# MODELOVÁNÍ KVALITY ŽIVOTA POMOCÍ ROZHODOVACÍCH STROMŮ

E + M

**Jiří Křupka, Miloslava Kašparová, Pavel Jirava**

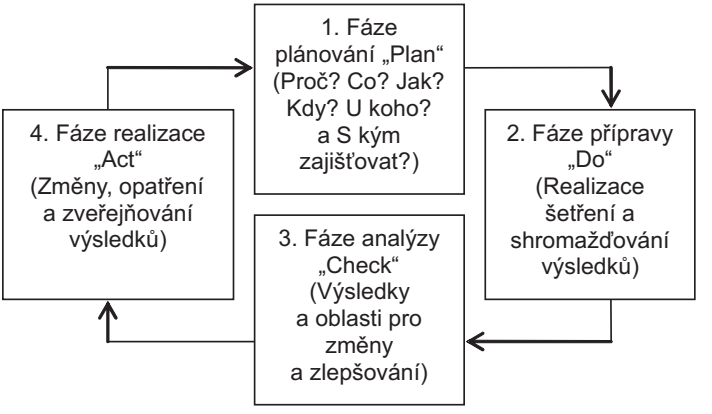
Příspěvek prezentuje možnost využití teorie rozhodovacích stromů pro modelování problematiky kvality života ve vybrané lokalitě České republiky. Jde o klasifikaci spokojenosti obyvatel města Chrudimi s kvalitou okolního životního prostředí. Navrhnuté modely pracují s reálnými daty daty získanými z dotazníkového šetření.

## Úvod

Definice regionálního managementu [2,15] je v podstatě analogií k definicím podnikového managementu. V [2] je uvedeno její zobecnění: „... cílem regionálního managementu je prospěch a rozvoj regionu, uspokojování zájmů a potřeb obyvatelstva a jeho skupin a veřejný zájem ...“, které je v souladu s definicí v zákonech č. 128/2000 Sb., o obcích (obecní zřízení) a č. 129/2000 Sb., o krajích (krajské zřízení). Ministerstvo vnitra podporuje zavádění nástrojů řízení kvality ve veřejné správě (VS). To se odráží i v návrhu Strategie Národní politiky kvality v České republice (ČR) na období let 2008 až 2013. Návrh Strategie vychází z analýzy výsledků dosavadního plnění Národní politiky podpory jakosti a na základě vyhodnocení současné situace definuje na příští období poslání, vizi, rámec a dlouhodobý strategický cíl. Tímto cílem je spoluvytvářet v ČR prostředí, ve kterém je úsilí o vysokou kvalitu trvalou součástí všech oblastí života společnosti i občanů včetně růstu kvality života, který je veden cestou udržitelného rozvoje [37]. Strategie má být přínosem pro podnikatelskou sféru, veřejnou správu i pro celou občanskou společnost. V oblasti veřejného sektoru má Strategie podporovat rozvoj kvality VS. Nejčastěji používané nástroje, metody a modely pro hodnocení kvality managementu VS jsou uvedeny např. v [14,15,19,24]. V dané oblasti je možné použít také přímé měření spokojenosti občanů, vycházející z dotazníkového šetření.

Způsob zjišťování informace pomocí přímého měření spokojenosti občanů poskytuje managementu zpětnou informaci o názorech občanů přímo od nich samotných. Přímé měření má své nesporné výhody, ale přináší i některé problémy. Ačkoliv je sám pojem „spokojenost“ definován

***Obr. 1: Projekt a cyklus šetření spokojenosti***




Fáze Projektu šetření spokojenosti

Zdroj: upraveno podle [24]

normou ČSN EN ISO 9000:2001, je zřejmá jeho nejednoznačnost [15]. Přes tuto nejednoznačnost jsou informace získané měřením spokojenosti občanů velice důležitou a mnohdy jedinou zpětnou vazbou managementu VS. Výzkumy, průzkumy a šetření týkající se spokojenosti občanů s různými stránkami života jsou v praxi často užívaným nástrojem získávání informací. Provádějí se průzkumy spokojenosti občanů [17] s osobním životem, životem v obci, fungováním samosprávy, politickou situací, bezpečností, životním prostředím, kulturou v obci atd. Jeden z možných modelů šetření - model [24] Projektu a cyklu šetření spokojenosti - vychází z obecného cyklu PDCA (Plan, Do, Check a Act), který je na obr. 1.

Na evropské, národní i regionální úrovni provádějí šetření spokojenosti Eurobarometr (Evropská komise), Český statistický úřad, Centrum pro výzkum veřejného mínění (Sociologický ústav AV ČR), agentury pro výzkum veřejného mínění, média i samotné orgány státní správy a územní samosprávy. V praxi městských úřadů se provádějí nejčastěji obecná šetření spokojenosti občanů s životem v obci a šetření spokojenosti „zákazníků" městského úřadu s jeho službami. Častější je tato praxe pochopitelně ve větších městech. Povinnost provádět tato šetření není explicitně vymezena zákonem. Významnou roli proto hraje iniciativa obecních zastupitelstev a rad, zejména pak osobní angažovanost městských manažerů (starostů, tajemníků, vedoucích odborů atd.).

Na základě konzultací s pracovníky městských úřadů z několika českých měst vyplynulo [15], že některá města dříve šetření spokojenosti ve vybraných letech prováděla. Jde např. o projekt občanského sdružení TIMUR (Týmové iniciativy pro místní udržitelný rozvoj), který pilotně aplikoval Hradec Králové a Vsetín v roce 2003. Následující řešený problém (případová studie) vychází z dotazníkového šetření v Chrudimi, město je administrativně začleněno do NUTS III - Pardubický kraj, NUTS II – Severovýchod.

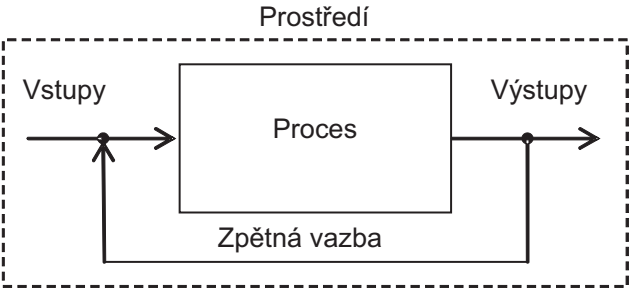
## 1. Formulace problému

Případová studie se týká modelování spokojenosti občanů s kvalitou životního prostředí ve městě Chrudim. Toto město využívá zveřejněných výsledků dotazníkových průzkumů ze sady Společné evropské indikátory jako zpětné vazby při řízení a rozvoji města. Dávají se na vědomí radě a zastupitelstvu města, odborům městského úřadu, organizacím města (Technické služby). Výsledky jsou využívány např. při rozhodování o přidělování dotací.

Vzhledem k povaze problematiky se jeví výhodné přistoupit ke zpracování dat o spokojenosti občanů jako k úlohám z oblasti dobývání znalostí. Standardem v dobývání znalostí je metodika CRISP-DM (CRoss-Industry Standard Process for Data Mining), která člení celý proces do šesti základních etap. Jde o [13,16]: porozumění problematice, porozumění datům, přípravu dat, modelování, hodnocení a využití výsledků. Podkladem pro modelování budou vybraná data získaná dotazníkovým šetřením Indikátoru A1 - Spokojenost s místním společenstvím. Tento indikátor patří v rámci „Indikátorů udržitelného rozvoje na místní úrovni" do skupiny indikátorů „Společné evropské indikátory". Společné evropské indikátory jsou sada deseti indikátorů, které odráží rozličné aspekty života a řízení města, při akceptování rovnováhy mezi pilíři udržitelného rozvoje - sociálním, ekonomickým a environmentálním [31]. Umožňuje sběr srovnatelných údajů v rámci celé Evropy a v rámci srovnatelně velkých sídelních útvarů. Projekt vychází z iniciativy skupin kolem Evropské komise [15,21]. Daný indikátor zjišťuje a vyčísluje subjektivní pocit spokojenosti občanů s městem, ve kterém žijí a pracují, a dílčí aspekty této spokojenosti.

