292	350	743	2385
Počet klientů vhodných pro klasifikaci – trénovací a testovací skupina	1164	316	668	2148
Počet klientů – validační množina	169	63	133	365
Počet klientů vhodných pro klasifikaci – validační množina	159	60	121	340

Během dalšího testování se prokázalo, že právní subjektivita je velmi významný atributem, který rozhoduje o chování klientů. Některé pokročilé metody jako NNs vykázaly výrazně lepší výsledky v případě, že data byla zkoumána v rámci jednotlivých skupin samostatně. Jelikož se i dle standardních statistických metod zjistil výrazný rozdíl v kvalitě stávající klasifikace, bylo v PI přistoupeno k úpravě procesů. Díky této úpravě bude dále postupováno tak, že se jednotlivé právní subjektivity budou zkoumat odděleně. Základní model bude rozvinut především u právnických osob, které tvoří základ podnikání PI.

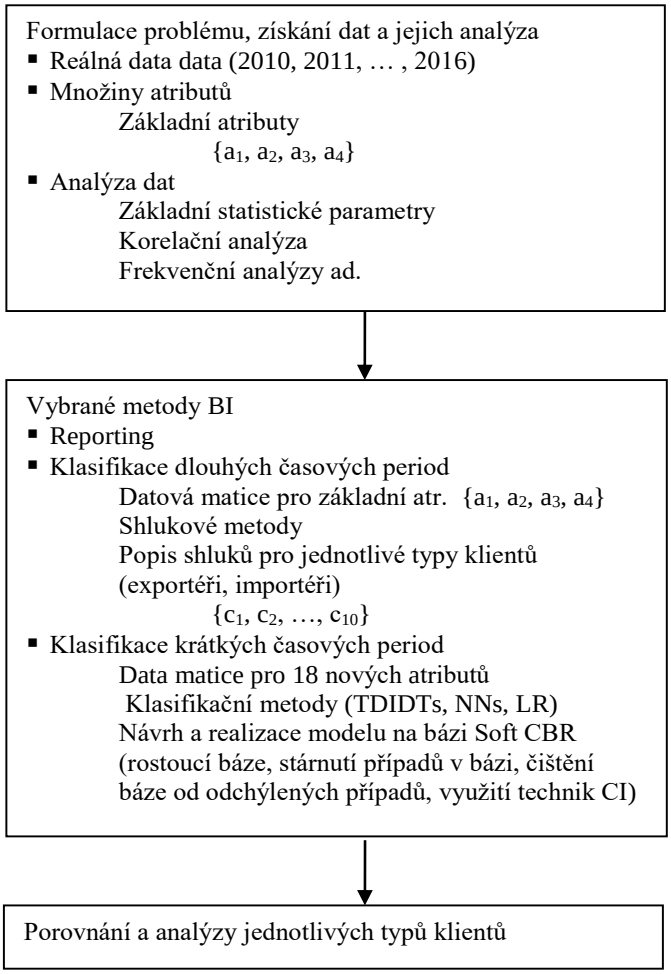
4.2 Data a jejich předzpracování

Pro analýzu klientů PI (profitabilní a neprofitabilní) byla použita shluková analýza. Vstupními daty byly hodnoty atributů a_1, a_2, a_3 and a_4 za roky 2010 až 2013. Analýza "nových" klientů na základě definovaných shluků podle ročních dat není dostatečně dynamická a nezohledňuje krátkodobější období.

Pro klasifikaci nových klientů bylo použito 18 vstupních atributů a 1 výstupní atribut (viz Příloha 11). Uvedené atributy přesněji popisují počátek vztahu mezi klientem a PI. Jde o:

- agregovaný výstupní atribut „bonitní klient" je stanoven na základě 6 ukazatelů: objem za prvních 6 a 12 měsíců od podpisu rámcové smlouvy (dále RS); profit za prvních 6 a 12 měsíců od podpisu RS; počet transakcí za prvních 6 a 12 měsíců od podpisu RS;
- vstupní klasifikační atributy (18 atributů), které jsou definované takto: objem první až šesté transakce klienta; profitové ukazatele (6 atributů); profit první až šesté transakce klienta (6 atributů); frekvenční ukazatele doba mezi podpisem RS a prvním obchodem ve dnech, doba mezi následujícími šesti transakcemi ve dnech (6 atributů). Všechny vstupní atributy mají charakter log-normálního rozdělení pravděpodobnosti.

Pokročilé analýzy se týkají buď zjištění dosud skrytých vazeb, nebo mají určit míru ovlivnění výstupního parametru jednotlivými vstupními parametry. Při klasifikaci se snažíme zařadit prvek na základě nám známých informací do skupiny podobných prvků. Dle zařazení pak je možné usuzovat na jeho pozdější chování [42]. U predikce se snažíme odhadnout výši výstupního parametru, která se nejčastěji řeší pomocí regresních metod. Je možné využít i jiné metody, jako jsou NNs [31, s. 98], TDIDTs, LR aj. V rámci těchto analýz je podstatné najít validní parametry, které skutečně ovlivňují výstup. S velikostí firmy však obvykle roste počet těchto parametrů, a tím i objem zpracovatelných dat. V rámci tohoto velkého množství jsou možné dva přístupy. Prvním je dekompozice velkých celků na menší a následně jejich řešení. To je méně náročné na kapacity a lze jednoduše využít všech možností statistiky či Business Intelligence (BI). Druhým je práce s velkými daty (Big data – velikost terabytů nebo petabytů, více viz [25] a [49]). Výhodou je, že se může přijít na skryté souvislosti, které se při dekompozici dat nedají objevit. Navíc se dá využít standardní metodologie BI – CRISP-DM (Cross-Industry Standard Process for Data Mining). Nevýhodou jsou vysoké nároky na technické i lidské kapacity firem. Proto se mnoho těchto služeb i u velkých firem outsourcuje. Výsledky takovéto analýzy jsou pak zpětně využity ve standardních reportech či dekomponovaném pozorování popsaných např. v [13].



Obrázek 11: Model analýzy prováděné v PI. Zdroj: autor.

V roce 2015 tak např. PI v rámci Ad Hoc analýzy zjistila, že na základě dostupných informací nesprávně klasifikuje v d_2 klasifikuje klienty – fyzické i právnické osoby. V rámci této analýzy bylo testováno 2385 klientů, z toho 1292 PO, 350 FOP a 743 FO. Pomocí běžných statistických metod (porovnání středních hodnot, mediánů a dalších parametrů jednotlivých atributů) byly změněny klasifikační parametry. Díky tomu byla kvalita klasifikace v čase d_2 zvýšena o cca 10 % (viz Tabulka 17). Zároveň byly otestovány další možnosti klasifikace. V Tabulce 17 jsou zobrazeny tři nejvíce úspěšné možnosti: TDIDTs – C5, NN – MLP a LR.

Výše zmíněné algoritmy posunuly kvalitu klasifikace od dalších 10 % na testovací množině zobrazené v Tabulce 17. Díky tomu došlo nejen ke zvýšení profitu. Méně profitabilním klientům byla přidělena vyšší marže či horší sazebník, zatímco profitabilní získali výhodnější marži a další podmínky, čímž byla snížena možnost jejich akvizice konkurencí. Zároveň se snížily náklady vzhledem k propadu lepší klasifikace od odměn obchodních zástupců.

Tabulka 17: Výsledky testovaných metod u statického modelu na validační skupině právnických a fyzických osob. Zdroj: [39].

Klasifikace klienta – testovací skupina – metoda:	Právnická osoba – správně / chybně klasifikováno v [%]	Fyzická osoba – správně / chybně klasifikováno v [%]
Původní systém dle interní analýzy PI	62,71 / 37,29	66,03 / 33,97
Standardní statistické metody	73,54 / 26,46	76,31 / 23,68
Klasifikace TDIDT – C5	81,76 / 18,24	87,15 / 12,85
Klasifikace NN – MLP	83,65 / 16,35	87,25 / 12,75
Klasifikace – LR	79,25 / 20,75	86,85 / 13,15

4.3 Model CBR

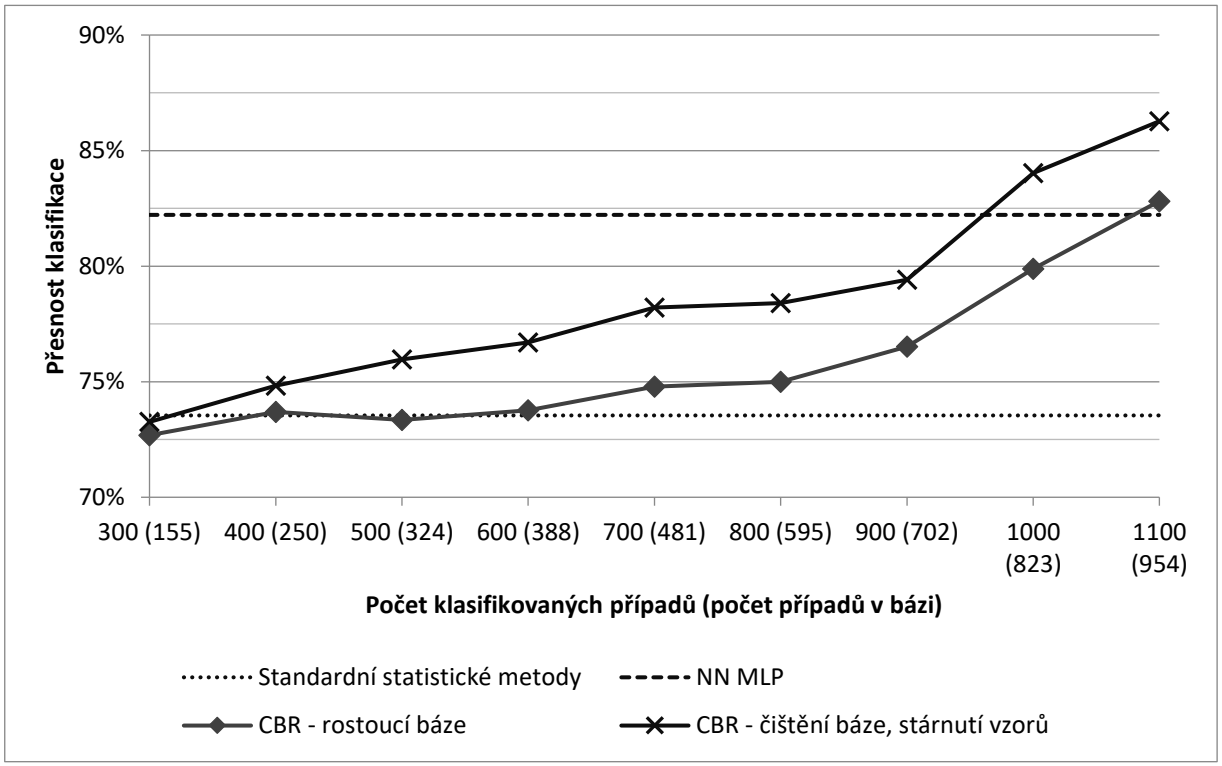
V rámci modelu CBR bude rozvinuto několik postupů, které by měly zvyšovat kvalitu CBR. Standardní model CBR lze nalézt např. v [42] nebo [58]. Jednotlivé základní fáze budou postupně upravovány:

- Fáze získání (Retrieve)
    - Měření vzdálenosti pomocí standardní Euklidovské vzdálenosti případů
    - Vzdálenosti měřené pomocí „hrubé vzdálenosti“ – nastavení bude podrobeno experimentálnímu zkoumání
- Fáze posouzení (Revise)
    - Rozhodne na základě shody vstupních atributů a vybere odpovídající výstupní parametr
    - Pokud bude vzdálenost odpovídat více případům v bázi, tak zvolí převažující výsledek y
    - Při opětovné shodě rozhodne dle vzdálenosti u klíčových atributů
- Fáze zachycení (Retain)
    - Bez omezení – všechny případy vstupují do báze
    - Případy jsou zpětně posuzovány, pokud vedou k chybám, jsou jako chybné z báze vyjmuty pomocí algoritmu, jehož nastavení projde experimentálním testováním
    - Omezení vstupu do báze, pokud v dané vzdálenosti již budou podobné případy se stejným výstupním výsledkem – podle podobnosti bude docházet k větší váze takovýchto případů

Fáze Revise bude aplikována v rámci všech úprav modelu. Fáze získání a zachycení budou postupně doplňovány či obměňovány. Jejich výsledky se pak porovnají a vybere se nejvhodnější model. Ten bude zároveň porovnán s rozhodovacími metodami NNs, TDIDTs a LR. Metody BI obvykle potřebují trénovací množinu případů (o velikosti od 50-70 %) a následně testovací množinu, která výsledky ověří. Jako nižší hranice bylo stanoveno 580 klientů a jako vyšší hranice 800 klientů v testovací množině.

Výsledky těchto metod lze najít v Tabulce 17. Z ní lze vypozorovat, že skupina právnických osob je výrazně obtížnější na klasifikaci, proto se bude model CBR věnovat převážně jí. U fyzických osob jsou výsledky TDIDTs algoritmu C5 natolik kvalitní, že není tak velká potřeba rozvíjet složitější model.

Jednotlivé případy tedy vstupují do báze případů postupně a stávají se aktivními až v případě, že jsou jednoznačně definovány jako profitabilní nebo neprofitabilní. CBR je v tomto ohledu pružnější než jiné rozhodovací nástroje, protože je schopné dávat validní výsledky již při poměrně malé bázi případů. Porovnání s původním systémem klasifikace v PI a klasifikací pomocí CBR je možné porovnat na dle Tabulky 17 a Obrázku 12.



Obrázek 12: Kvalita klasifikace pro CBR s rostoucí bází. Zdroj: [40].

4.3 Standardní CBR, časové hledisko a čištění báze

Jako řešení kvality klasifikace v rámci standardního CBR byly identifikovány dvě možné cesty. První z nich předpokládá, že vzory v bázi stárnou (stárnutí vzorů – SV) a představují skutečnost, která již byla překonána. Díky tomu může dojít ke špatné klasifikaci nových případů. Řešení bylo možné buď pomocí dalšího klasifikačního atributu (rozdíl mezi d_1 případu a d_1 případu v bázi případů, rozuměj vzoru). Tento způsob se projevil jako neefektivní. Kvalita predikce se zvýšila v řádu desetin procenta (maximum bylo 1 % u 1000 případů v bázi), přičemž tento atribut vyžadoval neustálý přepočet dynamického CBR kvůli standardizaci tohoto atributu. Dále bylo možné případ automaticky odebrat, pokud dosáhl nějaké hranice stáří. Tento přístup se však na datech pokrývajících 3 roky neosvědčil, protože snižoval počet případů v bázi a tím i kvalitu klasifikace, která následně nepřekračovala 75 %.

Byla zvolena jiná cesta pomocí násobitele (konstanty) vzdálenosti v závislosti na rozdílu d_1 případu a vzoru. Odhadem byla stanovena výše konstanty (2,6), jíž budou starší případy znevýhodněny. Následně došlo k prohledání prostoru časových vzdáleností pomocí iterační metody HC popsané v [48]. Lokální optimum bylo nalezeno na hodnotě 930 dní. Nyní se tou samou metodou prohledal prostor výše konstanty. Lokální optimum se nacházelo na hodnotě 3,6. Dále se otestovala další – nižší – hranice. Zatímco hranice 930 dní je dle mnoha testování optimální, nově nalezená hranice na hodnotě 548 se jeví pouze jako lokální optimum, přičemž další je na hodnotě 731. Zatím testy ukázaly jako vhodnější variantu s pásmem na hranici 548:

1. pásmo: 548> (d_1 případ – d_1 vzor)
2. pásmo: 548 ≤(d_1 případ - d_1 vzor) < 930
3. pásmo: (d_1 případ - d_1 vzor) ≥930

Pro první pásmo je konstanta vždy 1. Pro druhé pásmo byla nalezena pomocí HC lokální maxima u hodnot 1 a 1,4. U třetího pásma se při opětovném prohledání prostoru s nižší hranicí na 548 hodnota koeficientu nezměnila (3,6). Tímto bylo dosaženo kvalitnějších výsledků na testovací množině (Obrázek 12 – CBR – čištění báze, stárnutí vzorů). Při validaci však došlo k poklesu na 77,99 % při nastavení konstant 1/1/3,6 a na 75,44 % při nastavení 1/1,4/3,6. Výsledek se tedy při rostoucí bázi neustále zpřesňuje, až dosahuje rozdílu 3,14 % ve prospěch využití „časového“ aspektu v CBR. I přes dobré výsledky na validační množině je možné toto využití považovat za experimentální a je třeba dále otestovat jeho robustnost i při řešení jiných úloh [40].