První fází dobývání znalostí na základě metodiky CRISP-DM je pochopení cílů úlohy z manažerského hlediska a její převod na úlohu dobývání znalostí. V tomto případě, kde se zabýváme spokojeností občana s životním prostředím, je v manažerské roli regionální management. Manažerskou úlohou je hledání souvislostí mezi demografickými a jinými údaji o občanech, dílčími aspekty spokojenosti občanů, a především modelování spokojenosti občanů s životním prostředím s ohledem na regionální rozvoj a kvalitu života jedinců. Pokud by se podařilo definovat skupiny občanů, kteří jsou nespokojeni, mohou získané znalosti sloužit jako podklady pro stanovení priorit při dalším rozhodování a řízení rozvoje regionu. Hovoříme-li o řízení, musíme předpokládat, že jde o dynamický systém a akceptujeme platnost teorie systémů. Kybernetické principy řízení dynamického systému jsou obecně vyjádřeny na obr. 2. V procesu jsou dva prvky (řídící, který je představován regionálním managementem a řízený, který můžeme chápat jako region) a vazba

***Obr. 2: Model řízení systému podle Norberta Wiesnera***




Zdroj: převzato z [27]

mezi těmito prvky, která reprezentuje řídící zásah. Vstupy jsou plánované požadavky do řídícího prvku a vnější působení na řízený prvek. Výstupem je změna kvality života [28,34]. Ve zpětné vazbě se nachází model hodnocení spokojenosti jako subsystém systému řízení, podrobněji v [15].

Z hlediska dobývání znalostí lze formulovat v této oblasti řadu úloh. S ohledem na zaměření se jedná o nalezení takových demografických či jiných atributů o občanech, které mají vliv na určení spokojenosti občanů s životním prostředím, popřípadě na základě určení spokojenosti občana vytvořit klasifikátor, který na základě dostupných a použitých atributů zařadí každého ob-

***Tab. 1: Přehled původních 10 atributů***

| Atribut | Popis atributu | Měřitelnost |
|---|---|---|
| x1 | ID respondenta | Číslo |
| x2 | Pohlaví | Muž/žena |
| x3 | Věk | Číslo |
| x4 | Zaměstnání | student (1) zaměstnaný (2) nezaměstnaný (3) důchodce (4) |
| x6 | Počet cigaret denně | Číslo |
| x11 | Spokojenost s veřejnými službami | Bodové hodnocení v rozsahu 0 až 10, kde 0 je velice nízká, 10 je velmi vysoká) |
| x13 | Spokojenost s kvalitou okolního životního prostředí | Bodové hodnocení v rozsahu 0 až 10, kde: 0 je velice nízká, 10 je velmi vysoká) |
| x15 | Počet hodin v průměru týdně aktivního pohybu nebo sportu | 0, popřípadě výběr z intervalu <1,3>, <4,7>, <8,10>, <11,15>, <16,20> a více než 20 |
| x16 | Vyjádření, zda prostředí a životní podmínky ve městě mají vliv na zdraví dotazovaného | Pomocí stupnice v rozsahu 1 až 5, kde: 1 je ano velmi, 2 je spíše ano, 3 je nevím, 4 je spíše ne, 5 je určitě ne |
| x17 | Úroveň spokojenosti s možností relaxace a odpočinku | Bodové hodnocení v rozsahu 0 až 10, kde: 0 je velice nízká, 10 je velmi vysoká) |

Zdroj: převzato z [15]

čana podle uvedené spokojenosti do příslušné kategorie. Úloha proto může být pojata jako klasifikační, v které je cílem zařazení občanů do tříd podle určení jejich spokojenosti a v které bude využito rozhodovacích stromů.

## 1.1 Sběr a předzpracování dat

Spokojenost člověka je jednou ze základních podmínek určujících kvalitu jeho života. Je to však kategorie (veličina) značně subjektivní a měnící se v čase a sám pojem spokojenost je rozsáhlý a neurčitý. Naměřená data, získaná z dotazníkového šetření, jsou zatížena neurčitostí. Dotazníkové šetření bylo provedeno na území města Chrudim na podzim roku 2007. Na základě vybraných atributů vycházejících z indikátoru A1 byl vytvořen datový soubor, obsahující 701 záznamů (objektů, příkladů, respondentů) a 10 atributů (vlastností, proměnných, dotazníkových otázek).

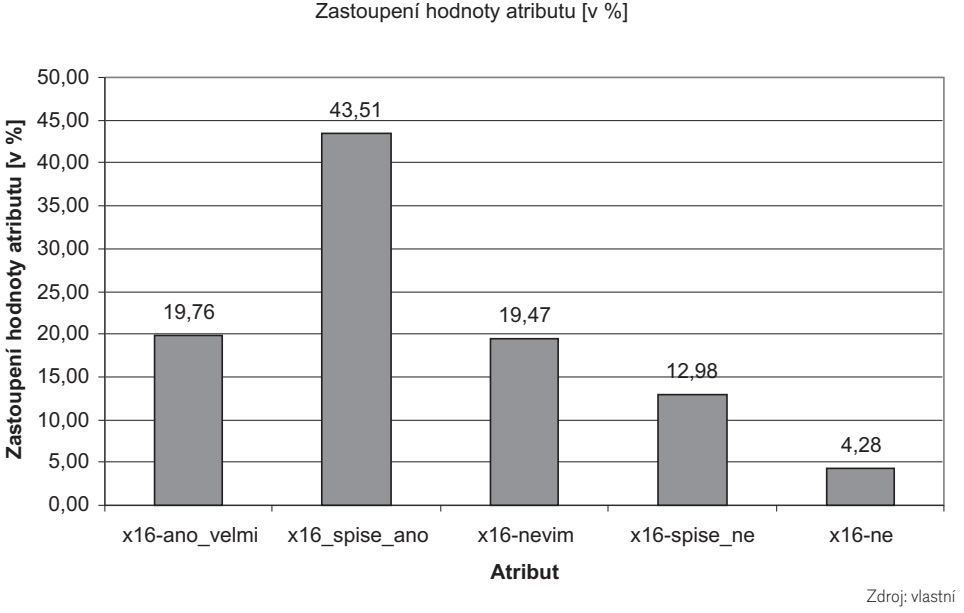
Přehled těchto atributů, které lze rozdělit na atributy obecné a specifické, je následující: obecné (pohlaví, věk, zaměstnání), specifické (počet vykouřených cigaret denně, spokojenost se základními veřejnými službami (zdravotní a sociální služby, školy, veřejná doprava atd.), spokojenost s kvalitou okolního životního prostředí, počet hodin v průměru týdně aktivního pohybu nebo sportu, vyjádření, zda prostředí a životní podmínky ve městě mají vliv na zdraví dotazovaného, úroveň spokojenosti s možností relaxace a odpočinku). Vzhledem ke skutečnosti, že dané atributy byly hodnoceny různě [15] (tab. 1), bylo nutné atributy i jejich hodnoty upravit do formy použitelné pro zvolený klasifikační algoritmus.

Z tab. 1 byly analyzovány uvedené atributy $x_1$, $x_2$, ..., $x_{17}$ z pohledu četností hodnot pro jednotlivé atributy. Příklad je uveden pro atribut $x_{16}$ (obr. 3), kde $x_{16}$ se dále dělí na: x16-ano_velmi, x16_spise_ano, x16-nevim, x16-spise_ne a x16-ne.

Na základě úprav souboru, tj. zastoupení hodnoty atributu, odstranění odlehlých a chybných hodnot, úpravy atributů (převod atributů na dichotomické proměnné a diskretizace proměnných věk, počet cigaret denně a proměnných hodnocených pomocí bodového hodnocení v rozsahu od 0 do 10 viz tab. 1) byl získán datový soubor obsahující 691 záznamů popsaných 33 atributy, které byly použity pro tvorbu klasifikačního modelu (tab. 2).

U atributu *počet cigaret denně* byla provedena diskretizace na předem zadaný počet intervalů,

***Obr. 3: Zastoupení výskytu hodnot atributu $x_{16}$ „Vyjádření, zda prostředí a životní podmínky ve městě mají vliv na zdraví dotazovaného“***




Zdroj: vlastní

a to: hodnota 2 pro interval <1,9>, hodnota 3 pro interval <10,19>, hodnota 4 pro 20 cigaret a více denně. V případě nekuřáků byla přidělena hodnota 1. Atribut *věk* byl diskterizován na předem zadaný počet ekvifrekvenčních intervalů, a to: hodnota 1 pro respondenty ve věku 15 až 29, respondenti ve věku od 30 do 41 let byli označeni hodnotou 2, hodnota 3 odpovídá intervalu <42,50> a hodnota

***Tab. 2: Přehled upravených 33 atributů***

| Atribut | Popis atributu | Rozsah hodnot | Počet |
|---|---|---|---|
| a1 | x4-student | {0,1} | 691 |
| a2 | x4-zam | {0,1} | 691 |
| a3 | x4-nezam | {0,1} | 691 |
| a4 | x4-duch | {0,1} | 691 |
| a5 | x16-ano_velmi | {0,1} | 691 |
| a6 | x16_spise_ano | {0,1} | 691 |
| a7 | x16-nevim | {0,1} | 691 |
| a8 | x16-spise_ne | {0,1} | 691 |
| a9 | x16-ne | {0,1} | 691 |
| a10 | x15-_sport_0 | {0,1} | 691 |
| a11 | x15-sport_1-3 | {0,1} | 691 |
| a12 | x15-_sport_4-7 | {0,1} | 691 |
| a13 | x15-_sport_8-10 | {0,1} | 691 |
| a14 | x15-_sport_11-15 | {0,1} | 691 |
| a15 | x15-_sport_16-20 | {0,1} | 691 |
| a16 | x15-_sport_vicnez20 | {0,1} | 691 |
| a17 | x3-vek_skup_29 | {0,1} | 691 |
| a18 | x3-vek_skup_41 | {0,1} | 691 |
| a19 | x3-vek_skup_50 | {0,1} | 691 |
| a20 | x3-vek_skup_82 | {0,1} | 691 |
| a21 | x6-nekurak | {0,1} | 691 |
| a22 | x6_kurak_9 | {0,1} | 691 |
| a23 | x6_kurak_19 | {0,1} | 691 |
| a24 | x6_kurak_40 | {0,1} | 691 |
| a25 | x2_zena | {0,1} | 691 |
| a26 | x2_muz | {0,1} | 691 |
| a27 | x17_relax_nizka | {0,1} | 691 |
| a28 | x17_relax_prumer | {0,1} | 691 |
| a29 | x17_relax_vysoka | {0,1} | 691 |
| a30 | x11_sluzby_nizka | {0,1} | 691 |
| a31 | x11-sluzby_prumer | {0,1} | 691 |
| a32 | x11_sluzby_vysoka | {0,1} | 691 |
| a33 | x13-kvalitaZP | {1,2,3} | 691 |

Zdroj: vlastní

4 respondentům ve věku od 51 do 82 let. V každém intervalu uvedeného atributu bylo zařazeno cca 175 objektů.

U proměnné *počet hodin v průměru týdně aktivního pohybu nebo sportu* bylo provedeno ohodnocení jednotlivých intervalů hodnotami 1 nebo 0. V případě platnosti vybraného intervalu z uvedených (viz tab. 2) byl atribut ohodnocen číslem 1, jinak 0. Stejným způsobem byl zpracován i atribut *vyjádření, zda prostředí a životní podmínky ve městě mají vliv na zdraví dotazovaného*. U atributů hodnocených pomocí bodů v rozsahu od 0 do 10 bylo provedeno následující: body v intervalu od 0 do 3 byly označeny číslem 1 (spokojenost nízká), od 4 do 6 číslem 2 (spokojenost průměrná) a body od 7 do 10 (spokojenost vysoká) číslem 3. Jednalo se o proměnnou *spokojenost s veřejnými službami, úroveň spokojenosti s možností relaxace a odpočinku a spokojenost s kvalitou okolního životního prostředí.*

Data jsou uložena v matici ***M***, která je tvořena *n* řádky a *m* sloupci, kde *n* = 691 a *m* = 33:

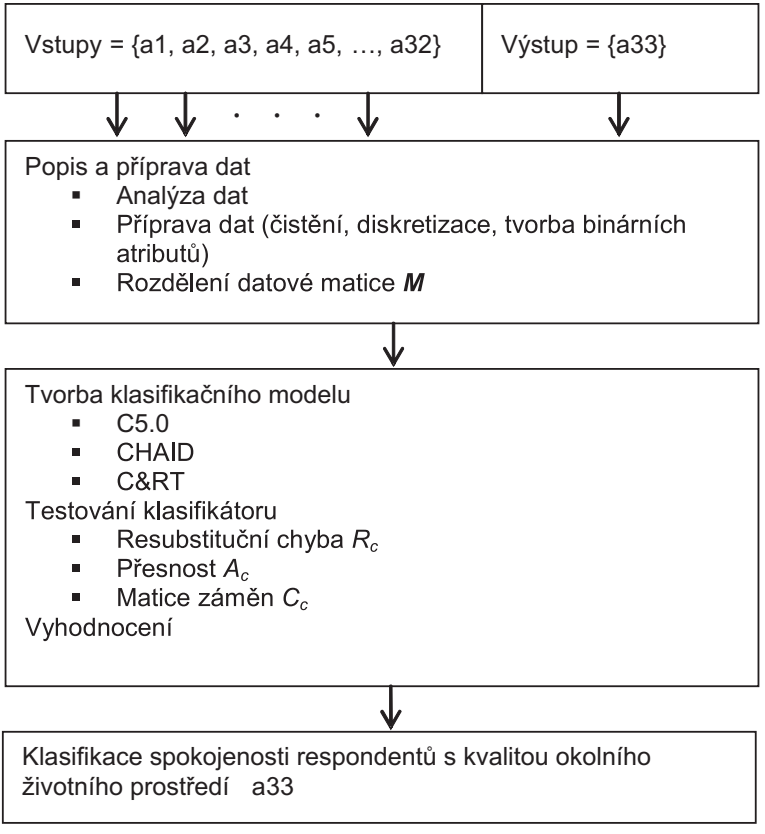
$$\boldsymbol{M} = \begin{bmatrix} h_{11} & h_{12} & \vdots & h_{1m} \\ h_{21} & h_{22} & \vdots & h_{2m} \\ \vdots & \vdots & & \vdots \\ h_{n1} & h_{n1} & \vdots & h_{nm} \end{bmatrix}. \qquad (1)$$

Jestliže řádky matice reprezentující sledované objekty, potom lze i-tý objekt zapsat jako:

$$o_i = [h_{i1}, h_{i2}, ..., h_{im}], \text{ kde } i = 1, 2, ..., 691. \qquad (2)$$

Data získaná při šetření spokojenosti občanů lze zpracovávat pomocí řady metod umělé a výpočetní inteligence. Pro řešení dané úlohy jsou zvo-

***Obr. 4: Návrh modelu klasifikátoru***




Zdroj: vlastní

leny klasifikační systémy založené na pravidlech. Jedná se o rozhodovací (klasifikační) stromy.

## 1.2 Návrh modelu klasifikátoru

K známým klasifikačním modelům patří algoritmy pro vytváření rozhodovacích stromů. Rozhodovací strom [3] můžeme definovat jako strom (stromový graf), kde každý nelistový uzel stromu představuje test hodnoty atributu a větve vedoucí z tohoto uzlu možné výsledky testu. Listové uzly stromu jsou ohodnoceny identifikátory tříd (výsledky klasifikace). Vlastní klasifikace pomocí stromu probíhá cestou záznamu od kořene stromu k jeho listu. V každém kroku je záznam testován podle testu v aktuálním uzlu rozhodovacího stromu a dále pokračuje po větvi shodné s konkrétním výsledkem testu. Pokud takto záznam dojde až do listového uzlu, je oklasifikován třídou identifikovanou hodnotou příslušného listu rozhodovacího stromu. Atribut vhodný pro větvení stromu vybíráme na základě jeho charakteristik převzatých z teorie informace a pravděpodobnosti: entropie, informačního zisku, poměrného informačního zisku, Chi-square testu, Giniho indexu a dalších.

Intuitivní vizuální zobrazení stromem napomáhá jasnějšímu pochopení výsledků a vztahů i laickým uživatelům a v praxi tak usnadňuje jejich rozhodování. Stromové grafy dovolují vizuálně prozkoumat výsledky a posoudit vhodnost modelu. Rozhodovací strom lze poměrně snadno převést na rozhodovací pravidla. Každé cestě stromem od kořene k listu odpovídá jedno pravidlo. Nelistové uzly jsou předpoklady, listový uzel pak závěrem pravidla. Mezi nejznámější a běžně užívané algoritmy pro vytváření rozhodovacích stromů [3,30,40] patří například: C4.5, C5.0, C&RT (Classification & Regression Trees), Chi-square Automatic Interaction Detection (CHAID)). C4.5 je rozšířením verze ID3, umožňuje práci s numerickými atributy, chybějícími hodnotami, převod na pravidla i prořezávání. Algoritmus C5.0 je modifikací uvedené verze. C&RT umožňuje vytvářet vedle klasifikačních stromů i stromy regresní. Uvedené C&RT jsou binární. Algoritmus CHAID používá jako kritérium pro větvení Chi-square test. Tento algoritmus seskupuje hodnoty kategoriálních atributů, při větvení se nevytváří tolik větví, kolik má atribut hodnot. Hodnoty atributu se postupně seskupují z původního počtu až do dvou skupin a poté se vybere atribut a jeho kategorizace, která je v daném kroku pro větvení nejlepší [3], více např. v [10,13,25,38]. Podrobnější popis algoritmů je uveden v [16,26].