Další možností, jak zkvalitnit bázi případů, bylo posuzovat kvalitu případů a vzorů v čase d_3. Pokud byly shodné, oba vstupují do báze případů. Pokud shodné nejsou pak:

- podobnost případů je příliš malá – oba případy jsou validní a vstupují do báze,
- případ vstupující do báze je nevalidní – jedná se pravděpodobně o odchýlený případ a do báze nevstoupí,
- vzor v bázi je nevalidní – jedná se odchýlený případ, který vstoupil do báze při malém počtu vzorů a na základě podobnosti s velmi vzdáleným případem, a bude odebrán z báze,
- případ i vzor jsou nevalidní – případ do báze nevstoupí a vzor z ní bude odebrán.

Tento přístup čištění báze (dále jako ČB) může být považován i za náhradu časového aspektu, protože v případě, že je nějaký vzor neodpovídající skutečnosti, dojde k jeho odstranění z báze. Výhodou přístupu je, že zcela odstraní odchýlené případy. Nevýhodami je, že některé případy resolveru jako odchýlené pouze připadají a jejich odstranění a zmenšení báze je nežádoucí. Dále pak k jejich odstranění dojde až po vyhodnocení případů v čase d_3. Ta doba může být až rok dlouhá. Během té doby se odchýlené případy v bázi mohou opakovaně stát vzory dalších případů. Zatímco v prvním případě SV by byly díky časovému rozdílu alespoň znevýhodněny.

Důležitý problém je zjištění hranice, od které lze říci, že případy jsou natolik vzdálené, že jejich podobnost nemusí vést ke stejnému výsledku. Čím více bude růst daná hranice, tím větší množství případů bude posuzováno, zda jsou validní či nikoliv. Za touto hranicí se již dá říci, že případy jsou tak vzdálené, že jsou pravděpodobně oba validní.

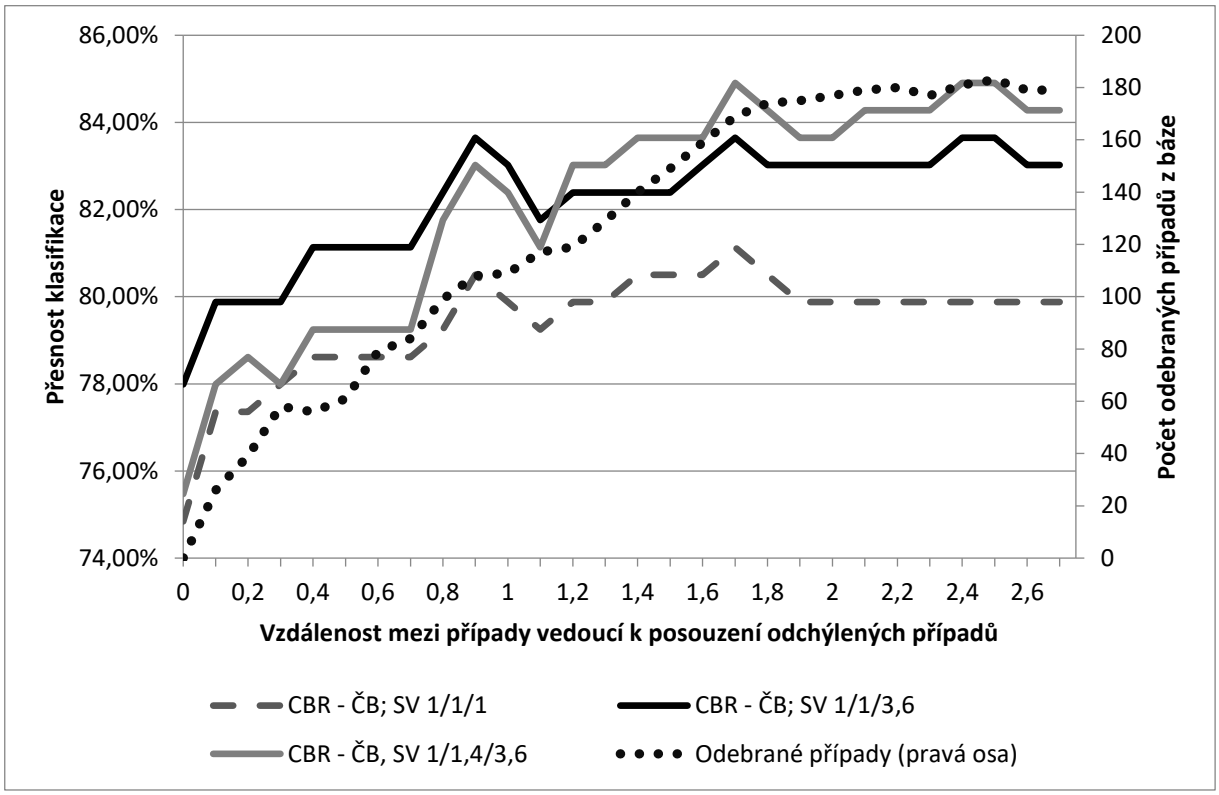
Tabulka 18: Výsledky přesnosti klasifikace na testovací množině v [%] dle jednotlivých přístupů v rámci CBR. Zdroj: [40].

Metoda	**Počet případů v bázi (Počet případů v bázi)**								
	300 (155)	**400 (250)**	**500 (324)**	**600 (388)**	**700 (481)**	**800 (595)**	**900 (702)**	**1000 (823)**	**1100 (954)**
Standardní CBR	72,69	73,69	73,34	73,76	74,78	75,00	76,52	79,88	82,81
Stárnutí případů	72,92	73,95	73,64	74,11	75,22	75,55	77,27	81,10	85,94
Stárnutí případů a čištění báze	73,26	74,83	75,96	76,70	78,21	78,40	79,41	84,02	86,27

Jak ukazuje Obrázek 13 na řadě CBR – ČB, SV 1/1/1 (tedy bez stárnutí vzorů), bylo pomocí metody HC [48] prohledána prostor vzdáleností pro čištění báze. První lokální maximum bylo zaznamenáno při nastavení této hranice na průměrné hodnotě 0,9 (tedy průměrná vzdálenost mezi atributy případu a vzoru je 0,05) a druhé na hodnotě 1,7 (průměrná vzdálenost atributů je 0,0944). Jedná se o normalizované a standardizované hodnoty atributů, jinak by podobné nastavení nebylo možné [40].

Algoritmus čištění báze tedy prověřoval okolí jednotlivých případů. Pokud případy v rámci definované hranice vedly ke stejným výsledkům, jako u zkoumaného případu, tak se případ

i přes jeho nesprávnou klasifikaci bral jako validní. Vzdálenost hranice byla nejdříve experimentálně testována. Celkový počet prověřovaných případů byl 396. Při nastavení vzdálenosti na 0 zůstávaly všechny případy v bázi (aby došlo k přezkoumání případu, musel by být zcela identický). Jak se rozšiřovala hranice, postupně rostl počet vyřazených případů z báze. Přímá úměra zde zcela neplatí, protože stejně jako roste vzdálenost, ve které posuzujeme případ a vzor, tak roste i vzdálenost, pro kterou posuzujeme jejich další okolí. Maximálně tak bylo z báze vyřazeno 183 případů v intervalu hodnot od 2,495 do 2,554.



Obrázek 13: Vliv nastavení vzdálenosti (osa x) pro zkoumání odchýlených případů na úspěšnost klasifikace validační množiny (levá osa y) a počet případů odebraných z báze (pravá osa y). Zdroj: [40].

Kvalita klasifikace přitom stoupala. Bylo však nutné stanovit optimální hranice, popř. zda se změní i nastavení hodnocení stárnutí případu. Opět byla použita iterační metoda HC. Výsledek metody na validační množině je vidět na Obrázku 13. Lokální maximum se vždy nacházelo na hodnotě 1,7, přičemž v případě nastavení koeficientů vzdáleností pro pásmo 1: 1, pro pásmo 2: 1 a pro pásmo 3: 3,6 (1/1/3,6) pak ještě na intervalu od 2,353 do 2,571. Při prozkoumávání prostorů pak bylo znovu jako optimální vyhodnoceno nastavení koeficientů pro pásmo 1: 1, pro pásmo 2: 1,4 a pro pásmo 3: 3,6 (1/1,4/3,6). Výsledky na testovací bázi lze porovnat v Tabulce 18. Výsledek klasifikace validační množiny je 84,91 % (při vyřazení 169 vzorů z báze), což je lepší výsledek, než jaký poskytuje NNs MLP (přesnost 83,65 %).

V případě výše uvedených výsledků se hodnoty, u kterých nebylo možné rozlišit jejich správnost, v bázi mazaly, popř. do ní vůbec nevstoupily. Při testování se rozdíl pohyboval okolo 0,63 % ve prospěch odebrání obou případů z báze, proto bylo dále standardně počítáno s touto variantou. Při zpětném ověřování na validační množině při nastavení stárnutí případů však průměrný rozdíl činil pouhá 0,62 % a průměrný rozdíl v počtu odebraných případů byl 18 případů (z 369 zkoumaných).

4.5 Využití metod výpočtu vzdálenosti ve fázi získání

Pro výpočty vzdáleností v rámci následujících metod byla využita Euklidovská vzdálenost (Euclidean distance, dále ED) se využívá v případech, kdy chceme najít nejbližší případ nebo *k*-nejbližších případů. Základem pro správný výpočet ED je standardizace, popř. i normalizace dat. V našem případě mají data log-normální rozdělení pravděpodobnosti. Tudíž je vhodné je normalizovat pomocí logaritmické funkce a následně standardizovat, protože pokud poměřujete objem a profit, který se pohybuje v promile objemu, tak by objem měl výrazně větší váhu. Vzorec pro výpočet vzdálenosti mezi případem *k* a případem *l* je níže:

$$d_E(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt{\sum_{j=1}^{m} \left(x_{kj} - x_{lj}\right)^2} \qquad 4.1$$

kde x_{kj} je hodnota *j*-tého pozorování na *k*-tém prvku a x_{lj} je hodnota *j*-tého pozorování na *l*-tém prvku [34, str. 238]. U standardního CBR se použije nejbližší případ z báze a dle něj se klasifikuje nový případ. Tento přístup je však velmi citlivý na šum, kdy jeden špatný vzor může rozhodnout o nevalidní klasifikaci mnoha nových případů.

4.5.1 Metody k-nejbližších sousedů

Oproti tomu metoda *k*-nejbližších sousedů (dále *k*-NNbs) vybere *k* nejbližších případů a dle určitých pravidel se nový případ klasifikuje. V našem případě, kde jsou pouze dvě hodnoty klasifikace (profitabilní a neprofitabilní klient), je nejjednodušší možností nový případ klasifikovat podle toho, zda převažují vzory neprofitabilních či profitabilních klientů. Pokud je počet stejný, tak se určí klasifikace podle nejbližšího případu. Možných úprav výpočtu pomocí ED je mnoho. Jednou z nich je rozlišit váhu jednotlivých případů dle vzdálenosti jednotlivých vzorů od současného případu. Omezí se tak význam velmi vzdálených případů pro klasifikaci. Nevýhodou tohoto přístupu dle [9] je jeho náročnost v případě rozsáhlé multidimenzionální báze případů.

Standardní přístup metody *k*-sousedů pro tři a pět nejbližších vzorů se dle výsledku jeví jako robustnější metoda oproti běžnému výpočtu CBR pomocí ED. Při porovnání s CBR – SV a ČB však při růstu báze začíná ztrácet a rozdíl je následně patrný především na validační množině. Přesto jsou tyto dvě metody srovnatelné s například s LR.

Tabulka 19: Vývoj klasifikace pro jednotlivé metody CBR v [%]. Zdroj: Autor

Metody	**Počet klasifikovaných případů (Počet případů v bázi)**								
	300 (155)	**400 (250)**	**500 (324)**	**600 (388)**	**700 (481)**	**800 (595)**	**900 (702)**	**1000 (823)**	**1100 (954)**
CBR – bez ČB a SV	73,3	74,1	73,9	74,0	75,0	74,8	75,2	76,9	76,9
CBR – SV a ČB	75,4	76,4	76,5	77,0	78,1	78,3	80,1	82,3	86,1
CBR – 3-NNbs	75,8	76,9	77,0	77,0	78,1	78,1	79,1	80,7	81,9
CBR – 5-NNbs	76,3	77,5	77,9	78,1	78,9	79,3	80,1	80,1	80,6
CBR – SNh	76,1	76,9	76,8	76,8	76,6	77,1	77,2	76,9	74,5
CBR – ENh1	77,4	78,9	78,7	79,3	79,6	80,6	81,7	83,2	82,9
CBR – ENh2	78,0	79,5	80,0	80,7	81,0	82,4	83,9	85,1	84,3
RdCBR - ČB	77,8	79,0	79,3	80,5	81,4	82,2	84,6	85,7	85,2

4.5.2 Experimentální metody zkoumání okolí případu

Další možností je prohledání okolí případu. Zde je nutné stanovit hranice pro vzdálenost, v rámci které je ještě porovnávání případů v bázi s novým případem validní, a kdy již nikoliv. Vzdálenost je možné stanovit jako statickou veličinu – konstantu (dále jako SNh), což je vhodné ve chvíli, kdy je báze již dostatečně plná a nové případy se již do ní neukládají nebo jen doplňují velký počet případů v bázi. Hrozí totiž dvě negativní situace. První z nich je, že některé případy nebudou mít v dané hranici žádný vzor z báze, a nebude je možné klasifikovat. Druhou negativní situací je opak, kdy se do porovnání dostane i velké množství vzdálených případů. Tomu se dá částečně zbránit stanovením většího počtu hranic a následným váhováním vzorů dle jejich vzdálenosti.

Případně je možné stanovit hranice elasticky v závislosti na charakteristice báze případů a vzdálenosti mezi novým případem a nejbližším vzorem. První možností (dále jako ENh1) je stanovit hranici pro *i*-tý případ následovně $H_i = \min d\left(X_i, X_j\right) \times K$, tedy vzdálenost k nejbližšímu vzoru *j* se vynásobí konstantou stanovenou na základě velikosti a charakteristik báze případů. Výhodou je, že každý případ má svůj vzor. U velmi blízkých případů *i* a *j* se prozkoumává jen relativně blízké okolí nového případu *i*. Tudíž se daří efektivně vyloučit nevalidní vzory. U případů s velmi vzdálenými vzory se naopak prozkoumává velký prostor, přičemž se počítá, že převáží případy vedoucí ke správné klasifikaci. Stejně jako v předchozím případě lze stanovit více hranic a případy dle vzdálenosti váhovat.

Posledním testovaným přístupem, který zohledňuje aktuální stav báze případů (ENh2), je stanovení hranice $H_i = \min d\left(X_i, X_j\right) + (\frac{1}{n} \times (\min d_{i-n} + \cdots + \min d_{i-1})) \times K$, kde se ke vzdálenosti k nejbližšímu vzoru přičte průměrná vzdálenost mezi vzory a případy v posledních *n* posuzovaných případech násobená konstantou *K*. Tento přístup je vhodný především pro dynamické systémy, kde báze roste, a tím se zmenšují průměrné vzdálenosti mezi případy a vzory. Opět je tam možné stanovit větší počet hranic s různou váhou vzorů.

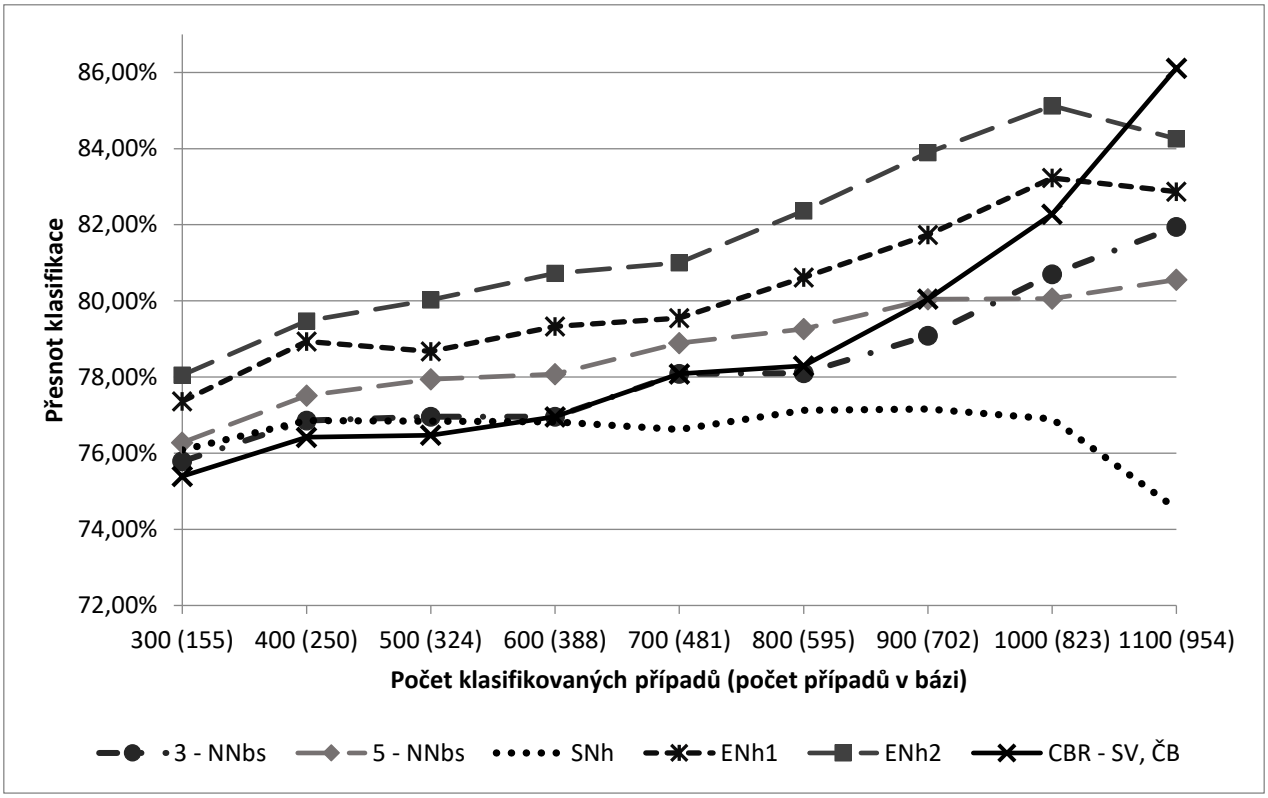
Při využití SNh bylo nutné nastavit hranice, v rámci které se případy porovnávají. Hranice byla nalezena pomocí metody HC a její optimální výše je 1,5. Bylo testováno, zda neexistuje ještě další hranice, kvůli níž by případy měly vyšší váhu. Tato hranice byla nalezena na hodnotě 0,705. Případy do této vzdálenosti dostaly při rozhodování váhu 1, případy mezi touto hranicí a vnější hranicí 1,5 měly váhu 0,5. Ostatní případy v bázi nebyly zahrnuty do rozhodování, resp. jejich váha byla 0. Na testovací množině byla úspěšnost při 700 případech v bázi 77,2 % a na validační množině 72,96 %. Přitom na testovací množině nebylo klasifikováno 26 případů z důvodu, že do vzdálenosti 1,5 se v bázi nenacházel jediný případ, podle kterého by mohl být nový případ klasifikován. Z daného vyplývá, že statická hranice není vhodné především proto, že jak se plní báze a roste počet případů, tak se do klasifikace propadají vzory, které nejsou pro srovnání vhodné – viz Tabulka 19. Průměrný počet vzorů na jeden případ pak je 319,68 mezi tisícím a tisícím stým případem.

Pokud využijeme ENh1, je nutné opět stanovit koeficienty, kterými se bude násobit vzdálenost k nejbližšímu případu. Čím je tato vzdálenost větší, tím větší okolí se prohledává. Opět byly určeny dvě hranice. V rámci první hranice mají vzory váhu 1, mezi první a druhou hranicí pak mají váhu 0,5. Lokální optima byla opět nalezena pomocí HC a jsou jimi koeficienty k_1 = 1,6 a k_2 = 2,6. Díky tomuto nastavení získáme kvalitní výsledky jak na testovací množině (81.7 % při 700 případech v bázi a následně dále roste), tak na validační množině (80.5 %).

Průměrný počet porovnávaných případů roste. Přesto je nárůst menší než u statického okolí. Je to z důvodu, že jak roste počet případů v bázi, klesá vzdálenost mezi případem a vzory, a tudíž se zmenšuje i zkoumané okolí. Tato metoda se tedy jeví jako robustnější než předchozí výpočet SNh. Průměrná vzdálenost mezi vzorem a klasifikovaným případem je 0,2884 (Medián je 0,2019 a maximum 2,146). Celkem 48 případů má vzdálenost větší než 1. Tím, jak se zvyšuje počet případů v bázi, klesá průměr vzdálenosti z hodnot 0,8238 u první stovky případů až po 0,1533 pro posledních sto klasifikovaných případů.

Při použití metody ENh2 se postupovalo podobně jako u ENh1. Hodnoty konstant k_1 a k_2 byly stanoveny pomocí metody HC na 0,5 (váha vzorů 1) a 1,4 (váha vzorů 0,5). Výsledky byly

lepší než u metody ENH1 jak na testovací množině (Obrázek 14), tak na validační množině, kdy přesnost dosáhla 82,39 % (Tabulka 20). Pro porovnání – 1000 až 1100 nový případ mají průměrně 22,21 vzorů oproti cca 320 u metody SNh.



Obrázek 14: Dosahované přesnost klasifikace dle použité metody výpočtu vzorů v dynamickém modelu dle počtu klasifikovaných případů, v závorce počet případů v bázi. Zdroj: autor.

Výsledky zobrazené v Tabulce 20 ukazují, že především metody ENh2 může konkurovat jiným klasifikačním metodám. Ostatní metody považujeme za průměrné, SNh dokonce za velmi podprůměrnou. Je to dáno tím, že čím je plnější báze, tím větší množství případů je porovnáváno a tím se ztrácí přesnost klasifikace.

Tabulka 20: Výsledky jednotlivých metod vyhledání vzorů v bázi případů na validační množině. Zdroj: autor

Klasifikace klienta – metoda:	**Přesnost klasifikace na validační množině – správně / chybně klasifikováno v [%]**
3-NNbs	79,87 / 20,13
5-NNbs	78,62 / 21,38
SNh	72,96 / 27,04
ENh1	80,50 / 19,50
ENh2	82,39 / 17,61
CBR – SV, ČB	84,28 / 15,72

4.5.3 Hrubá vzdálenost

Hrubá vzdálenost je zatím pouze experimentální pojem. Vychází z pojetí neurčitosti – nerozlišitelnosti, jak ho definuje Gotlob Frege, a který následně rozvíjí RST (viz [44]). Hrubá vzdálenost předpokládá, že čísla sama o sobě mohou být v dynamicky se měnícím prostředí (zde kurzové výkyvy) ne zcela přesná. Proto, pokud se od sebe liší do určité vzdálenosti(D_*), tak je bere jako stejná (nerozlišitelná). Naopak od určité vzdálenosti(D^*) můžeme prohlásit, že čísla jsou natolik rozdílná (nestejná), že není třeba o nich jako o podobných uvažovat. Mezi těmito dvěma regiony se nachází prostor, kde sice máme určitou představu o míře podobnosti, ale ta nemusí být přímo úměrná rozdílu velkosti těchto dvou čísel. Proto je hodnotám vzdálenosti, které se nacházejí v tomto regionu, přidělena hodnota 0,5.

Vzdálenost mezi jednotlivými atributy mezi objekty lze definovat následovně:

$$\left|x_{ik} - x_{jk}\right| \leq D_* \; pak \; d_k = 0 \qquad 4.2$$

$$\left|x_{ik} - x_{jk}\right| \geq D^* \; pak \; d_k = 1 \qquad 4.3$$

$$D_* < \left|x_{ik} - x_{jk}\right| < D^* \; pak \; d_k = 0{,}5 \qquad 4.4$$

Celková vzdálenost pak bude:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} d_k \qquad 4.5$$

Zcela zásadní je v tomto případě stanovit dolní (D_*) a horní hranice (D^*). V našem modelu se využívaly standardizované hodnoty. Tudíž se počáteční nastavení stanovilo na D_*=0 a D^*=0,03. Následně se metodou HC experimentálně otestovaly další hodnoty. Na testovací množině byly určeny optimální hodnoty na D_*=0,28 a D^*=0,82. Pro porovnání je možné nalézt jak na Obrázku 14, tak v Tabulce 21.

Tabulka 21: Porovnání výsledků testovací množiny pro standardní CBR, CBR s čištěním báze a stárnutím případů a CBR využívající hrubou vzdálenost v [%]. Zdroj: autor

Metoda	**Počet případů v bázi (Počet případů v bázi)**								
	300 (155)	**400 (250)**	**500 (324)**	**600 (388)**	**700 (481)**	**800 (595)**	**900 (702)**	**1000 (823)**	**1100 (954)**
Standardní CBR	72,69	73,69	73,34	73,76	74,78	75,00	76,52	79,88	82,81
CBR – ČB a SV	75,39	76,42	76,47	76,96	78,08	78,29	80,05	82,28	86,11
Hrubá vzdálenost	77,81	78,98	79,34	80,50	81,38	82,22	84,63	85,76	85,20

Obecně se CBR s využitím hrubé vzdálenosti (dále RdCBR) jeví jako robustní, protože pro jakýkoliv počet případů v bázi je výsledná hodnota klasifikace lepší než u předcházejících testovaných typů CBR, přičemž na hodnotě 800 případů v bázi je již srovnatelná s NN MLP i vzhledem k tomu, že NN MLP potřebuje podobný počet případů jako trénovací množinu.

Následovalo ověření na validační množině, kde NN MLP dosáhlo ještě lepšího výsledku a to 83,65 %. CBR se stárnutím případů a čištění báze pak hodnoty 84,28 %. RdCBR vykazuje na validační množině přesnosti klasifikace 83,65 % (oboje viz Tabulka 26, s. 77[2]). To je výrazně více než standardní CBR s rostoucí bází (74,84 %), ale méně než dosáhlo NN MLP a CBR s čištěním báze a stárnutím případů. Z toho vyplývá poznatek, že ačkoliv RdCBR často porovnává více vzorů vůči současnému případu, tak není zcela imunní vůči odchýleným hodnotám.

Proto byl vytvořen model, kde se využilo CBR s čištěním báze pro identifikaci odchýlených hodnot, které pak byly odstraněny z báze, kterou využívá RdCBR. Zatímco na testovací množině se hodnoty zvýšily maximálně o procento, tak na validační množině byla výsledná hodnota 84,91 %.

Přes to, že na zpracovávaných datech metoda RdCBR vypadá jako velmi robustní a schopná přinést kvalitní výsledky, bude nutné její obecnou využitelnost potvrdit i na dalších datech. V rámci modelu klasifikace klienta se nabízí možnost přenést tuto klasifikaci i na jiný trh (Polsko). Stále se však bude jednat o podobný typ dat.

Jako příklad lze uvést následující případy. Případ i vzor jsou oba klienti, kteří inkasují pátý den každého měsíce z Velké Británie 1000 GBP, které obvykle ihned smění na CZK a používají pro svou potřebu. Takto se jeví jako zcela totožné případy. Pokud však vzor má d_1 *z 1. října 2015 (klasifikační atributy tak jsou z října až prosince 2015), pak bude mít až na poslední měsíc pravidelnou měsíční periodu s celkovým objemem 112.200,- CZK. Zatímco případ bude spadat do období červen až srpen 2017 (*d_1 *= 1. 6. 2017), tak objem bude 88.500,- CZK, což je jen na objemu v CZK[3] rozdíl 21 %. Stejně tak do směny mohou zasáhnout svátky 5. a 6. července a rozdíl na měsíční bázi může být i 16 % v čase (přičemž jedno posunutí ovlivní pauzu mezi dvěma transakcemi – jednou směrem k delšímu a podruhé ke kratšímu prostoji. Takže i zcela obdobné případy mohou na datech vykazovat značně odlišné hodnoty. Právě z tohoto důvodu je nutné ověření i na datech z jiných oborů.*

Zatím se jako optimální možnost jeví souběh dvou klasifikačních metod. Jednou z nich je modifikované CBR, které zároveň čistí bázi od nevalidních případů. Jako druhá metoda se nabízí buď výše zmíněné CBR s hrubou vzdáleností, CBR s ENh2, nebo NN MLP. Oba tyto postupy jsou v rámci disertační práce rozvinuty a popsány. Výhodou CBR s hrubou vzdáleností

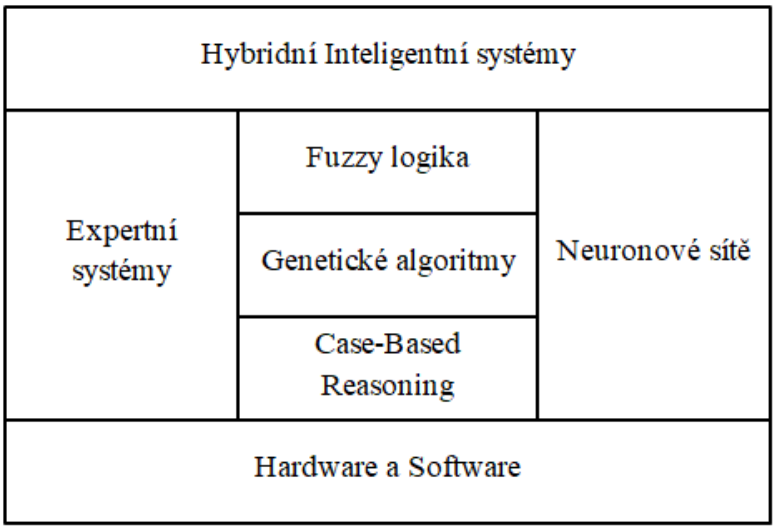
[2] Rozdíl mezi výsledky NN MLP uvedenými zde a v Tabulce 26 (s. 77) je způsobený tím, že v tabulce jsou uvedené výsledky dosažené pomocí naprogramovaného modulu ve VBA v rámci dynamického modelu, zatímco zde jsou výsledky získané pomocí IBM SPPS Modeler se standardním rozložením dat na testovací, trénovací a validační množinu.