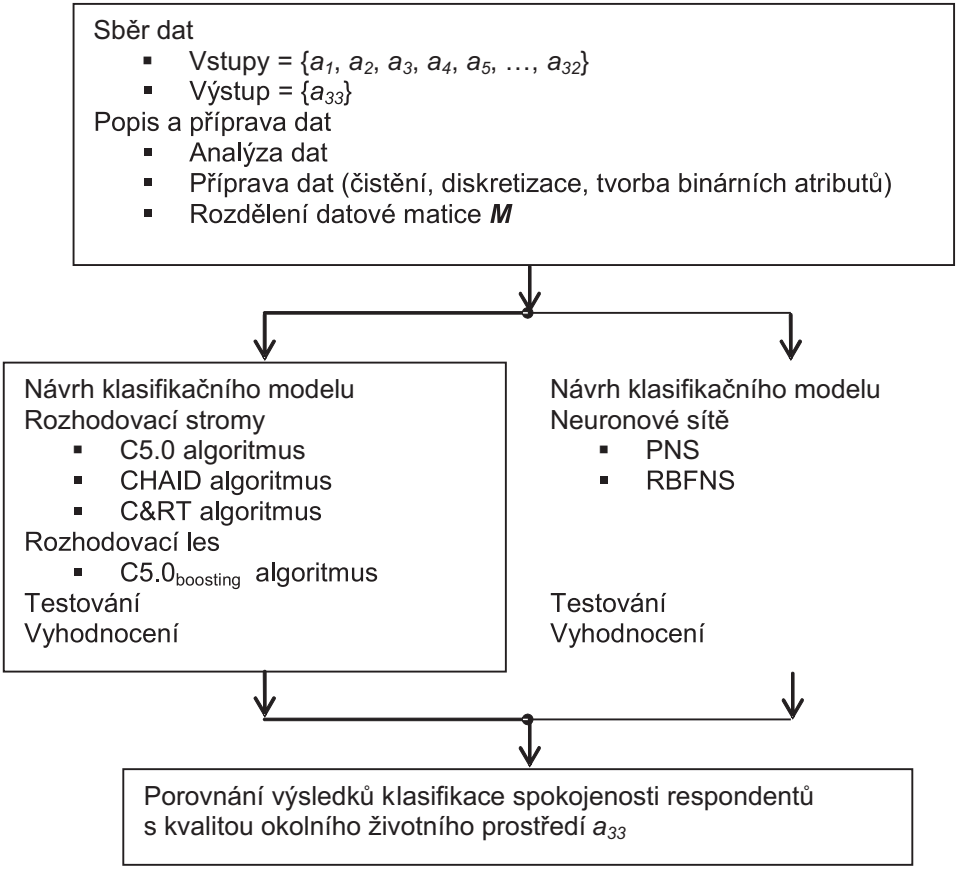
Modelování je rozšířenou metodou nacházející uplatnění v řadě oblastí společenské praxe. V našem případě se jedná o modelování spokojenost občanů s okolním životním prostředím na základě atributů uvedených v kapitole 2.1. Jednotlivé objekty jsou popsány pomocí 32 nezávislých atributů a 1 závislým atributem (*spokojenost respondentů s kvalitou okolního životního prostředí*) [15].

Pro vytvoření modelů a jeho ověření je nutné rozdělit datovou matici na matici, obsahující data trénovací a matici určenou k testování. Trénovací data jsou uložena v matici $\boldsymbol{M}_{TR}$, tvořené *n* řádky a *m* sloupci, kde *n* odpovídá náhodnému výběru 66,67 % trénovacích příkladů z $\boldsymbol{M}$ a testovací data v matici $\boldsymbol{M}_{TE}$, která odpovídá 33,33 % z výchozí datové matice $\boldsymbol{M}$.

Pro tvorbu modelu byly zvoleny algoritmy C5.0, CHAID a C&RT. Obecné schéma klasifikačního modelu spokojenosti „občanů" – „respondentů" s kvalitou okolního životního prostředí uvádí obr. 4.

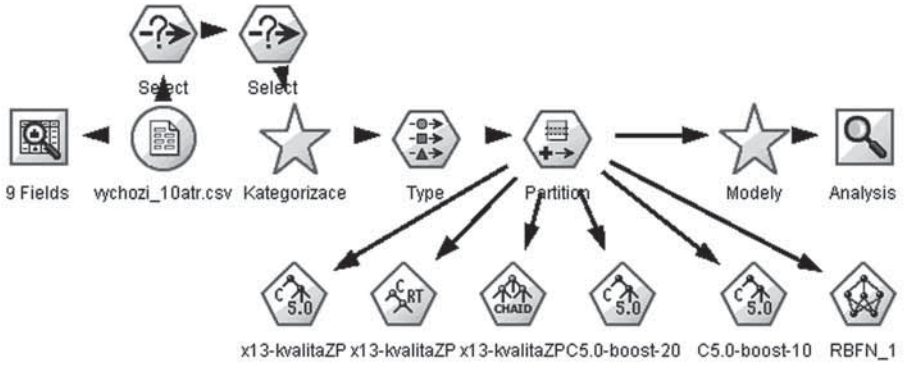
Výsledky klasifikátoru na bázi rozhodovacích stromů - v tomto případě je možné hovořit i o rozhodovacím lese (z ang. decision forest [26] nebo decision wood [1]), který využívá myšlenku skupinového7 rozhodování pro množinu rozhodovacích stromů. Skupinové rozhodování může být realizováno paralelním nebo hierarchickým spojením modelů rozhodovacích stromů. V prvním případě je možné pro jejich výsledné ohodnocení použít ohodnocení výstupů jednotlivých modelů na základě průměru, váženého průměru, mediánu, modusu, váženého modusu atd. V druhém případě, kdy se jedná o hierarchickou stavbu modelu [26], je možné použít algoritmus C5.0 zahrnující speciální proceduru nazývanou boosting, která ve většině případů vede ke zvýšení přesnosti klasifikace. Tato metoda pracuje na bázi tvorby série modelů. První model je vytvořen obvyklým způsobem. Při budování druhého modelu je pozornost věnována převážně objektům, které byly v prvním modelu zaklasifikovány chybně. Následující modely se vždy vytváří na základě chyb v předchozích modelech. Na závěr jsou objekty klasifikovány použitím celé sady modelů a váženého hlasování, více v [26,30,40].

***Obr. 5: Porovnání klasifikačních modelů***




Zdroj: vlastní

***Obr. 6: Návrh klasifikačního modelu modelu v Clementine***




Zdroj: vlastní

Boosting může významně zvýšit přesnost modelů vytvořených pomocí algoritmu C5.0 avšak tento process vyžaduje delší čas pro trénování. Tento algoritmus je v textu označen jako $C5.0_{boosting}$ (C5.0-boost) algoritmus.

Následně byla porovnána dosažená přesnost klasifikátorů založených na rozhodovacích stromech s typy klasifikátorů na bázi neuronových sítí. Jde o Radial Basis Function (RBF) neuronovou síť (RBFNS) a pravděpodobnostní neuronovou síť (PNS), z ang. Probabilistic Neural Network, jak je uvedeno na obr. 5.

Oba typy těchto sítí byly již použity k porovnání výsledků klasifikace v [39] pro dvacet datových sad získaných z Machine Learning Repository na Kalifornské Irvine univerzitě. Z výsledků výzkumu vyplynulo, že PNS dosahovala vyšší přesnosti klasifikace než RBFNS a PNS měla průměrnou přesnost klasifikace 76,01 % [39].

## 1.3 Analýza výsledků

Navržený klasifikátor (obr. 5) byl realizován ve dvou prostředích. V prvním případě šlo o Clementine ver.10.0 a realizaci algoritmů „C5.0“, „C&RT“, „CHAID“, „C5.0-boost“ a „RBFNS“ (obr. 6) a ve druhém o MATLAB ver.7.3, Toolbox Neural Nets (algoritmy „RBFNS“ a „PNS“). Na ohodnocení dosažených výsledků byly použity známé metody, uvedené např. v [3,13,16,26,38]. V modelu Clementine (obr. 6) byly použity standardní typové uzly.

Uzly realizují: proces analýzy dat „9 Fields“; přípravy dat (čistění, diskretizace, tvorba binárních atributů) „Kategorizace“; rozdělení datové matice na trénovaní a testovací „Partition“, modelování na bázi rozhodovacích stromů (C5.0, C&RT, CHAID, C5.0-boost) a neuronové sítě (RBFNS) „Modely“ a jejich testování „Analysis“. Výstupy z uzlu „Partition“ reprezentují již výše uvedené algoritmy: „C5.0“ (uzel s označením C5.0 a popiskem x13-kvalitaZP), „C&RT“ (uzel s označením CRT a popiskem x13-kvalitaZP), „CHAID“ (uzel s označením CHAID a popiskem x13-kvalitaZP), „C5.0-boost“ (uzly s označením C5.0 a popisky C5.0-boost-20 a C5.0-boost-10) a „RBFNS“ (uzel s popiskem RBFN_1). Jednotlivé modely pracují s reálnými daty, které jsou tvořeny nezávislými vstupními atributy {a1, a2, ..., a32} a závislým atributem {a33}, tak jak je uvedeno v tab. 2. Rozdělení datové matice na trénovaní a testovací bylo uskutečněno na základě náhodného výběru a potom použito shodně pro všechny algoritmy.

Cílem testování a vyhodnocení je určit, v kolika případech se klasifikátor shoduje s učitelem a v kolika se dopustil chyby. V daném případě lze vyžít matici záměn $C_c$, celkovou přesnost $A_c$ popřípadě celkovou chybu $E_c$ [3,13,38]. Matice záměn $C_c$ sleduje počty správně a nesprávně zařazených příkladů. Celková přesnost $A_c$ patří mezi jednoduché charakteristiky určující, jak jsou nalezené znalosti (pravidla) kvalitní. Obecně je výpočet celkové přesnosti $A_c$ definován následovně:

$$A_c = (SP + SN)/(SP + SN + FP + FN), \qquad (3)$$

kde: $SP$ znamená správně pozitivní, SN znamená správně negativní, $FP$ znamená falešně pozitivní, $FN$ je falešně negativní zařazení.

Z charakteristiky $A_c$ lze získat celkovou chybu $E_c$, která se vypočte jako relativní počet chybných rozhodnutí klasifikátoru:

$$E_c = (FP + FN)/(SP + SN + FP + FN). \qquad (4)$$

Při zjišťování pouze počtu správných či chybných rozhodnutí klasifikátoru je:

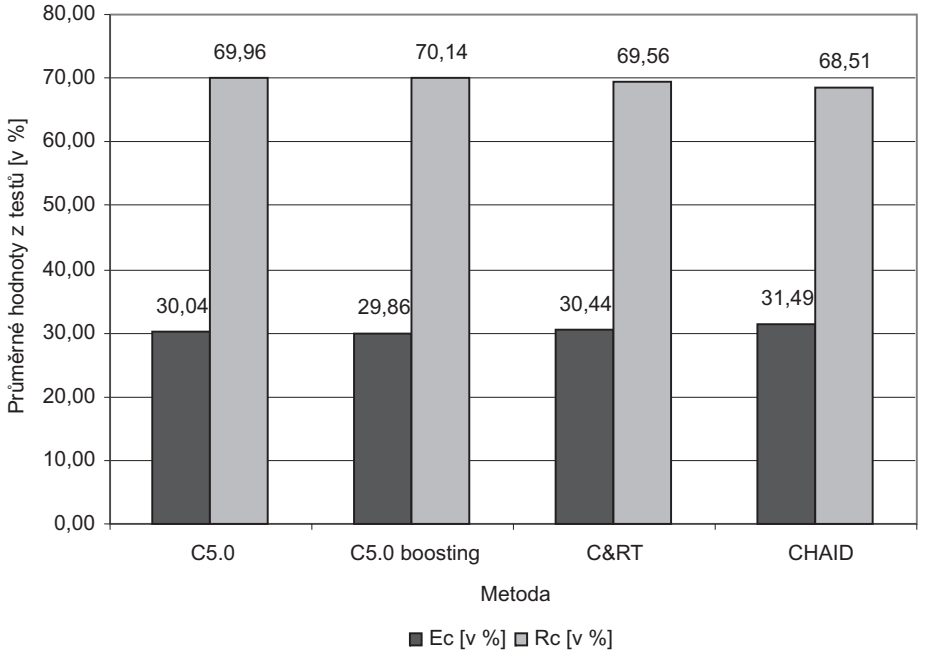
$$E_c = 1 - A_c. \qquad (5)$$

***Tab. 3: Průměrné hodnoty 30 testů algoritmů rozhodovacích stromů – testovací data***

| Algoritmus | Celková chyba $E_c$ [%] | Celková přesnost $A_c$ [%] |
|---|---|---|
| C5.0 | 30,00 | 70,00 |
| $C5.0_{boosting\_10}$ | 30,00 | 70,00 |
| $C5.0_{boosting\_20}$ | 30,00 | 70,00 |
| CHAID | 31,82 | 68,18 |
| C&RT | 32,27 | 67,73 |

Zdroj: vlastní

***Obr. 7: Průměrné hodnoty z 30 testů algoritmů rozhodovacích stromů – trénovací data***




Zdroj: vlastní

V případě, že provádíme testování na trénovacích datech, je možné získat také resubstituční chybu $R_c$ [13,38]. Tento způsob testování má však malou vypovídací schopnost z hlediska nalezených znalostí použitelných pro klasifikování nových případů. Přesto je tento ukazatel v praxi používán i když není vhodným ukazatel správnosti klasifikace [38].

Výsledky dosažené pomocí navrženého klasifikátoru vytvořeného na základě algoritmu C5.0, CHAID, C&RT získané na testovacích datech (33,3 % respondentů z datové matice M) jsou uvedeny v tab. 3. V rámci algoritmu C5.0 byl využit i tzv. boosting - vznikají postupně modely s rostoucí váhou hlasu; každý z modelů v řadě se zaměřuje jen na ty případy, které předcházející modely nedokázaly správně klasifikovat. Testovali jsme tvorbu deseti ($C5.0_{boosting\_10}$) a dvaceti ($C5.0_{boosting\_20}$) hierarchických modelů. Při kombinování modelů se obvykle zvyšuje správnost klasifikace. V našem případě byla ovšem dosažena shodná přesnost klasifikace (70,00 %) pomocí základního klasifikátoru založeného na pravidlech C5.0 i $C5.0_{boosting\_10}$ a $C5.0_{boosting\_20}$.

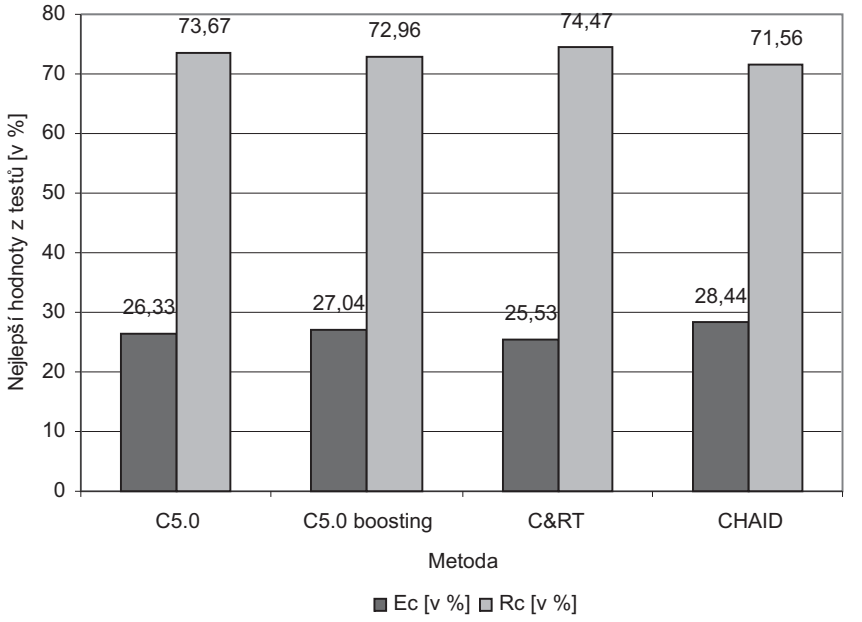
Rozdíl mezi dosaženými výsledky přesnosti klasifikace na testovacích a trénovacích datech nebyl větší než 2,5 %, což nás vede k závěru, že generované rozhodovací stromy nebyly přeučeny.

Výsledky získané z trénovacích datech (v tomto případě hovoříme o resubstituční chybě Rc) uvádí obr. 7 a obr. 8. Na obr. 7 jsou zachyceny průměrné hodnoty Rc a obr. 8 znázorňuje nejlepší dosažené hodnoty klasifikátorů při použití uvedených algoritmů. Lze konstatovat, že algoritmy dosahují podobných výsledků.

Na základě následné analýzy generovaných pravidel uvedenými algoritmy je možné rozdělit tato pravidla do tří skupin. První skupina pravidel je definována algoritmy C5.0, $C5.0_{boosting\_10}$ a $C5.0_{boosting\_20}$ (množina dvanácti pravidel), druhá pomocí C&RT (množina šesti pravidel) a třetí prostřednictvím CHAID (množina šesti pravidel).