[3] V CZK jsou počítány jak objemy, tak profit, protože je to PI referenční měna, ke které vztahuje jak veškeré hospodářské ukazatele, tak svou pozici na trhu a rizikové váhy.

by mělo být malý počet případů v bázi nutných pro kvalitní klasifikaci. Pro NN MLP naopak hovoří robustnost metody a to, že pomocí vedle probíhajícího CBR eliminují největší nevýhody NNs – těžkou ověřitelnost výsledků a zachovávání nevalidních případů v bázi.

4.6 Hybridní model dynamického CBR

Při zkoumání klasifikace byl zjištěn rozdíl mezi kvalitou klasifikace jednotlivých metod vzhledem ke konečné hodnotě zařazení. Z tohoto důvodu bylo rozhodnuto o vytvoření hybridního modelu, který by propojoval jednotlivé metody a dle pravidel by případ finálně klasifikoval. Tento model vychází z konceptu, který ve své knize Hybrid Intelligence Systems představil Larry Medsker. Jednotlivými částmi položenými mezi takový systém a hardwarové a softwarové vybavení jsou expertní systémy, FST, Genetické algoritmy, CBR a NNs [33 s. 1]. Tento koncept je dále rozvíjen v řadě odvětví. Pro finanční sféru se těmito systémy zabývá S. Goonatilake v [18] a M. Sonar (například v [50]).

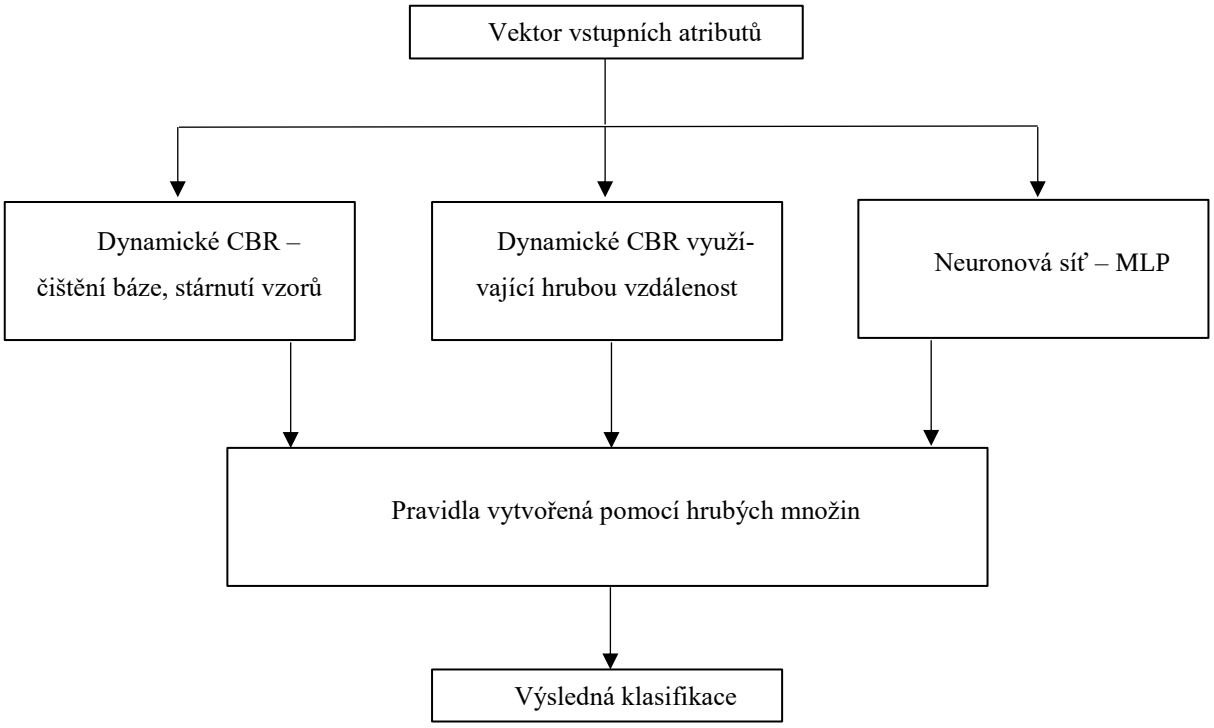


Obrázek 15: Inteligentní technologie využívané v hybridních inteligentních systémech. Zdroj: [33, s. 1].

V rámci tohoto přístupu bylo nutné vyřešit tři dosud neřešené problémy:

- počet a typ metod, se kterými bude tento model vytvořen,
- jak vytvořit pravidla, podle kterých by se tento model řídil,
- vzhledem k tomu, že jako velmi rozdílná byla od počátku identifikována NNs – MLP, bylo nutné vyřešit, jakým způsobem bude tato síť vytvořena a zpravována v rámci dynamického modelu.

Při řešení první z těchto otázek byly brány v potaz výsledky, které byly dosaženy dle předešlých dynamických modelů, a to nejenom z pohledu kvality klasifikace. Proto byla zvolena metoda CBR s čištěním báze a se stárnutím případů. Podává kvalitní výsledky a zároveň upravuje bázi pro další metody, které se mohou v rámci modelu využít. Dále bylo vhodné vybrat metodu, která minimalizuje určitý typ chyby. Proto byla vybrána NNs – MLP, protože má na rozdíl od ostatních metod velmi dobré výsledky v hodnocení případů, které neplní podmínku profitability (dle CM). Dále se zvažovalo několik dalších metod. Nakonec byly vybrány dvě, které mají vysokou přesnost klasifikace již při malém počtu případů v bázi a to jsou RdCBR a CBR – Enh2. Tyto dvě metody mají velmi podobné výsledky, proto bylo zvoleno řešení, že se otestuje model pokaždé se třemi metodami a tyto dvě výše zmíněné metody budou alternovat. Následně se zvolí ten model, který vykazuje vyšší přesnost. Na Obrázku 16 je vidět tento model s využitím RdCBR.



Obrázek 16: Model dynamické klasifikace pomocí více klasifikačních metod. Zdroj: autor.

Vytvoření pravidel, podle kterých se následně bude klasifikovat, je možné několika způsoby. Primární otázkou je, jak rychle chceme pravidla vytvořit a dále jak moc složitý mechanismus využijeme, resp. kolik pravidel vytvoříme. Nejjednodušší metodou je zvolit konečnou hodnotu na základě převažující shody klasifikace jednotlivých metod. Další možností je vytvořit pravidla na základě expertního odhadu, který je možné si utvořit na aktuálními CM. Poslední možností je zapojit do rozhodování nejenom výsledky jednotlivých klasifikací, ale i další atributy (např. míru podobnosti případu se vzorem), a k rozhodnutí bude využito dalších vědeckých

metod – TDIDT nebo RST. V rámci výzkumu budou prověřeny jak přístup pomocí expertního odhadu, tak i dle RST.

Posledním popsaným problémem byla nutnost připravit NNs takovým způsobem, aby byla využitelná na denní bázi. V rámci toho bylo rozhodnuto o tom, že se NN natrénuje na aktuální bázi vždy jednou za 14 dní. Na tomto základě byla vytvořena NNs – MLP v kódu VBA. Jako bylo následně otestováno, tato neuronová síť sice nevykazuje tak kvalitní výsledky, jako je tomu u klasické NN a výpočtech na trénovací, testovací a validační množině, ale zůstala ji hlavní schopnost velmi dobře klasifikovat případy následně vyhodnocené jako neprofitabilní.

4.6.1 Dynamický model využívající jednoduchý model metaklasifikace

Jednoduchý model metaklasifikace využívá jednoduché pravidlo, že v případě, že alespoň dvě ze tří metod klasifikují případ stejně, odpovídá tomuto výsledku i výsledná klasifikace (princip majority). Takovýto model je velmi jednoduchý na obsluhu, nejsou potřeba žádné další podpůrné nástroje. Porovnání dvou modelů – jednoho s RdCBR a druhého s ENh2 (oba dva modely doplněny NNs – MLP a CBR s čištěním báze a stárnutím případů).

RdCBR je mírně robustnější (viz Tabulka 22), což se potvrzuje na validační množině, kde je přesnost u modelu s RdCBR 84,28 %, zatímco u modelu s ENh2 83,65 %. Podobnost výsledků je způsobená tím, že dvě metody jsou totožné a poslední může ovlivnit pouze 13,15 % případů, které nejsou metodami NN MLP a CBR – ČB, SV shodně klasifikované.

Zároveň je však třeba upozornit na to, že v rámci dynamického modelu je přesnost přes 82 % velmi vysoká a dá se poměřovat s výsledky nejpřesnějších metod na statickém modelu se standardní trénovací, testovací a validační množinou. Například rozdíl u NN MLP mezi statickým (83,65 %) a dynamickým modelem (79,87 %) na validační množině činí 3,78%.

Tabulka 22: Porovnání přesnosti dvou modelů jednoduchého hybridního systému v [%]. Zdroj: autor.

Jednoduchý hybridní model tvořený CBR – ČB, SV; NN – MPL a třetí metodou (řízeno principem majority)	**Počet klasifikovaných případů (Počet případů v bázi)**								
	300 (155)	**400 (250)**	**500 (324)**	**600 (388)**	**700 (481)**	**800 (595)**	**900 (702)**	**1000 (823)**	**1100 (954)**
Třetí metoda – RdCBR	78,4	79,6	79,7	80,0	81,0	81,4	82,9	84,5	85,7
Třetí metoda – ENh2	78,1	79,2	78,9	79,3	79,9	81,0	82,0	82,9	84,7
Třetí metoda – ENh1	78,0	79,0	78,8	79,2	79,7	80,8	81,7	82,6	84,3

4.6.2 Dynamický model využívající expertní odhad v modelu metaklasifikace

Expertní odhad je založen na předpokladu, že expert zná data a je schopen posoudit výsledky. Již během klasifikace prvních případů je zřejmé, že metody mají jinou úspěšnost u jednotlivých výsledků klasifikace. NNs MLP má vysokou úspěšnost při klasifikování klienta jako neprofitabilního. Tedy minimalizuje chybu, že takto označený klient se nakonec ukáže jako profitabilní. Stejně tak, že metody RdCBR a ENh2 mají naopak výbornou klasifikaci u profitabilních klientů, takže minimalizují chybu, že by se takto označený klient následně ukázal jako neprofitabilní. Proto byla stanovena následující pravidla v této posloupnosti:

1. Když NNs MLP klasifikují případ jako neprofitabilní => neprofitabilní
2. Když RdCBR / ENh2 klasifikují případ jako profitabilní => profitabilní
3. V případě, kdy NNs MLP klasifikují jako profitabilní a RdCBR / ENh2 jako neprofitabilní, rozhoduje výsledek klasifikace pomocí CBR se stárnutím případů a čištěním báze

Tabulka 23: Matice záměn pro 250 až 349 případ. Zdroj: autor.

Skutečnost	**Klasifikace NNs**		**Klasifikace RdCBR**	
	Neprofitabilní	**Profitabilní**	**Neprofitabilní**	**Profitabilní**
Neprofitabilní	14	33	3	44
Profitabilní	10	43	3	50

Tato pravidla vychází z CM, kde je na skupině 100 případů (250 až 349 případ) možné vidět, že NNs dokáží dříve správně určit alespoň část neprofitabilních klientů. Výsledky jsou zobrazeny v Tabulce 23. Hybridní model s využitím RdCBR má zpočátku podobnou úspěšnost jako hybridní model s využitím ENh2, při více hodnotách však překonává tento model o jedno a více % (postupné výsledky přesnosti klasifikace v dynamickém modelu jsou v Tabulce 24). Celkově však výsledky nejsou optimální, což se ukazuje na validační množině. Hybridní model s využitím RdCBR má hodnotu pouze 81,76 % a s využitím ENh2 81,13 %. Samotné metody mají přesnost RdCBR 83,56 % a ENh2 82,39 % a metody založené na jednoduchém hybridním modelu mají 84,28 % (s RdCBR), 83,65 % (s ENh2).

Tabulka 24: Dynamický hybridní systém s expertem nastavenými pravidly klasifikace v [%]. Zdroj: autor.

Experimentální hybridní model CBR – ČB a SV, NN – MLP a třetí metodou (řízeno expertním odhadem)	**Počet klasifikovaných případů (Počet případů v bázi)**								
	300 (155)	**400 (250)**	**500 (324)**	**600 (388)**	**700 (481)**	**800 (595)**	**900 (702)**	**1000 (823)**	**1100 (954)**
Třetí metoda – RdCBR	80,1	80,6	80,8	81,0	81,2	81,0	81,5	82,6	83,8
Třetí metoda – ENh2	80,0	80,5	80,3	80,2	80,2	80,2	80,3	81,0	82,9

4.6.3 Dynamický model využívající RST v modelu metaklasifikace

Z předchozích experimentů vyplývá, že expertní odhad bez hlubší analýzy dat není optimálním řešením. Nabízí se možnost využít nějakou metodu, která je schopná z výstupů jednotlivých metod vytvořit pravidla a následně je aplikovat tak, aby se zvýšila přesnost klasifikace. Mezi takové metody patří TDIDT a RST. Vzhledem k tomu, že je nutné v této oblasti předpokládat vyšší míru nejistoty, byly vybrány RST a algoritmus LEM 2 pro vytvoření pravidel. Takto vytvořená pravidla by se dále spravovala v bázi pravidel. Pro experiment byl vybrán postup, kdy se pravidla neaktualizují, ale samozřejmě možné vytvořit i zde dynamickou strukturu, která se bude učit dle nárůstu počtu případů v bázi, nebo dle aktuálních trendů v datech.

Výsledky této metody pro využití CBR s ČB a SV, NNs a RdCBR / ENh2 vidíme v Tabulce 25. Pravidla byla vytvořena na datech prvních 324 zařazených případů – tedy od 500 případu je lze využít ke standardní klasifikaci nových případů (výsledky pro předchozí případy jsou dle zpětné klasifikace pomocí těchto pravidel).

Vzhledem k nižšímu počtu případů a nastavení samotného algoritmu LEM2 bylo vygenerováno pouze 5 pravidel. Přesto došlo ke zpřesnění výsledků, což je vidět v Tabulce 24 a na Obrázku 18.

Tabulka 25: Dynamický hybridní systém s pravidly klasifikace získanými pomocí RST. Zdroj: autor.