U rozhodovacích pravidel, která jsou výstupem z klasifikačních modelů rozhodovacích stromů, lze s do jisté míry konstatovat, že se přibližují realitě. Závislý výstupní atribut (tab. 1) „Spokojenost s kvalitou okolního životního prostředí“ závisí významně na atributu „Úroveň spokojenosti s mož-

***Obr. 8: Nejlepší hodnoty z 30 testů algoritmů rozhodovacích stromů – trénovací data***




Zdroj: vlastní

ností relaxace a odpočinku“ a pouze částečně na „Vyjádření, zda prostředí a životní podmínky ve městě mají vliv na zdraví dotazovaného“. Uvedená skutečnost odpovídá výsledkům studie [6], kde kvalitní/poškozené životního prostředí – kvalita vzduch a vody, hygiena, kvalitní potraviny [8] atd. má na zdraví populace (kvalitu života) vliv pouze 15 -20 %, ale životní styl je ovlivňuje z 50 % [6].

Na základě datové sady z projektu TIMUR pro Chrudim a použití výstupů rozhodovacích stromů ve formě množiny pravidel, byla toto pravidla interpretována pro možné využití managementem na regionální úrovni VS. Klasifikační algoritmy C5.0 a $C5.0_{boosting}$ vytvořily vždy 12 pravidel, C&T a CHAID 6 pravidel. Všechny algoritmy shodně označily za nejdůležitější atribut $x_{17}$ „Úroveň spokojenosti s možností relaxace a odpočinku“, jinými slovy $x_{17}$ má nějvětší vliv na výstupní závislý atribut $x_{13}$ „Spokojenost s kvalitou okolního životního prostředí“ (tab. 1). Uvedený atribut $x_{17}$ na základě tab. 2 odpovídá atributům a29 „x17_relax_vysoka“ respektive a27 „x17_relax_nizka“. Atribut a28 „x17_relax_prumer“ nemá v modelech takovou významnost jako a29 a a27.

Z takto definované množiny pravidel využitím obecného zápisu n-tého pravidla Pn

$$P_n : IF \text{ “předpoklad” } THEN \text{ “závěr”}, \qquad (6)$$

je možné:

- Na základě Occamovy břitvy [29,33] vycházet z jednoduššího vyjádření pravidel, které dává algoritmus C&RT a CHAID, takto:
  *IF* “Úroveň spokojenosti s možností relaxace a odpočinkujevysoká ”*THEN*“ Spokojenost s kvalitou okolního životního prostředíjevysoká”.
  Jinými slovy, spokojenost s kvalitou okolního životního prostředí je přímo úměrná pouze s úrovni spokojenosti s možností relaxace a odpočinku.
  Nedostatkem tohoto pojetí je, že se úplně vytratil atribut $x_{16}$ „Vyjádření, zda prostředí a životní podmínky ve městě mají vliv na zdraví dotazovaného” (tab. 1).
- Najít průnik pravidel u všech algoritmů (C5.0, C5.0 boosting, C&RT and CHAID). V tomto případě můžeme pravidla zapsat následujícím způsobem s využitím tab. 1 a 2:
  *IF* “Úroveň spokojenosti s možností relaxace a odpočinkujevysoká ”*AND*“ Vyjádření, že pro-

středí a životní podmínky ve městě mají vliv na zdraví dotazovaného jeano-velmi "*THEN*" Spokojenost s kvalitou okolního životního prostředí je vysoká" nebo
*IF* "Úroveň spokojenosti s možností relaxace a odpočinkujevysoká "*AND*" Vyjádření, že prostředí a životní podmínky ve městě mají vliv na zdraví dotazovaného jeano "*THEN*" Spokojenost s kvalitou okolního životního prostředí je vysoká"nebo
*IF* "Úroveň spokojenosti s možností relaxace a odpočinkujevysoká "*AND*" Vyjádření, že prostředí a životní podmínky ve městě mají vliv na zdraví dotazovaného jeneurčeno "THEN" Spokojenost s kvalitou okolního životního prostředí je vysoká".
Jinými slovy, spokojenost s kvalitou okolního životního prostředí je přímo úměrná úrovni spokojenosti s možností relaxace a odpočinku a současně vyjádření, že prostředí a životní podmínky ve městě mají vliv na zdraví dotazovaného není hodnoceno jako ne nebo spíše-ne.
S ostatními proměnnými je potřebné pracovat až v případě, že se regionální management v průběhu rozhodovacího procesu zaměřuje na zlepšení spokojenosti konkrétní cílové skupiny, např.: ženy, studující nebo osoby vyznávající zdravý životní styl.

Jak bylo již zmíněno výše, dosažené výsledky navrhnutého klasifikátoru pomocí rozhodovacích stromů (rozhodovacího lesa) jsou porovnány se dvěma typy klasifikátorů na bázi neuronových sítí [7]. V prvním případě jde o PNS a v druhém o RBFNS.

Původní PNS [11,23] byla navržena za účelem řešit především nedostatky v neuronových sítích, jejichž učení bylo založeno na algoritmu zpětného šíření chyby (traditional back-propagation neural network). Je možné ji charakterizovat jako dopřednou (feed-forward), čtyřvrstvou neuronovou síť, která patří do skupiny algoritmů založených na metodě nejbližšího souseda (nearest-neighbor-like) [18] a je vhodná k řešení klasifikačních úloh. Podle [18] je nevhodné využít danou síť pro klasifikační problém, který obsahuje i irelevantní data. Je založena na dobře zjistitelných, prokazatelných statistických principech odvozených z bayesovského rozhodovacího pravidla a neparametrického jádra (kernel) vycházejícího z odhadů funkcí hustoty pravděpodobnosti.

Uvažujme *m* rozměrný vektor vzoru (objektu) **x** v klasifikačním problému. Bayesovské rozhodovací pravidlo implikuje, že **x** patří do třídy k, právě tehdy, když:

$$h_k l_k f_k(x) > h_i l_i f_i(x) \text{ ,pro } \forall\ i \neq k, \qquad (7)$$

kde: $h_i$ a $h_k$ jsou nepodmíněné pravděpodobnosti výskytu vzorů z třídy i a třídy k; $l_i$ a $l_k$ jsou ztrátové funkce asociované s rozhodnutím chybného zařazení **x** do dané třídy; $f_i(x)$ a $f_k(x)$ jsou funkce hustoty pravděpodobnosti pro třídy *i* a *k*.

Nepodmíněné pravděpodobnosti (apriori probability) jsou často známy nebo mohou být přesně odhadovány a ztrátová funkce vyžaduje subjektivní hodnocení. Ztrátové funkce a pravděpodobnosti mohou být v mnoha případech považovány za shodné. Klíčem k použití rozhodovacího pravidla podle (7) je proto odhadnout pravděpodobnost funkcí hustoty pravděpodobnosti vycházejících z trénovacích příkladů (vzorů, výběru).

Pravděpodobnostní neuronová síť se učí aproximovat funkci hustoty pravděpodobnosti z trénovacích vzorů. Přesněji řečeno PNS je interpetována jako funkce, která aproximuje hustotu pravděpodobnosti příslušného výběrového rozdělení. Metoda neparamaterického odhadu známá jako Parzen Window je použita ke konstrukci třídně závislé funkce hustoty pravděpodobnosti na základě bayesovského pravidla.

Jestliže $\mathbf{x}_j$ je *j*-tý trénovací vzor pro *i*-tou třídu, potom Parzenův odhad funkce hustoty pravděpodobnosti [6] pro *i*-tou třídu je:

$$f_i(x) = \left[\frac{1}{(2\pi)^{m/2}\sigma^m n}\right] \sum_{j=1}^{n} \exp\ , \ \frac{(\mathbf{x}-\mathbf{x}_j)^T(\mathbf{x}-\mathbf{x}_j)}{2\sigma^2} \qquad (8)$$

kde: *n* je počet trénovacích vzorů, *m* je rozměr vektoru vzorů a σ je vyhlazovací (smooth) parametr, který je možné upravit experimentálně.

V našem případě jde o *n* řádků a *m* sloupců matice ***M*** (kapitola 2.1, (1)).

Architektura PNS [11,39] zahrnuje 4 vrstvy (vstupní, vzorová (vrstvu vzorů nebo RBF vrstva), sumační a výstupní). První, vstupní vrstva je složena pouze z distribučních neuronů, které zabezpečují propojení vstupů do druhé vrstvy. Počet neuronů je v této vrstvě stejný jako rozměr vektoru vstupních vzorů. Druhá vrstva je nazývána vzorovou, která je složena z takzvaných RBF neuronů rovnajících se počtu trénovacích vzorů. Vstupem ke každému RBF neuronu této vrstvy je vektor

vzdáleností mezi váhovým vektorem **w** a vstupním vektorem **x**, násobeným biasem **b**. Jejich výstupy jsou odvozeny podle:

$$\mathbf{Y}(\mathbf{x}) = \exp(-\mathbf{n}^{-2}),\ \mathbf{n} = b\|w - x\|, \qquad (9)$$

kde: $||\bullet||$ označuje Euklidovskou vzdálenost a každý bias b je ve vzorové vrstvě nastaven takto:

$$\sqrt{-\log(0{,}5)} / Spread. \qquad (10)$$

Parametr *Spread* určuje šíři oblasti ve vstupním prostoru, které odpovídá každý neuron, tzn. že tento parametr ovlivňuje velikost oblasti okolo vstupního vektoru a tomu odpovídající výstup.

Sumační neurony pouze sčítají vstupy ze vzorových neuronů, které korespondují se třídou, ze které byl vybrán trénovací vzor. Počet neuronů v této vrstvě je proto stejný jako počet tříd trénovacích vzorů. Výstupní vrstva se skládá z neuronů, které vytvořily výstupy odpovídající nejvyšší hustotě pravděpodobnosti. Více o PNS je uvedeno v odborné publikaci[11].

Struktura RBFNS je složena ze tří vrstev, a to ze vstupní, skryté (RBF) a výstupní (lineární) vrstvy [4,22,35]. Vstupní vrstva je tvořená *m* prvkovým vektorem $\mathbf{x} = [x_1, x_2, ..., x_m]$. Spojení mezi vstupem a skrytou vrstvou zahrnuje kromě prvků vektoru **x** i prvky matice vstupních vah $\boldsymbol{IW}(m \times p)$, kde *p* je počet neuronů ve skryté vrstvě. Výstup j-tého neuronu ve skryté vrstvě $h_j(\mathbf{x})$ je následující:

$$h_j(\mathbf{x}) = \phi\left(\frac{||\mathbf{x} - \mathbf{iw}_j||}{iw_{0j}}\right), \qquad (11)$$

kde: $||\bullet||$ označuje Euklidovskou vzdálenost mezi vstupním vektorem **x** a *j*-tým řádkem $\mathbf{iw}_j$ matice ***IW*** pro $j = 1, 2, ..., p$; $iw_{0j}$ je bias *j*-tého neuronu ve skryté vrstvě; $\phi$ je přenosová funkce skryté vrstvy.

Výstup RBFNN $f(\mathbf{x})$ je dán lineární kombinací vektoru výstupu skryté vrstvy $\mathbf{h}(\mathbf{x})$ s vektorem vah $\mathbf{w} = [w_1, w_2, ..., w_p]$ podle:

$$f(\mathbf{x}) = \mathbf{w}_0 + \mathbf{w}^T\mathbf{h}(\mathbf{x}) = \mathbf{w}_0 + \sum_{j=1}^{p} w_j \phi\left[\frac{1}{w_{0j}}\sqrt{\sum_{q=1}^{m}(w_{jq} - x_q)^2}\right] \qquad (12)$$

kde: $\mathbf{w}_0$ je bias – vektor výstupní vrstvy.

Funkce $\phi$ je monotónně radiální funkce [22] s parametry - středem *c* a poloměrem *r*. Typickým příkladem jsou Gaussova, multikvadratická a inverzní-multikvadratická RBF [4,22]. Existuje několik přístupů jak optimalizovat parametry *c* a *r* [35,39].

V experimentech byly RBFN použity pro možnost porovnání s dalšími metodami. Pro potřeby tvorby modelu jsme použili funkci [7] „newrb", byla tak vytvořena dvouvrstvá síť, jejíž obecný popis najdeme taktéž v [7] . Jednotlivé neurony jsou přidávány do sítě do té doby než součet čtverců chyb klesne pod určenou úroveň nebo bylo dosaženo maximálního počtu neuronů. Maximální počet neuronů je dán hodnotou n (počet řádků vstupní matice (vstupů do NS). V tomto případě jde o jednoduchý algoritmus a bylo by možné RBFNS optimalizovat, tak jak je uvedeno např. v [5,12,20].

Pro navržené modely klasifikátorů PNS a RBFNN byla data rozdělena na trénovací a testovací množinu v poměru 2/3 a 1/3. U PNS byla testována změna parametru Spread, nejlepších výsledků na testovacích datech bylo dosaženo pro *Spread* = 0,7. Byla dosažena celková přesnost $A_c$ = 66,52 % a celková chyba $E_c$ = 33,48 %. V případě RBFNS jsme testovali změnu parametrů v RBF vrstvě, nejlepších výsledků bylo dosaženo s defaultním nastavením sítě. I přes snahu měnit parametry RBFNS jsou nejlepší hodnoty $A_c$ a $E_c$ horší než u PNS, jako je např. uvedeno v [39] a v [18], s. 1116. Pro komparaci byl také klasifikátor RBFNS (obr. 6, uzel s popiskem RBFNS_1). Nejlepší dosaženou hodnotou $A_c$ = 68,31 %. V tomto případě má struktura RBFNS ve vstupní vrstvě 32 neuronů, ve skryté 20 a ve výstupní 3 neurony.

Problémem netradičního přístupu k návrhu klasifikátoru pomocí NS, kde v procesu přípravy dat byla data diskretizována pomocí disjunktních intervalů, je, že může dojít ke zhoršení klasifikace. Navrhli jsme nový model RBFNS (shodně jako dílčí model RBFNS_1 na obr. 6,) pro reálná data z dotazníkových šetření. Jeho přesnost byla daleko horší než pro již navržený klasifikátor RBFNS ($A_c$ = 68,31 %). Předpokládáme, že uvedené se odvíjí od toho, že reálná data získaná z dotazníkového šetření nemají fyzikální „reálnou" interpretaci.

## Závěr

Mezi základní cíle regionálního managementu patří rozvoj regionu a růst kvality života jeho občanů. Působení na zlepšování kvality života občanů a tím i na zvyšování míry jejich spokojenosti nejen s životním prostředím ve kterém žijí vyžaduje, aby byly k dispozici nástroje, které umožní posoudit

úspěch tohoto působení. Informace o spokojenosti občanů jsou pro regionální management významným podkladem pro rozhodování a sebehodnocení, a proto je třeba spokojenost občanů hodnotit a měřit.

Případová studie, která se zabývala modelováním spokojenosti občanů s kvalitou životního prostředí, byla pojata jako úloha klasifikační, v které bylo cílem zařazení občanů do tříd podle určení jejich spokojenosti s kvalitou okolního životního prostředí (spokojenost nízká, průměrná a vysoká). Byly zde navrženy modely klasifikátorů vytvořené na základě algoritmů C5.0, CHAID, C&RT a $C5.0_{boosting}$. Nejlepších výsledků bylo dosaženo pomocí algoritmu C5.0 (tab. 3) a potvrdil se náš předpoklad, který je postavený na základě našich zkušeností, že algoritmus C5.0 dává nejlepší výsledky v širokém spektru klasifikačních úloh. S ohledem na výsledky ostatních algoritmů rozhodovacích stromů lze říci, že bylo dosaženo srovnatelných výsledků.

Možnost použitelnosti rozhodovacích stromů na klasifikaci spokojenosti respondentů s kvalitou okolního životního prostředí je možné zdůraznit i porovnáním výsledků klasifikace s PNS a RBF. Hodnota $A_c$ je pro PNS ovšem nižší než průměrná přesnost klasifikace, jak uvádí [39] (viz. kapitola 2.2). Problematika RBF a PNS je však velmi rozsáhlá a byla v tomto článku pouze zmíněna a musíme na tomto místě říci, že si do budoucna zaslouží dalšího zkoumání.

Ke zvýšení přesnosti modelů by bylo vhodné využití dalších vstupních atributů (dotazníkových otázek) charakterizujících sledovanou problematiku.

V budoucnosti lze použít data ze Sociologického ústavu Akademie věd České republiky. Použitelnost klasifikačních modelů je v budoucnu možné řešit pomocí metod, které využívají také metody výpočetní inteligence, např. fuzzy logiku [9,32,36] atd.

***Tento příspěvek vznikl za podpory projektu MŽP č. SP/4i2/60/07 „Indikátory pro hodnocení a modelování interakcí mezi životním prostředím, ekonomikou a sociálními souvislostmi“ a projektu GAČR č. 402/08/0849.***

**Literatura**

[1] ABRAHAMS, A. S. *Seeing the woods for the decision trees* [online]. c2003 [cit. 2008-04-09]. Dostupné z: <http://www.computer.org/portal/pages/dsonline/2003_Archives/0301/d/bks_b.html>.