Experimentální hybridní model CBR – ČB a SV, NN – MLP a třetí metodou (řízeno pravidly vytvořenými RST)	**Počet klasifikovaných případů (Počet případů v bázi)**								
	300 (155)	**400 (250)**	**500 (324)**	**600 (388)**	**700 (481)**	**800 (595)**	**900 (702)**	**1000 (823)**	**1100 (954)**
Třetí metoda – RdCBR	81,6	82,5	81,8	82,2	82,5	84,0	85,4	86,4	87,1
Třetí metoda – ENh2	81,1	82,0	81,5	82,3	82,5	83,8	85,3	84,8	86,7

4.6.4 Vytvoření aplikačního dynamického modelu využívající RST

Vytvoření modelu, který by byl v praxi využitelný, závisí především na dvou věcech. První z nich je cena vytvoření takovéhoto modelu. Ta může být vyjádřena buď penězi za nákup požadovaného programového vybavení, nebo množstvím programátorské práce (tzv. člověkohodin). V tomto případě byly využity již dříve vytvořené části CBR (báze případů, čištění báze, samotná klasifikace). K tomu byly naprogramované části:

Část standardizace a normalizace případů. Ta vycházela z předpokladů, že maximální délka $d_1 – d_2$ je 91 dní a minimální 0. U objemu a profitu pak z toho, že objem / profit nad určitou hranici vedou automaticky ke klasifikování klienta jako profitabilního, tudíž tato hranice byla

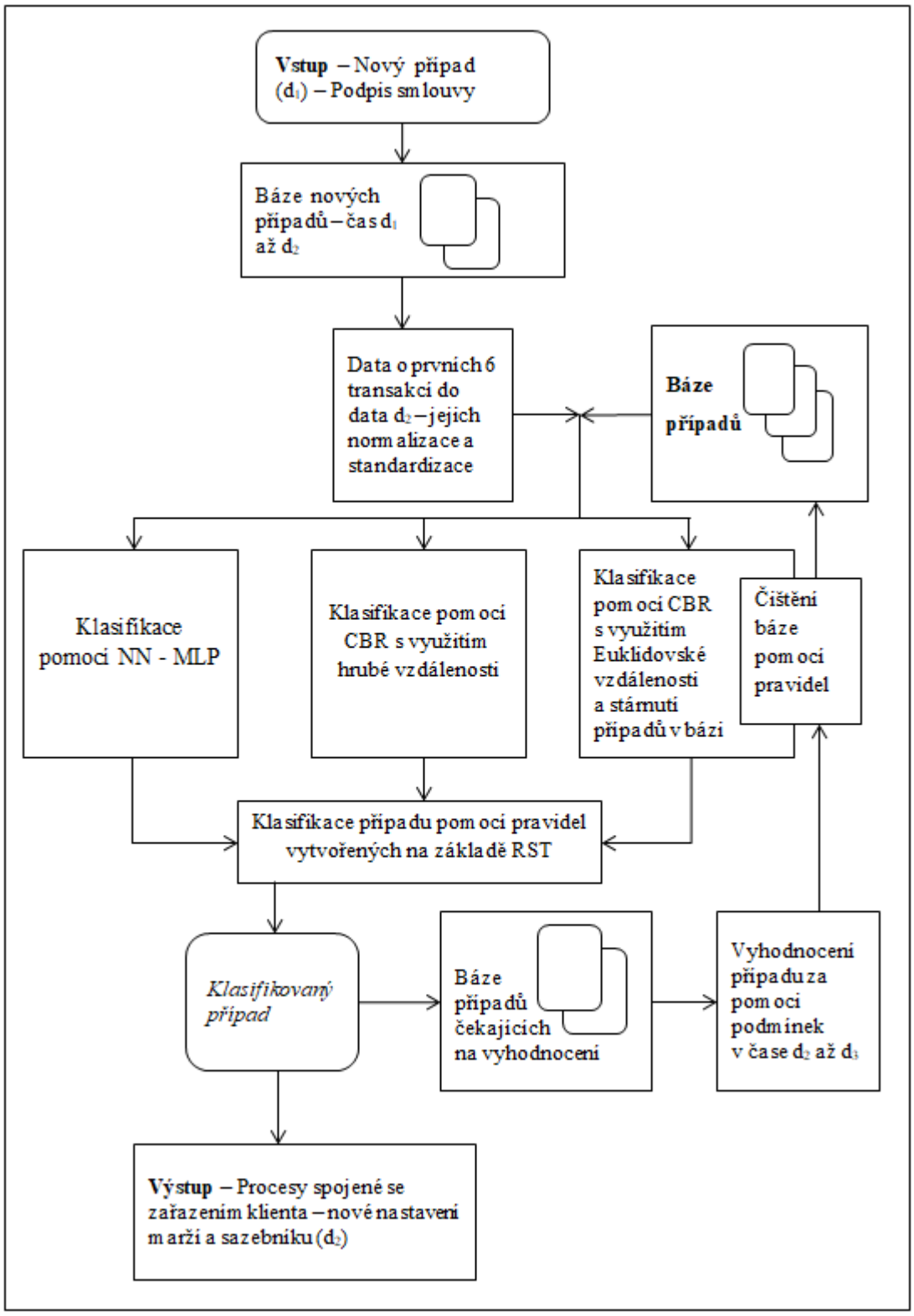
označena jako maximální. Minimální hranice se pak limitně blíží 0. Tím bylo dosaženo toho, že se nemusí při vstupu nového případu s novými maximálními hodnotami přepočítávat zpětně celá báze.

1. Část klasifikace pomocí NNs MLP. Využity byly části kódu uvedené v [5]. Rozdíly mezi výpočetními přístupy vedly i k rozdílům mezi výsledkem metody počítaným pomocí IBM SPSS Modeler (který není volně dostupný) a pomocí kódu ve VBA. Kvalita klasifikace se snížila průměrně o tři procenta. Základní výhoda, tedy kvalitní klasifikace neprofitabilních klientů, však NNs zůstala, a tudíž bylo možné tuto metodu dále využívat. Navíc je v tomto případě vidět velká síla NNs, kdy již s velmi malou trénovací množinou je schopná dosahovat přesnosti okolo 80 %.
2. Metaklasifikace jednotlivých případů na základě vstupů z jednotlivých klasifikačních metod a dle pravidel vytvořených pomocí RST. Samotný vznik pravidel nebyl do programu zabudován z důvodu, že se jejich výpočet neopakuje. Jejich výpočet tak proběhl v programu RSES pomocí algoritmu LEM2.

Z již vytvořených modelů byly použity části finálního vyhodnocení případu dle pravidel v čase d_2 až d_3, čištění báze pomocí pravidel vyvinutých v rámci modelu klasického CBR a taktéž práce s buffery, ve kterých jsou případy čekající na posun do další fáze.

Obecně lze říci, že takto využité VBA v rámci aplikace Excel je velmi levným řešením, co se týče softwaru. Kromě programu RSES nebylo nutné využít dalších nástrojů. Jedinou nevýhodou, která se v rámci zkoumané množiny klientů neprojevila, je čas zpracování, který s růstem báze případů narůstá. V případě využití pro velký počet případů je pak možné realizovat buď přepis kódu do výkonnějšího jazyka (nabízí se C++), nebo omezení počtu případů v bázi podobně, jako to bylo ukázáno na modelu kvality ovzduší.

V tomto případě by však byly nutné stanovit pravidla redukce báze s tím, že u hraničních případů by k redukci téměř nedocházelo, zatímco u jasných případů vedoucích k jedné nebo druhé hodnotě vyhodnocení by redukce byla velká. Dalo by se přitom využít modulu čištění báze pomocí pravidel, do kterého by byly doprogramovány další části, jejichž základem by byla vzdálenost mezi jednotlivými případy. Na pozadí by se pak stále držela báze všech případů a jednou za předem stanovený čas by došlo k přepočítání případů v aktivní bázi (vzorů).



Obrázek 17: aplikační model klasifikace klienta. Zdroj: autor.

4.7 Shrnutí dosažených výsledků pro klasifikační modely klienta nebankovní finanční instituce

Při porovnání modelů klasifikace klientů založených na metodě CBR s model jiných klasifikačních metod (TDIDT, NNs, LR) se ukázalo, že modely založené na CBR jsou vhodným klasifikačním nástrojem. Oproti jiným metodám má velkou výhodu v jednoduchosti tvorby dynamického modelu. Zároveň umožňuje jednoduchou správu báze případů, a tudíž se umí vyrovnat s odchýlenými případy, které nebylo možné odstranit ještě před vstupem do klasifikačního modelu.

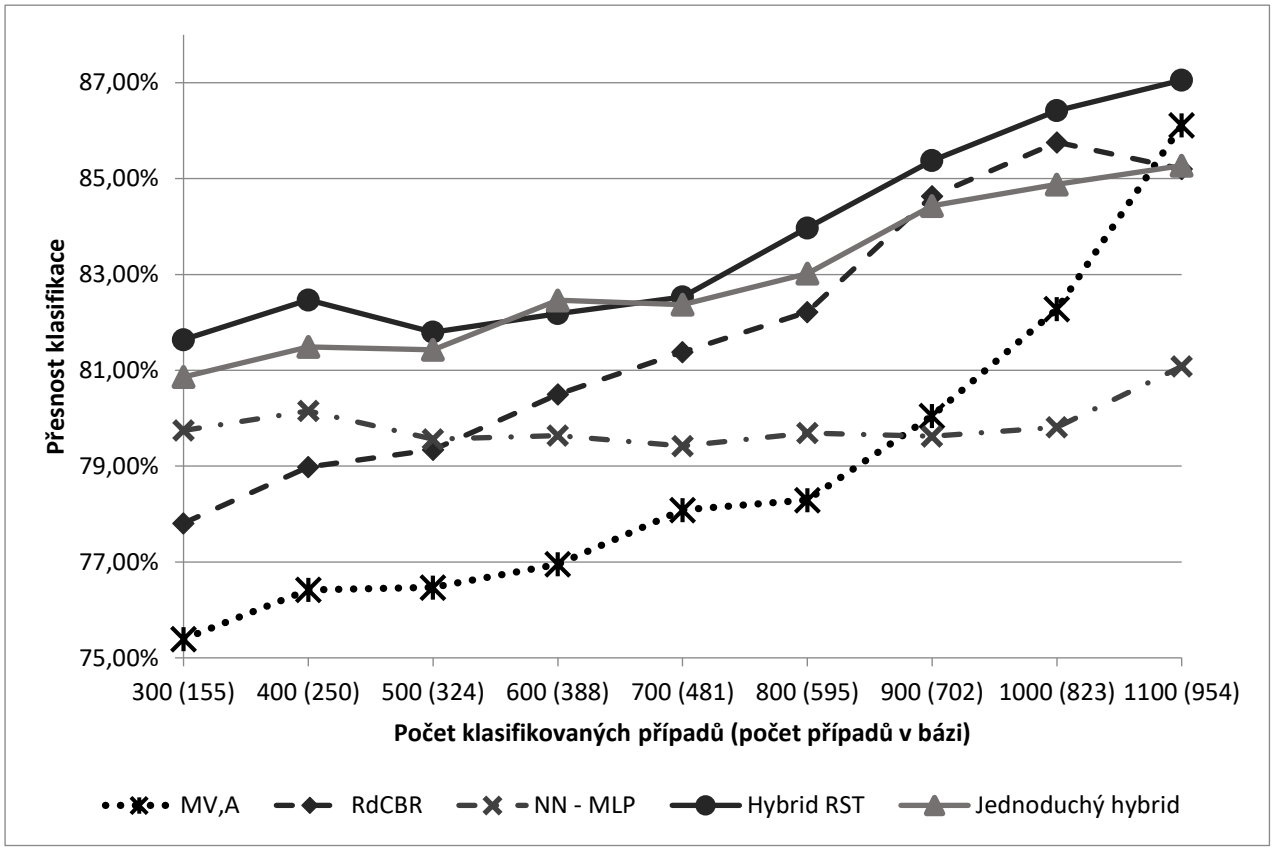
Z hlediska přesnosti a obecné schopnosti pracovat s neurčitostí tuto metodu zřejmě předčí pouze NNs. I ty však fungují lépe, pokud zároveň pracují s výstupy z CBR. K propojení těchto dvou metod došlo v rámci hybridního modelu představeného v předchozích kapitolách. Je třeba podotknout, že kvalita klasifikace vyšší jak 86 % je velmi kvalitním výsledkem. Původní klasifikace v PI se pohybovala okolo 62 až 63 %. Výsledkem je tak zlepšení o více jak 23 %, což je obrovský kvalitativní posun oproti původnímu stavu.

Tabulka 26: Rozdíl mezi kvalitou klasifikace posledních 165 případů z testovací množiny a 159 případů validační množiny v [%]. Zdroj: autor.

Množiny	**CBR – ČB, SV**	**3 – NNbs**	**5 – NNbs**	**SNh**	**ENh1**	**ENh2**	**Rd-CBR**	**NN – MLP**	**Hybrid. model – RST pravidla**
Testovací množina	86,11	81,94	80,56	74,54	82,87	84,26	85,20	81,08	87,05
Validační množina	84,28	79,87	78,62	72,96	80,50	82,39	83,65	79,87	86,16
Rozdíl	-1,83	-2,07	-1,94	-1,58	-2,37	-1,87	-1,55	-1,21	-0,89

Ověření kvality na validační množině je možné považovat za dostatečné. Je to z toho důvodu, že na validační množině se velmi silně projevila trendová složka dat, která snížila přesnost u všech metod (rozdíl mezi metodami založenými na CBR je možné pozorovat v Tabulce 25). Trendová složka se projevuje v datech po celou dobu, kdy roste báze případů. Novější případy reflektují změny v chování klientů. Bohužel k jejich vyhodnocení dochází postupně, a tudíž je není možné ihned využít jako vzory. Proti trendové složce, která snižuje kvalitu klasifikace, naopak působí nárůst počtu případů v bázi, a tudíž se zvyšuje i podobnost případů s jejich vzory a roste tím i kvalita klasifikace. Na obrázku 18 je vidět, že trendová složka převažuje nad pozitivním nárůstem báze případů ve dvou případech, a to mezi čtyřstým a pětistým případem a následně právě na validační množině (viz Obrázek 18).

Právě velikost rozdílu mezi posledními klasifikovanými případy na testovací množině a případy na validační množině může sloužit jako ukazatel robustnosti jednotlivých metod, jak ukazuje Tabulka 26. Zcela jednoznačně je tam prokázána jak kvalita, tak robustnost právě hybridního systému založeného na pravidlech generovaných dle RST.



Obrázek 18: Vývoj jednotlivých metod CBR, NNs a inteligentních hybridních systémů na testovací množině. Zdroj: autor.

5 NAPLNĚNÍ CÍLŮ DISERTAČNÍ PRÁCE

Základním cílem předložené disertační práce bylo vytvoření modelů s využitím metody CBR. Uvedený cíl byl naplněn vytvořením dvou klasifikačních modelů. První byl klasifikační model kvality ovzduší založený na bázi Huntova modelu CBR.

Jako metoda omezující počet vstupních atributů bylo zvoleno RST. Z celkové množiny atributů tak bylo vybráno 6 nejdůležitějších, se kterými bylo dále počítáno jak v metodě CBR, tak i v RFA (kap. 3.4 s. 37–39). Výsledek CBR byl podroben zkoumání. Následovala analýza vah atributů provedená pomocí dvou experimentů. První z nich určoval váhu pomocí zastoupení atributů v pravidlech generovaných pomocí RST. Druhým experimentem byla citlivostní analýza, přičemž ta byla opět provedena pro všechny tři výše zmíněné typy výpočtu vzdáleností. Výsledek ukázal, že první postup můžeme označit za nepřesný. Jeho využití je možné pro rychlé vyhodnocení, běžnou citlivostní analýzu však nahradit nedokáže (kap. 3.5.2 s. 43-44).

Současně byly ověřeny různé možnosti struktury báze případů, kdy jednotlivé sobě podobné případy (podobnost určena pomocí shlukování metodou k-průměrů) byly reprezentovány jedním případem. Byly zkoumány dva přístupy. Prvním z nich bylo vybrání případu, který byl nejblíže středu daného shluku, druhým přístupem bylo vytvoření umělého případu, přičemž hodnota jeho atributů byla rovna střední hodnotě atributů obsažených v daném shluku. Mírně vyšší přesnosti bylo využití prvního přístupu, přičemž tento fakt nebyl dále analyzován (kap. 3.5.2 s. 43).

Fáze získání nejbližšího případu byla testována v obou modelech. Na modelu kvality ovzduší byly využity různé metody výpočtu vzdálenosti mezi případy (Euklidovská, Manhattan a Čebyševova vzdálenost), (kap. 3.5.2 s. 44). U modelu klasifikace klienta byly zkoumány jednotlivé přístupy k vybrání nejbližšího případu včetně několika experimentálních. Jako první byl vyvinut výpočet pomocí standardního výběru nejbližšího případu. Následovaly úpravy tohoto přístupu. První bylo čištění báze případů od odchýlených hodnot pomocí prozkoumání okolí případu a následné řešení pomocí nastavených pravidel. Druhým bylo zachycení časové rozdílnosti mezi případy a tím eliminování trendové složky. Nastavení parametrů proběhlo metodou HC. Řízení čištění báze je pak řešeno pomocí jednoduchých pravidel, které porovnávají další případy v rámci stanoveného okolí (kap. 4.5 s. 58–61). Dále byly zkoumány metody tří a pěti nejbližších sousedů, zkoumání statického a dynamického okolí případu a experimentální hrubá vzdálenost. Všechny metody byly porovnány a byly analyzovány z pohledu jejich rychlosti učení se a celkové robustnosti. Dynamický model byl připraven pro všechny výše zmíněné metody. Zásadním se ukázalo především propojení s čištěním báze, které bylo vyvinuto v rámci

metody nejbližšího souseda. U všech těchto metod došlo ke zlepšení klasifikace, přičemž nejmenší vliv mělo na metody 3 a 5 nejbližších sousedů. Výsledkově nejzajímavější metody určení vzdáleností byla upravená metoda nejbližšího souseda, metoda 5 nejbližších sousedů, metoda dynamického prozkoumávání okolí (ENh2) a hrubá vzdálenost (kap. 4.6 s. 61-67) Vzhledem k integraci těchto metod kvůli čištění báze bylo možné následně jednoduše připravit hybridní model s využitím NNs.

U dynamického modelu CBR, který vyhodnocoval profitabilitu klienta, byly nejdříve řešeny standardizace a normalizace v dynamicky rostoucí bázi. U atributů obratu a profitu byly určeny maximální hodnoty, které automaticky vedou ke klasifikaci klienta jako profitabilního. U atributů týkajících se frekvence obchodování byla maximální hodnota nejvyšší možná hodnota mezi podpisem smlouvy a dobou tří měsíců, tedy 92 dní. Tudíž se stala i nejvyšší možnou hodnotou. Zahrnutí časového hlediska bylo řešeno pomocí hranic a násobitelů vzdálenosti mezi případy, nikoliv pomocí atributů (kap. 4.7.4 s. 72–73).

Hybridní model byl sestaven dle teoretického konceptu popsaného v [33]. První částí řešení bylo přeprogramovat NNs – MLP do jazyka VBA. Využito bylo vzoru, který lze najít na internetu. Pouze došlo k úpravě, aby NNs odpovídala dříve využité NNs v programu SPSS Modeler. I přes tuto úpravu došlo k mírnému snížení přesnosti, kterou vzhledem k tomu, že kód SPSS Modeler není dostupný, není možné jednoduše odstranit. Konečným úkolem bylo relevantně sloučit výsledky všech rozhodovacích metod. Bylo využito několik metod, přičemž se osvědčilo využití RST. Pomocí něj byla vytvořena pravidla. Tento přístup se ukázal jako optimální jak vzhledem k přesnosti, tak robustnosti výsledků.

Posledním krokem obou modelů bylo porovnání CBR s dalšími metodami – NNs, TDIDTs, RST, LR a RFA. Nejlepších výsledků dosahovaly kromě CBR ještě NNs – MLP a RFA u modelu klasifikace kvality ovzduší. U modelu klasifikace klienta bylo dosaženo nejlepších výsledků pomocí hybridního modelu kombinujícího CBR, NNs a RST.

6 ZÁVĚR

V této práci byly představeny klasifikační modely založené na případovém usuzování a dalších metodách výpočetní inteligence, přičemž byly ukázány možnosti propojování jednotlivých metod tak, aby výsledné modely vykazovaly vysokou míru přesnosti i robustnosti. Toto propojení je možné v podstatě v jakékoliv fázi zpracování dat či modelování. Od základního přehledu o datech, kde kromě statistických metod můžou pomoci například neuronové sítě bez učitele (Kohonenovy mapy), přes určení vstupních atributů (pomocí RST), jednotlivé fáze případového usuzování, až po integraci výsledků více klasifikačních metod.

Takto kombinované metody se dají nazvat hybridním inteligentním systémem, jak je chápe Medsker (v [33]). Důvodem, proč zkoumat takovéto hybridní inteligentní systémy, je nutnost řešit velmi složité problémy pomocí umělé inteligence. Ze znalosti silných slabých stránek jednotlivých metod jsme schopni sestavit takový hybridní systém, který obejde omezení jednotlivých metod, a naopak využije jejich výhod tak, že výsledný systém bude přesnější, robustnější, rychlejší nebo méně náročný na zdroje než pokud by byla použita jen jedna metoda. Původním cílem takovýchto systémů bylo simulovat inteligenci podobnou lidské, přičemž význam byl kladen na množství a různost zpracovávaných dat.

V současnosti se množství zpracovávaných dat neustále rozšiřuje. Jejich zpracování v rámci technologií (Big data) se neustále zrychluje, a tak již není primárním cílem inteligentních hybridních systémů zpracování co největšího množství dat, ale schopnost nakládat naopak s malými množinami dat nebo s daty, která jsou zatížena velkými trendovými odchylkami. Popř. mají nahrazovat systémy, které by sice přinesly větší přesnost, ale za cenu neefektivního zvýšení nákladů. Toto dokládá druhý model, kdy z původní přesnosti okolo 63 % bylo dosaženo přesnosti 86 %. Další mírné navýšení je možné očekávat s tím, jak dále poroste báze dat. Očekávaná přesnost velkých dat dle studií obdobných dat se pohybuje od 93 do 97 % [46]. Cena hardwarových i softwarových prostředků je však násobně vyšší. Navíc nelze očekávat, že by tato technologie přinesla tak jednoduché řešení, aby jej bylo možné implementovat například do odměňovacího systému a nedalo se jednoduše obejít.

Oba dva v práci představené modely mohou být využity v praxi (u druhého z nich tomu taky skutečně je). Kromě samotného splnění cílů práce překvapil u modelu ovzduší především velmi jednoznačný rozdíl v typu znečištění mezi Ostravou (primárně lokální topení) a Pardubicemi (doprava). Ten je vidět na čase špiček znečištění. Odhad výhledu na následujících 24 hodin však není lokalitou ovlivněn. Je to z důvodu, že vstupní atributy se vždy vážou k dané lokalitě

(a tedy reflektují automaticky specifika lokality) a podobné povětrnostní podmínky vedou k podobným imisním situacím.

Od 1. ledna 2017 taktéž platí nová a přísnější legislativa [64], počítá s dvanácti hodinovými průměry. Je to z důvodu zrychlení vyhlašování stavů regulace. Stále se však jedná o velkou časovou prodlevu, která se může negativně projevit na zdraví nejzranitelnějších skupin obyvatel, tedy seniorů a dětí. Navíc, jak je vidět z grafů, dvanáctihodinový průměr může být více zavádějící než dvaceti čtyřhodinový. Především pak na Ostravsku, kde je v rámci dne jen jedna špička. Přijetí této právní normy však také znamená, že si zákonodárci uvědomují existenci problému se zpožděním informování veřejnosti a vyhlášení stavu regulace, který se v představeném modelu řeší.

Model klasifikace klientů se již úspěšně využívá v praxi jako systém pro podporu rozhodování, které je však stále svěřeno expertovi (v tomto případě obchodníkovi s devizami), který na základě shromážděných informací vyhodnocuje profitabilitu klienta a přiděluje mu sazebník a případně upravuje obchodní marži. Eliminace experta není zatím možná, protože existují klienti, kteří se vyznačují specifickými potřebami či kladnými externalitami, jako je například vlastnictví většího počtu firem. Méně zajímavé firmy tak dostávají stejné podmínky, jako ty skutečně zajímavé. Všechny tyto informace musí expert posoudit a následně klienta finálně kategorizovat.

Podobný systém lze vyvinout v každém oboru, který pracuje s frekvencemi akcí uživatelů a jejich výší, které nejsou cíleně stanoveny. V privátní sféře je to tedy spíše záležitost velkoobchodu (maloobchod využívá spíše metody vedoucí ke křížovému prodeji – cross-selling). Ve veřejné sféře je to jakékoliv využívání veřejných statků, které není zcela pravidelné a kde je nutné rozhodnout o efektivní nabídce daného statku uživatelům.

Přínosy této práce v teoretické rovině lze označit:

- Porovnání velkého množství přístupů k měření vzdálenosti (podobnosti) mezi případy, a to včetně experimentálních, které dosud nebyly nikde popsány. Ty mohou být klíčové zejména v oblasti s velkou neurčitostí v datech. Přesto je nutné některé přístupy ještě dále ověřit na datech z jiných oborů.
- Propojení jednotlivých metod do jediného celku tak, že vzniknul robustní, přesný a kompaktní model schopný řešit úlohy v dynamicky se měnícím prostředí.

Jako praktické přínosy pro obor informatiky ve veřejné správě lze určit:

- Oba dva modely se dají využít ve veřejné správě. Model klasifikace kvality ovzduší lze využít k omezení dopadů smogové situace na děti a seniory. Klasifikační model klientů lze využít pro daňové či jiné účely popsané výše.

- Oba dva modely založené na CBR byly vytvořeny běžnými nástroji. Lze je velmi jednoduše upravit a následně využít, přičemž jejich používání bude levné a zároveň efektivní.

Rozvoj podobných hybridních inteligentních modelů lze i nadále očekávat vzhledem k tomu, že ačkoliv počet dat, a tedy i využití technologií Big dat, bude nadále růst, tak zároveň bude přibývat dat, která nebudou dostupná z důvodu ochrany osobních údajů, nebo jiných zákonných úprav. Právě v těchto případech je využití hybridních inteligentních systémů s využitím CBR jednou z možností, jak zpracovat kvalitní klasifikační či predikční model.

7 LITERATURA

[1] AAMODT A. – PLAZA E. Case-based reasoning: foundational issues, methodological variations, and system approaches. In: *Artificial Intelligence Communications*, IOS Press, 1994, 7 (1), s. 39–59. Dostupné také z: https://ibug.doc.ic.ac.uk/media/uploads/documents/courses/CBR-AamodtPlaza.pdf

[2] ACORN, T. – WALDEN, S. SMART: support management cultivated reasoning technology Compaq customer service. In: *Proceedings of the Fourth Innovative Applications of Artificial Intelligence Conference on Artificial Intelligence* (IAAI-92), San Jose, CA, AAAI Press, Menlo Park, CA, 1992, s. 3-18.

[3] ALLEN, B. P. Case-based reasoning: business applications. In: *Communications of the ACM 37*, 1994, 37 (3), s. 40-42.

[4] ALTHOFF, K. D. et al. A Review of Industrial Case-Based Reasoning Tools. In: *AI Perspectives Report*, AI Intelligence. 1995.

[5] *Artificial Neural Network with Backpropagation Training in VBA* [online] bquanttrading ©2015 [cit. 30. 10. 2017]. Dostupné z: https://quantmacro.wordpress.com/2015/08/13/artificial-neural-network-with-backpropagation-training-in-vba/

[6] BAILLIE, R. T. – BOLLERSLEV, T. Intra-day and inter-market volatility in foreign exchange rates. In: *The Review of Economic Studies*, 1991, 58 (3), s. 565-585.

[7] BAILLIE, R. T – BOLLERSLEV, T. The message in daily exchange rates: a conditional-variance tale. In: *Journal of Business &Economic Statistics*, 2002, 20 (1), s. 60-68.

[8] BELLANDER, T. et al. Using geographic information systems to assess individual historical exposure to air pollution from traffic and house heating in Stockholm. In: *Environmental Health Perspectives*, 2001, 109 (6), s. 363–369.

[9] BONZANO, A. – CUNNINGHAM, P. – SMYTH B. Using introspective learning to improve retrieval in CBR: a case study in air traffic control. In: *Proceedings of the Second International Conference on Case-based Reasoning* (ICCBR-97), Providence, RI, Springer-Verlag, Berlin, 1997, s. 291-302. ISBN: 978-3-540-69238-6.

[10] BRAUER, M. et al. Air pollution from traffic and the development of respiratory infections and asthmatic and allergic symptoms in children. In: *American journal of respiratory and critical care medicine*, 2002(166), s. 1092–1098.

[11] BRUNINGHAUS, S. – ASHLEY, K. D. Using machine learning for assigning indices to textual cases. In: *Proceedings of the Second International Conference on Case-based Reasoning* (ICCBR-97), Springer-Verlag, Berlin, Heidelberg, 1997, s. 303-314.

[12] BURNETT, K. *Klíčoví zákazníci a péče o ně: koncepce, metody a postupy, jak utvářet a řídit vztahy s klíčovými zákazníky.* Praha: Computer Press, 2002. Business books (Computer Press). ISBN: 80-7226-655-1.

[13] CALEGARI, S. – CIUCCI, D. Granular computing applied to ontologies. In: *International Journal of Approximate Reasoning*, 2010, 51 (4), s. 391-409.

[14] ČESKÁ NÁRODNÍ BANKA. *Kurzy devizového trhu* [online], dostupné z: http://www.cnb.cz/cs/financni_trhy/devizovy_trh/kurzy_devizoveho_trhu/index.html

[15] ČESKÝ HYDROMETEOROLOGICKÝ ÚSTAV. *Portál CHMI* [online], dostupné z: http://portal.chmi.cz/

[16] ČESKÝ ROZHLAS. *Pardubicko trápí smog.* [online]. 14. 11. 2011 [cit. 2018-03-24]. Dostupné z: https://pardubice.rozhlas.cz/pardubicko-trapi-smog-6067989

[17] FINNIE, G. – SUN, Z. R5 model for case-based reasoning. In: *Knowledge-based systems*. 2003, (16), s. 59-65.

[18] GOONATILAKE, S. Intelligent Systems for Finance and Business: An Overview. In: *Goonatilake, S., Treleaven, P. C. (Eds).Intelligent Systems for Finance and Business*, 1995, John Wiley and Sons, s. 1-28, ISBN: 0-471-94404-1.

[19] GORMLEY, F. M. – MEADE, N. The utility of cash flow forecasts in the management of corporate cash balances. In: *European Journal of Operational Research*. Elsevier, 2007 (182), s. 923-935. ISSN: 0377-2217.

[20] GRZYMALA-BUSSE, J. W. – WANG, A. Z. Modified algorithms LEM1 and LEM2 for Rule Induction from Data with Missing Attribute Values. In*: 5th Int. Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences (JCIS'97)*, Research Triangle Park, NC, 1997, s. 69-72.

[21] HERSH, M. *Mathematical Modelling for Sustainable Development*. Springer, Heidelberg, 2006. ISBN: 978-3-642-06341-1.

[22] HINDERER, K. – WALDMANN, K. H. Cash management in randomly varying environment. In: *European Journal of Operational Research*. Elsevier, 2001, (130), s. 468-485. ISSN: 0377-2217.

[23] HORÁK, J. et al. Bilance emisí znečišťujících látek z malých zdrojů znečišťování se zaměřením na spalování tuhých paliv. In: *Chemické listy*. 2011, (105), s. 851-855.

[24] HOSSEINI, S. M. S. – MALEKI, A. – GHOLAMIAN, M. R. Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. In: *Expert Systems with Applications*, 2010, 37 (7), s. 5259-5264.

[25] JEMAL, D. – FAIZ, R. What if mixing technologies for big data mining and queries optimization. In: *Computational Collective Intelligence*. Springer International Publishing, 2015, s. 619-627.

[26] JIRAVA, P. – KŘUPKA, J. – KAŠPAROVÁ, M. Application of rough sets theory in air quality assessment. In: *International Conference on Rough Sets and Knowledge Technology*. Springer, Berlin, Heidelberg, 2010, s. 371-378.

[27] JIRAVA, P. – KŘUPKA, J. – KAŠPAROVÁ, M. System Modelling based on Rough and Rough-Fuzzy Approach. In: *WSEAS Transactions on Information Science and Applications*, 2003, 10 (5), s. 1438 -1447.

[28] KAŠPAROVÁ, M. – KŘUPKA J. Air Quality Modelling by Decision Trees in the Czech Republic Locality. In: *8th WSEAS Int. Conf. on Applied Informatics and Communications (AIC'08)*, WSEAS Press, Greece, 2008, s. 196-201.

[29] KAŠPAROVÁ, M. – KŘUPKA, J. – JIRAVA, P. Approaches to Air Quality Assessment in Locality of the Pardubice Region. In: *5th Int. Conf. Environmental Accounting Sustainable Development Indicators* (EMAN2009), Praha, 2009, s. 1-12.

[30] KOLODNER, J., L. *Case-based reasoning.* Morgan Kaufmann, San Francisco, 1993. ISBN: 1-55860-237-2

[31] KVASNIČKA, V. et al. *Úvod do teórie neurónových sietí*. Bratislava, 1997. ISBN: 80-88778-30-1. Dostupné také z: https://encyklopediapoznania.sk/data/eknihy/informatika/uvod_do_teorie_neuronovych_sieti.pdf

[32] LEAKE, D. E. *Case-Based Reasoning: Experiences, Lessons and Future Directions.* MIT Press, Cambridge, MA, USA, 1996, ISBN: 026262110X.

[33] MEDSKER, L. B: *Hybrid Intelligence Systems.* 1995 Springer Science+Business Media New York. ISBN: 978-1-4613-5998-2.

[34] MELOUN. M. – MILITKÝ, J. – HILL, M. *Kompendium statistického zpracování dat*. Karolinum, Praha, 2012. ISBN: 978-80-246-2196-8.

[35] MELVIN, M. – YIN, X. Public information arrival, exchange rate volatility and quote frequency. In: *The Economic Journal*, 2000, 110 (465), s. 644-661.

[36] MEZERA, F. – KŘUPKA, J. Cash flow management model of payment institution on the basis of system approach. In: *Proceedings of the International Conference Hradec Economic Days 2015*, Hradec Králové, 2015, 5, s. 16-22. ISBN: 978-80-7435-550-9.

[37] MEZERA, F. – KŘUPKA, J. Environmental Modelling based on Rough-Fuzzy Approach. In: *Man-Machine Interactions*, Berlin, Springer, 2014, s. 407-414. ISBN: 978-3-319-02308-3.

[38] MEZERA, F. – KŘUPKA, J. Local model of the air quality on the basis of rough sets theory. In: *Soft Computing Models in Industrial and Environmental Applications.* Springer Berlin Heidelberg, 2013, s. 277-286.

[39] MEZERA, F. – KŘUPKA, J. Selected Business Intelligence Methods for Decision-making Support in a Finance Institution. In: *Scientific Papers of the University of Pardubice*, 2/2017, 24 (40), s. 154 - 164. ISSN: 1211-555X.

[40] MEZERA, F. - KŘUPKA, J. Classification of Clients on the basis of Modifying Case-based Reasoning Algorithms. In: *Man-Machine Interactions 5*, Berlin: Springer, 2017, s. 311-319. ISBN: 978-3-319-67792-7.

[41] NERI, M. et al. Children's exposure to environmental pollutants and biomarkers of genetic damage: II. Results of a comprehensive literature search and meta-analysis. In: *Mutation Research/Reviews in Mutation Research.* 2006, 612 (1), s. 14-39.

[42] PAL, S. K. – SHIU, S. C. K. *Foundation of Soft Case-Based Reasoning*. Hoboken: New Jersey, 2004. ISBN: 0-471-08635-5.

[43] PARDUBICKÝ KRAJ. *Město je ohroženo smogem* [online]. Pardubický kraj ©2018 [cit. 30-06-2017], dostupné na: http://www.pardubice.eu/urad/radnice/pro-media/tiskove-zpravy/mesto-je-ohrozeno-smogem-byl-vyhlasen-signal-upozorneni/

[44] PAWLAK, Z. *Rough Sets –Theoretical Aspects of Reasoning about Data*. Boston, London, Dordrecht: Kluwer, 1991.

[45] PAWLAK, Z. Rough set approach to knowledge-based decision support. In: *European Journal of Operational Research 99*, 1997, s. 48-57.

[46] PROFINIT. *Data science identifikuje zaměstnanecké I rodinné vztahy, a řekne vám, za kolik máte půjčovat.* [online] Profinit ©2018 [cit. 12. 3. 2018]. Dostupné online: https://profinit.eu/blog/data-science-identifikuje-zamestnanecke-i-rodinne-vztahy-a-rekne-vam-za-kolik-mate-pujcovat/

[47] QIN, H. – LUO, D. New Uncertainty Measure of Rough Fuzzy Sets and Entropy Weight Method for Fuzzy-Target Decision-Making Tables. In: *Journal of Applied Mathematics*, 2014, s. 1-7.

[48] RUSSELL, S. J. – NORVIG, P. *Artificial Intelligence: A Modern Approach* (2nd ed.). Upper Saddle River, New Jersey: Prentice Hall, 2003, s. 111–114, ISBN 0-13-790395-2

[49] RUSSOM, P. et al. *Big data analytics. TDWI best practices report*. Fourth quarter, 2011, 19 (4), s.1-34.

[50] SONAR, R. M. Business Intelligence through Hybrid Intelligent System Approach: Application to Retail Banking. In: *CONTRIBUTIONS – Volume I - A collection of papers on banking, finance & technology*. Banknet India Publications, 2006.

[51] SIMOUDIS, E. Using case-base retrieval for customer technical support. *In: IEEE Expert*, 1992, 7 (5), s. 7-12.

[52] SKOWRON, A. – BAZAN, J. – SZCZUKA, M.S. – WROBLEWSKI, J. *Rough Set Exploration System* (version 2.2.2) [online], 2009, [cit. 29-09-2017], dostupné z: http://logic.mimuw.edu.pl/~rses/

[53] SMYTH, B. – KEANE, M. T. Adaptation guided retrieval: questioning the similarity assumption in reasoning. In: *Artificial Intelligence*, 1998, (102), s. 249-293.

[54] SUN, Z. – HAN, J. – DONG, D. Five Perspectives on Case-based reasoning. In: *Advanced Intelligent Computing Theories and Applications with Aspects of Artificial Intelligence.* Berlin / Heidelberg, 2008, s. 410-419. ISBN: 978-3-540-85983-3.

[55] ŠOLTÉS, E. *Regresná a korelačná analýza s aplikáciami*. Jura Edition, 2008. ISBN: 978-80-8078-163-7.

[56] TOPINKA, J., et al. Influence of GSTM1 and NAT2 genotypes on placental DNA adducts in an environmentally exposed population. In: *Environmental and molecular mutagenesis*, 1997, 30 (2), s. 184-195.

[57] VOSS, A. Towards a Methodology for Case-based Adaptation. In: *Proceedings of 12th European Conference on Artificial Intelligence (ECAI´96)*, Wiley, Chichester, 1996, s. 147-151.

[58] WATSON, I. *Applying Case-Based Reasoning: Techniques for Enterprise Systems*. 1st edition. San Francisco: Morgan Kaufmann Publisher, Inc. 1997. ISBN 1-55860-462-6.

[59] WU, H. – CHANG, E. – LO, CH. Applying RFM model and K-means method in customer value analysis of an outfitter. In: *Global Perspective for Competitive Enterprise*, Economy and Ecology, 2009, s. 665-672.

[60] – ZADEH, L. A. Outline of a new approach to the analysis of complex systems and decision processes. In: *IEEE Transactions on systems, Man, and Cybernetics*, 1973, (1), s. 28-44.

[61] ZADEH, L. A. The concept of a linguistic variable and its application to approximate reasoning. In: *I. Information sciences*, 1975, 8 (3), s. 199-249.

[62] Zákon č. 86/2002 Sb., o ochraně ovzduší a o změně některých dalších zákonů (zákon o ochraně ovzduší). In: *Sbírka zákonů*. 2002 (38), s. 1786-1839. Také dostupné z: http://www.psp.cz/sqw/sbirka.sqw?r=2002&cz=86

[63] Zákon 201/2012 Sb., o ochraně ovzduší. In: *Sbírka zákonů*. 2012 (69). Také dostupné z: https://www.zakonyprolidi.cz/cs/2012-201

[64] Zákon ze dne 19. října 2016, kterým se mění zákon č. 201/2012 Sb., o ochraně ovzduší, ve znění pozdějších předpisů, a zákon č. 634/2004 Sb., o správních poplatcích, ve znění pozdějších předpisů. In: *Sbírka zákonů* č. *369 / 2016*. 2016 (147) s. 5766-5789. Také dostupné z: http://aplikace.mvcr.cz/sbirka-zakonu/ViewFile.aspx?type=z&id=61288

8 PŘÍLOHY

Příloha 1: Tabulka s hodnotami proměnných počasí modelu klasifikace kvality ovzduší; Zdroj: CHMI

Název proměnné	**Zkratka**	**Typ proměnné**	**Průměr**	**Medián**	**Min**	**Max**
Průměrná denní teplota ve městě	m_1	Spojitá proměnná	9,36	9,00	-10,92	24,01
Maximální denní teplota ve městě	m_2	Spojitá proměnná	13,84	13,50	-5,23	30,62
Minimální denní teplota ve městě	m_3	Spojitá proměnná	4,89	5,00	-15,17	17,98
Průměrná denní teplota na stanici Lysá Hora	m_4	Spojitá proměnná	3,95	3,50	-15,70	18,90
Maximální denní teplota na stanici Lysá Hora	m_5	Spojitá proměnná	7,40	7,15	-11,80	23,90
Minimální denní teplota na stanici Lysá Hora	m_6	Spojitá proměnná	1,52	0,9	-17,30	16,40
Průměrná rychlost větru	m_7	Spojitá proměnná	10,39	9	1	32
Maximální rychlost větru	m_8	Spojitá proměnná	23,6	22	6	57
Vlhkost vzduchu	m_{11}	Spojitá proměnná	77,23	77	53	98
Atmosférický tlak	m_{12}	Spojitá proměnná	1015,3	1015,9	986,9	1038,8
Průměrný minimální rozdíl v denní teplotě změřený na stanicích ve městě a na Lysé Hoře	d_1	Spojitá proměnná	-5,4	-5,7	-9,9	4,4
Maximální rozdíl v denní teplotě změřený na stanicích ve městě a na Lysé Hoře	d_2	Spojitá proměnná	-6,44	-6,60	-12,30	3,90
Minimální rozdíl v denní teplotě změřený na stanicích ve městě a na Lysé Hoře	d_3	Spojitá proměnná	-3,38	-3,85	-9,10	5,40
PM_{10} - průměrná hodnota za posledních 24 hodin	d_4	Spojitá proměnná	41,02	33,59	10,44	165,38

Příloha 2: Tabulka s hodnotami binárních proměnných výskyt mlhy (v_1), srážek (v_2), bouřky (v_3) dle záznamů z letiště Ostrava – Mošnov; Zdroj: CHMI

Název parametru	Značka proměnné	Výskyt – počet
Mlha	v_1	61
Srážky	v_2	184
Bouřka	v_3	28

Příloha 3: Tabulka s hodnotami proměnných – směr větru v dopoledních (v_4) a odpoledních (v_5) hodinách dle stanice na letišti Ostrava – Mošnov; Zdroj: CHMI

Popis – směr větru	Dopolední proudění v_4	Odpolední prodění v_5	Celkem
Bezvětří	77	28	105
Sever	26	42	68
Severoseverovýchod	30	34	64
Severovýchod	10	19	29
Východoseverovýchod	2	9	11
Východ	2	5	7
Východojihovýchod	0	0	0
Jihovýchod	1	0	1
Jihojihovýchod	0	1	1
Jih	3	25	28
Jihojihozápad	36	46	82
Jihozápad	107	65	172
Západojihozápad	36	45	81
Západ	4	10	14
Západoseverozápad	1	0	1
Severozápad	0	2	2
Severoseverozápad	8	11	19
Proměnlivý	22	21	43

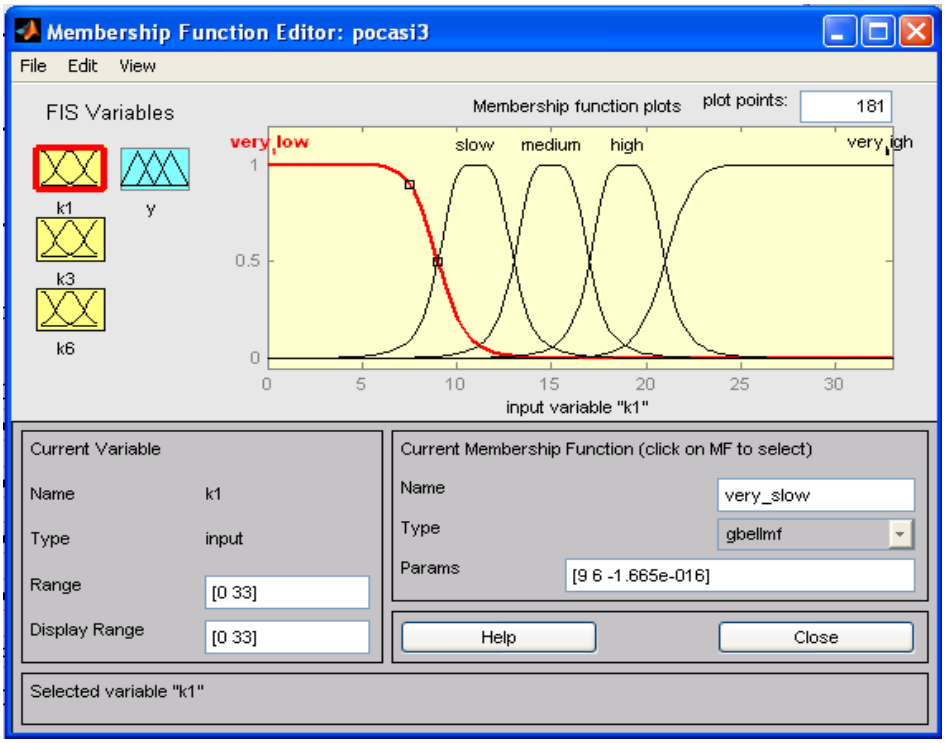
Příloha 4: Tabulka s hodnotami PM10 na stanicích v Ostravě v roce 2008 (trénovací a testovací množina modelu klasifikace kvality ovzduší); Zdroj: CHMI

Stanice	Typ	Průměr	Medián	Minimum	Maximum
Bartovice	Průmyslová	48,62	42,00	8,40	180,00
Českobratrská	Dopravní	43,13	34,00	8,00	231,00
Fifejdy	Pozaďová	40,55	33,10	8,40	188,40
Mariánské Hory	Průmyslová	41,81	37,00	7,70	156,10
Poruba	Pozaďová	29,98	24,00	6,00	146,00
Přívoz	Průmyslová	47,00	38,50	7,30	211,00
Zábřeh	Pozaďová	37,18	29,70	5,80	190,20

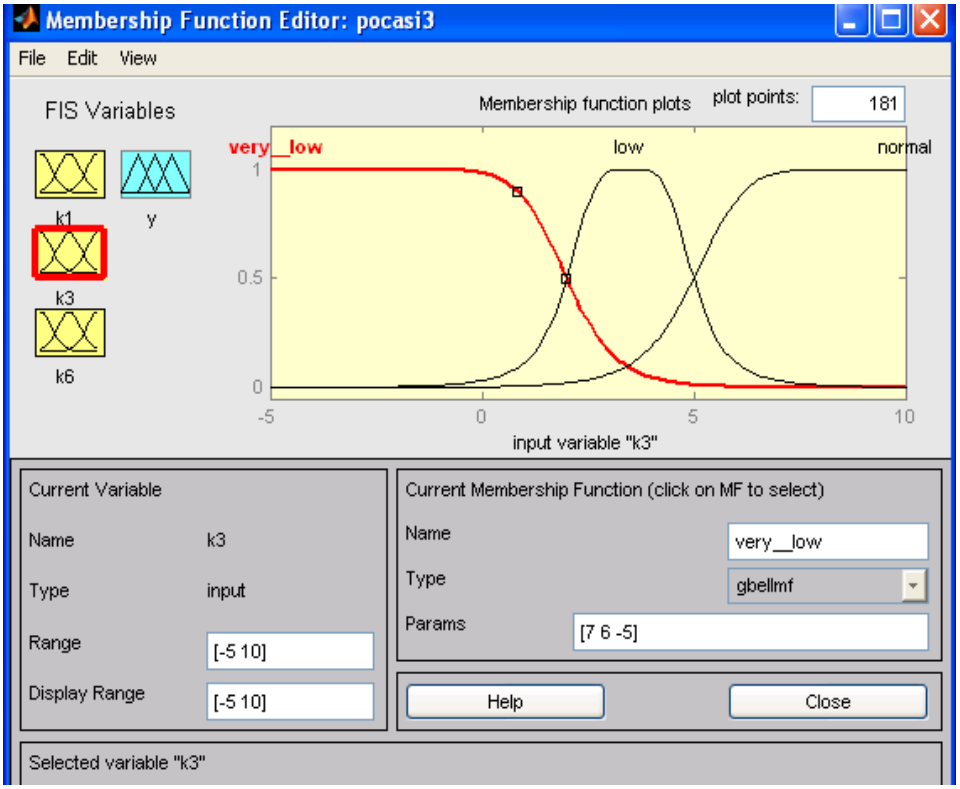
Příloha 5: Tabulka kategorizované kvality ovzduší pro jednotlivé dny 2008 na měřících stanicích v Ostravě a typ stanic; Zdroj: CHMI

Stanice	Typ	1	2	3	4	5	6
Bartovice	Průmyslová	29	59	169	59	46	4
Českobratrská	Dopravní	18	130	137	35	40	6
Fifejdy	Pozaďová	24	128	147	28	33	1
Mariánské Hory	Průmyslová	20	110	147	49	39	1
Poruba	Pozaďová	74	173	73	19	26	1
Přívoz	Průmyslová	16	96	152	51	42	9
Zábřeh	Pozaďová	48	140	117	27	28	6

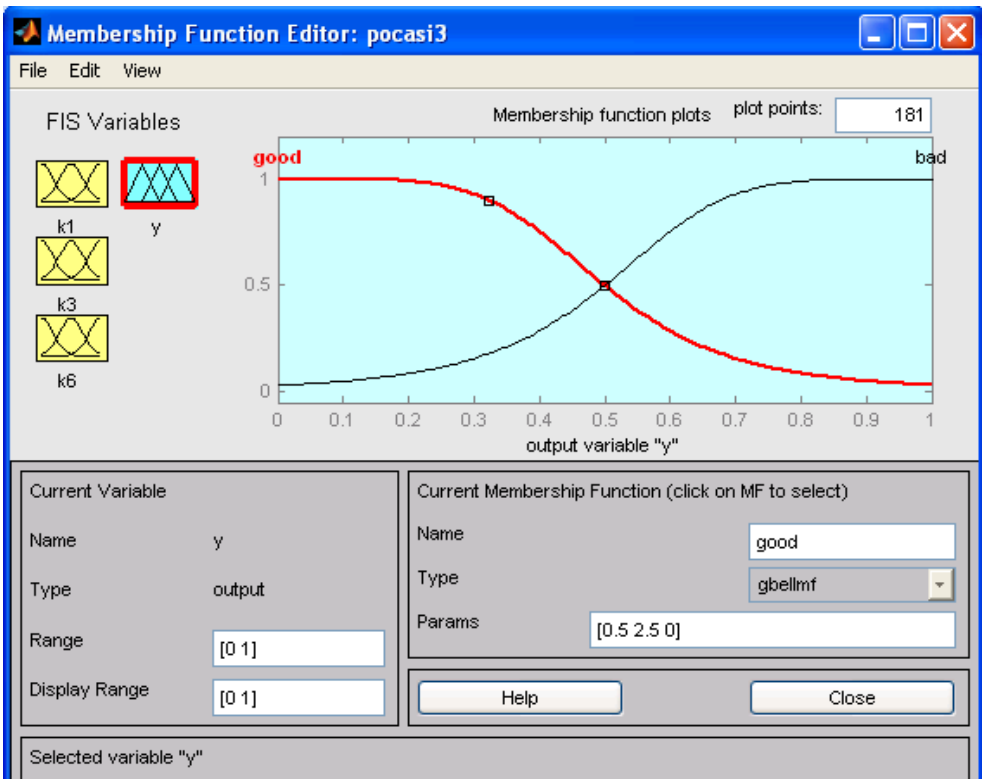
Příloha 6: Obrázek vstupní funkce příslušnosti parametru K_1 (Denní průměrná rychlost větru v m/s) modelu klasifikace kvality ovzduší pomocí RFA; Zdroj: autor



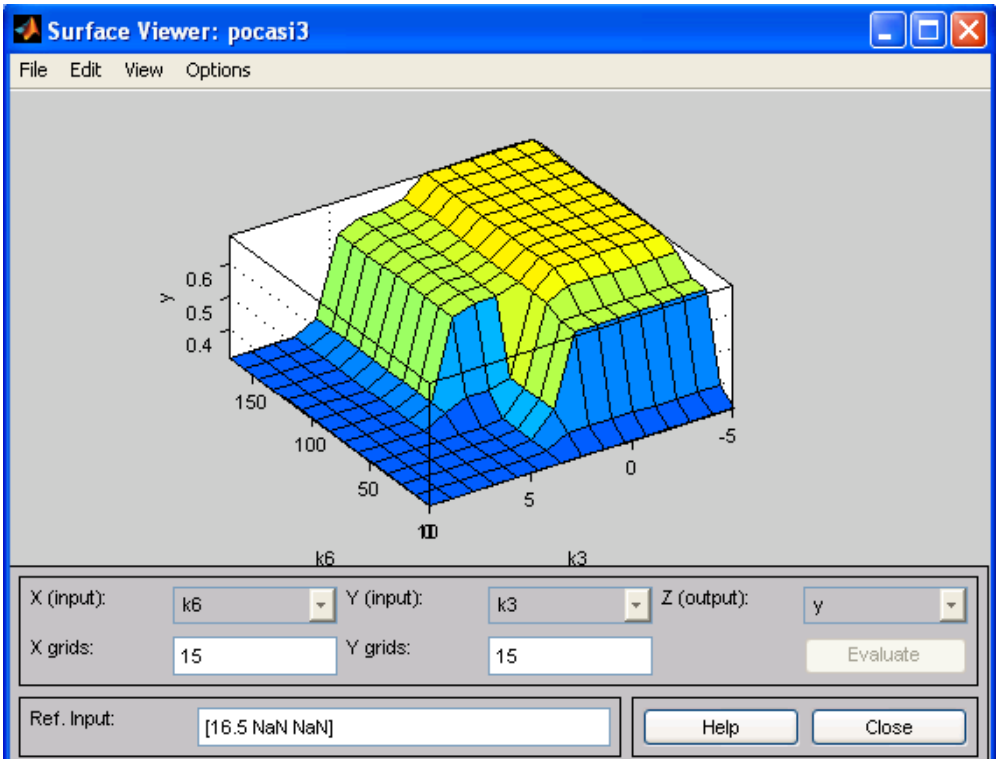
Příloha 7: Obrázek vstupní funkce příslušnosti parametru K_3 (inverzní charakter počasí v předcházejících 24 hodinách) modelu klasifikace kvality ovzduší pomocí RFA; Zdroj: autor



Příloha 8: Obrázek výstupní funkce příslušnosti parametru Y modelu klasifikace kvality ovzduší pomocí RFA; Zdroj: autor



Příloha 9: Vizualizace pravidel pomocí nástroje Surface Viewer pro vstupní parametry k_3 a k_6; Zdroj: autor



Příloha 10: CBR citlivostní analýza s porovnáním jednotlivých metrik využitých pro výpočet nejbližšího souseda (v [%]), šedě označené výsledky mají v rámci daného počtu vstupních parametrů nejvyšší přesnost; Zdroj: autor

počet vstupů	**Vítr - prům.**	**Vítr – max.**	**Vlhkost**	**Tlak**	**Inverzní charakter poč.**	**Stav PM_{10} v T-1**	**Euklidovská metrika**	**Manhattan metrika**	**Čebyševova metrika**
6	1	1	1	1	1	1	**94,69 %**	93,51 %	93,81 %
5	0	1	1	1	1	1	**94,69 %**	94,40 %	94,40 %
5	1	0	1	1	1	1	**94,10 %**	93,22 %	93,51 %
5	1	1	0	1	1	1	86,43 %	**93,22 %**	85,84 %
5	1	1	1	0	1	1	**94,10 %**	85,55 %	92,92 %
5	1	1	1	1	0	1	92,33 %	**93,81 %**	92,04 %
5	1	1	1	1	1	0	91,15 %	90,56 %	**91,45 %**
4	0	0	1	1	1	1	**94,10 %**	92,33 %	91,74 %
4	0	1	0	1	1	1	**87,91 %**	86,73 %	87,32 %
4	0	1	1	0	1	1	**94,40 %**	94,40 %	93,51 %
4	0	1	1	1	0	1	**92,92 %**	92,63 %	91,74 %
4	0	1	1	1	1	0	**93,51 %**	92,63 %	92,04 %
4	1	0	0	1	1	1	85,25 %	84,66 %	**86,14 %**
4	1	0	1	0	1	1	**93,22 %**	92,63 %	92,63 %
4	1	0	1	1	0	1	**92,33 %**	92,33 %	91,74 %
4	1	0	1	1	1	0	89,38 %	89,97 %	**91,15 %**
4	1	1	0	0	1	1	84,07 %	**84,96 %**	82,89 %
4	1	1	0	1	0	1	**83,48 %**	82,89 %	82,60 %
4	1	1	0	1	1	0	**80,53 %**	79,06 %	79,35 %
4	1	1	1	0	0	1	**94,40 %**	92,63 %	93,22 %
4	1	1	1	0	1	0	91,15 %	**90,56 %**	90,27 %
4	1	1	1	1	0	0	**86,14 %**	85,55 %	85,55 %
3	0	0	0	1	1	1	87,02 %	**87,91 %**	86,14 %
3	0	0	1	0	1	1	93,22 %	**93,81 %**	92,33 %
3	0	0	1	1	0	1	89,09 %	**91,15 %**	87,61 %
3	0	0	1	1	1	0	**89,09 %**	88,79 %	**89,09 %**
3	0	1	0	0	1	1	**86,43 %**	86,14 %	84,37 %
3	0	1	0	1	0	1	87,02 %	**87,32 %**	82,89 %
3	0	1	0	1	1	0	82,01 %	**82,60 %**	82,30 %
3	0	1	1	0	0	1	**93,81 %**	93,22 %	92,92 %
3	0	1	1	0	1	0	**92,92 %**	92,33 %	91,45 %
3	0	1	1	1	0	0	84,96 %	**85,55 %**	83,19 %
3	1	0	0	0	1	1	**84,66 %**	84,37 %	84,37 %
3	1	0	0	1	0	1	**83,48 %**	82,89 %	82,89 %
3	1	0	0	1	1	0	76,11 %	**84,66 %**	76,70 %
3	1	0	1	0	0	1	**93,22 %**	92,04 %	92,33 %
3	1	0	1	0	1	0	89,09 %	87,61 %	**90,86 %**
3	1	0	1	1	0	0	84,37 %	**84,66 %**	83,78 %
3	1	1	0	0	0	1	83,78 %	**84,66 %**	81,42 %
3	1	1	0	0	1	0	77,29 %	**79,65 %**	75,22 %
3	1	1	0	1	0	0	78,47 %	**79,35 %**	76,11 %
3	1	1	1	0	0	0	85,84 %	85,84 %	**86,14 %**

Příloha 11: Datový slovník atributů modelu klasifikace klientů v nebankovní instituci; Zdroj: autor

Atribut	Minimum	Maximum	Průměr	Medián	Počet hodnot
Objem – první 3 měsíce (v CZK)	200	560 499 770	2 859 328,3	698 130	1 323
Objem – prvních 6 měsíců (v CZK)	8261	607 994 970	4 803 449	1 143 212,5	1 323
Objem – druhých 6 m. (v CZK)	0	266 545 066	3 005 337,5	337743	1 323
Objem – prvních 12 m. (v CZK)	9940	841 019 559	7 808 786,5	1 888 110	1 323
Profit – první 3 m. (v CZK)	0,47	545 109	4 139,1	1 500	1 323
Profit – prvních 6 m. (v CZK)	16,68	564 309	7 400	2 500	1 323
Profit – druhých 6 m. (v CZK)	0,00	236 218	5 702,4	1 020,8	1 323
Profit – prvních 12 m. (v CZK)	36,40	674 184	13 102,4	4 140,9	1 323
Počet transakcí za první 3 měsíce	1	71	5,29	3	1 323
Počet transakcí za prvních 6 m.	1	130	9,68	5	1 323
Počet transakcí za druhých 6 m.	0	260	7,56	3	1 323
Počet transakcí za prvních 12 m.	1	356	17,24	9	1 323
Doba od podepsání RS do první transakce (dny)	0	91	14,55	6	1 323
Doba mezi první a druhou transakcí (dny)	0	91	19,08	13	996
Doba mezi druhou a třetí transakcí (dny)	0	79	14,68	10	772
Doba mezi třetí a čtvrtou transakcí (dny)	0	86	11,79	8	605
Doba mezi čtvrtou a pátou transakcí (dny)	0	57	9,81	7	472
Doba mezi pátou a šestou transakcí (dny)	0	39	7,14	6	398
Objem první transakce (v CZK)	199,74	25 624 000	597 666,1	220 000	1 323
Objem druhé transakce (v CZK)	24,98	25 075 000	522 485,2	161 327,7	996
Objem třetí transakce (v CZK)	59,23	24 595 000	438 319,5	130 108,7	772
Objem čtvrté transakce (v CZK)	14,82	16 300 000	385 302,1	127 116,9	605
Objem páté transakce (v CZK)	694,63	12 605 000	381 128,8	124 046,7	472
Objem šesté transakce (v CZK)	58,69	7 696 500	327 358,0	106 610,5	398
Profit první transakce (v CZK)	0,47	26 012	859,6	381,8	1 323
Profit druhé transakce (v CZK)	0,07	26 576	752,2	349,6	996
Profit třetí transakce (v CZK)	0,18	36 784	732	312,5	772
Profit čtvrté transakce (v CZK)	0,05	13 120	630,8	297,4	605
Profit páté transakce (v CZK)	2,84	10 583	654,8	285,2	472
Profit šesté transakce (v CZK)	0,18	9 081	586,4	259,4	398

Příloha 12: Tabulka se středními hodnotami atributů klientů v rámci jednotlivých shluků za rok 2013 vytvořených pomocí nehierarchické shlukovací metody *k*-means; Zdroj: autor

Shluk	**Typ shluku**	**Obrat klienta (v CZK)**	**Profit klienta (v CZK)**	**Počet transakcí**	**Průměrný profit na transakci**
k_1	Import	52 427 378	148 156	272.06	550.70
k_2	Export	34 836 204	81 686	68.55	1 188.70
k_3	Import	8 257 069	28 148	94.46	299.20
k_4	Export	9 024 410	20 901	23.61	887.90
k_5	Import	2 216 415	7 316	32.80	220.50
k_6	Export	3 777 535	7 549	7.21	1 041.10
k_7	Import	698 142	1 993	12.24	164.70
k_8	Export	978 580	1 931	3.13	612.40
k_9	Nedefinováno	194 181	472	3.48	135.00
k_{10}	Nedefinováno	32 836	82	3.18	25.40