[2] ADAMČÍK, S. *Regionální politika a management regionů, obcí a měst.* 1. vyd. Ostrava: Technická univerzita Ostrava, 2000. ISBN 80-7078-837-0.

[3] BERKA, P. *Dobývání znalostí z databází.* 1. vyd. Praha: Academia, 2003. ISBN 80-200-1062-9.

[4] BROOMHEAD, D., LOWE, D. *Multivariable functional interpolation and adaptive networks. Complex Systems.* 1988, Vol. 2, s. 321-355.

[5] BUHMAN, M. D. *Radial Basis Functions: Theory and Implementations.* Cambridge Monographs on Applied and Computational Mathematics, 2003.

[6] CLARK, N. M. *Population health and the environment. Essays on the Future of Environmental Health Research* [online]. c2007 [cit. 2009-10-13]. Dostupné z: <http://www.ehponline.org/docs/2005/7644/7644.pdf>.

[7] DEMUTH, H., BEALE, M., HAGAN, M. *Neural Network Toolbox for Use with MATLAB.* The MathWorks, 2005.

[8] ENHIS. *Methodical guidelines for a core and extended set of indicators. European Environment and Health Information System* [online]. c2008 [cit. 2009-10-07]. Dostupné z: <http://www.enhis.org/.../file/enhis_Guidelines_indicator_methodology_V3_uneditedVersion.pdf>.

[9] ESPINOSA, J., VANDEWALLE, J., WERTZ, J. *Fuzzy Logic, Identification, and Predictive Control.* London: Springer - Verlag, 2004. 263 s. ISBN 1-85233-828-8.

[10] GUIDICI, P. *Applied Data Mining: Statistical Methods for Business and Industry.* West Sussex: Wiley, 2003. 376 s. ISBN 0-47084679-8.

[11] GUO, J., LIN, Y., SUN, Z. A Novel Metod for Protein Subcellular Localization Based on Boosting and Probabilistic Neural Network. In: *Proc. of the 2nd Asia-Pacific Bioinformatics Conference.* Dunedin, New Zealand: Australan Computer Society, 2004, Vol. 29, 7 s.

[12] HLAVÁČKOVÁ, K., NERUDA, R. Radial basis function network. *Neural Network World,* 1993, roč. 3, č. 1, s. 93-101.

[13] HAN, J., KAMBER, M. *Data Mining: Concepts and Techniques.* San Francisco: Mor-

gan Kaufman Publishers, 2001. 770 s. ISBN 1-55860-901-6.
[14] JALOVECKÁ, M. 5. konference kvality ve veřejné správě – Paříž 2008. In: *Veřejná správa: týdeník vlády České republiky*. 2008, roč. 11, č. 5, s. vii-viii. ISSN 1213-6581.
[15] KAŠPAROVÁ, M., KŘUPKA, J., PÍRKO, J. Modelování spokojenosti občanů ve vztahu k regionálnímu rozvoji a kvalitě života. *Scientific Papers – Series D*, Univerzita Pardubice, roč. 13, 2008, s. 109-120. ISBN 978-80-7395-040-8, ISSN 1211-555X.
[16] MAIMON, O., ROKACH, L. *Decomposition Metodology for Knowledge Discovery and Data Mining.* London: World Scientific Publishing, 2005. 323 s. ISBN 978-981-256-079-7.
[17] MANDYS, J. *Analýza poskytovatelů sociálních služeb ve městě Pardubice.* Realizována zakázka v rámci přípravy návrhu 1. komunitního plánu sociálních a souvisejících služeb v Pardubicích. Pardubice, 2007.
[18] MONTANA, D. J. A Weighted Probabilistic Neural Network. In: MOODY, J. E.: HANSON, S. J., LIPPMANN, R. (Eds.): *Advances in Neural Information Processing Systems 4*, [NIPS Conference, Denver, Colorado, USA, December 2-5, 1991]. San Francisco: Morgan Kaufman Publishers, 1992, s. 1110-1117. ISBN 1-55860-222-4.
[19] MV ČR. *Kvalita ve veřejné správě* [online]. c2005 [cit. 2008-04-09]. Dostupné z: <http://www.mvcr.cz/odbor/moderniz/koncepce/kvalita.html>.
[20] NERUDA, R., KUDOVÁ, P. Learning methods for radial basis function network. *Future Generation Computer Systems*, 2005, roč. 21, č. 7, s. 1131-1142, ISSN 0167-739X.
[21] NOVÁK, J. *Evropské indikátory udržitelného rozvoje v praxi měst České republiky* [online]. 25.09.2006, poslední revize 01.02.2007 [cit. 2007-09-02]. Dostupné z: <http://www.timur.cz/index2.php?option=com_docman&gid=13&task=doc_view&Itemid=38>.
[22] ORR, M. et al. Assessing RBF Networks Using DELVE. *International Journal of Neural Systems*, 2000, Vol. 10, s. 397-415.
[23] POSPÍCHAL, J. *Neurónové siete a umelá inteligencia*. In: KVASNIČKA, V. et al.: Úvod do teórie neurónových sieté. Bratislava: IRIS, 1997. s. 43-56. ISBN 80-88778-30-1.
[24] PŮČEK, M. et al. *Měření spokojenosti v organizacích veřejné správ - soubor příkladů.* [online]. c2005, [cit. 2008-04-09]. 1.vyd. Praha: Ministerstvo vnitra České republiky, úsek veřejné správy, odbor modernizace veřejné správy, 2005. ISBN 80-239-6154-3. Dostupné z: <http://www.mvcr.cz/odbor/moderniz/spokojenost_final.pdf>.
[25] PYLE, D. *Business Modeling and Data Mining.* San Francisco: Morgan Kaufman Publishers, 2003. 693 s. ISBN 1-55860-653-X.
[26] ROKACH, L., MAIMON, O. *Data Mining with Decision Trees. Theory and Applications.* London: World Scientific Publishing, 2008. 244 s. ISBN 978-981-277-171-1.
[27] SHAFRITY, J. M., RUSSELL, E. W., BORICK, Ch. P. *Introducing Public Administration.* New York: Pearson, 2008.
[28] SHINNERS, S. M. *Modern Control System Theory and Design.* 2nd ed. New York: John Wiley and Sons, 1998. ISBN 0-471-24906-8.
[29] SHIRN, M. (ed.) *The philosophy of mathematics today.* New York: Oxfor University Press Inc., 2003.
[30] SPSS. SPSS *Classification Trees* [online]. 05.02.2008 [cit. 2007-11-22]. Dostupné z: <http://www.spss.cz/sw_mcla.htm>.
[31] SUR. *Strategie udržitelného rozvoje České republiky* [online]. 21.12.2004 [cit. 2007-12-06]. Dostupné z: <http://www.env.cz/AIS/web-pub.nsf/$pid/MZPISF7Z6L7V>.
[32] TKÁČ, J., CHOVANEC, A. The Application of Neural Networks for Detection and Identification of Fault Conditions. *Metalurgija*, 2010, Vol. 49, No. 2, pp. 566-569. ISSN 0543-5846.
[33] TREFIL, J. S. *The nature of science: an A-Z guide to the laws and principles governing our.* Boston: Houghton Mifflin, 2003.
[34] TURBAN, E., ARONSON, J.E., LIANG, T. P. *Decision Support Systems and Intelligent Systems.* 7th ed. Upper Saddle River: Pearson Education, 2005. ISBN 0-13-046106-7.
[35] VASSILAKIS, H., HOWELL, A. J., BUTON, H. Comparison of Feedforward (TDRBF) and Generative (TDRGBN) Network for Gesture Based Control. In: WACHSMUTH, I., SOWA,T. (Eds.): *Gesture and Sign Language in Human-Computer Interactions. Lecture Notes in Computer Science.* Berlin Heidelberg: Springer-Verlag, 2002, s. 87-104. ISBN 978-3-540-43678-2.

[36] VAŠČÁK, J., RUTRICH, M. Path Planning in Dynamic Environment Using Fuzzy Cognitive Maps. In: *SAMI - 6th International Symposium on Applied Machine Inteligence and Informatics*, Herľany, Slovakia, pp. 5-9, 2008.
[37] VORLÍČEK, Z. Strategie Národní politiky kvality v České republice na období let 2008 až 2013 pro vyšší kvalitu života občanů České republiky. *Veřejná správa: týdeník vlády České republiky.* 2008, roč. 11, č. 5, s. iii-v. ISSN: 1213-6581.
[38] WITTEN, I.H., FRANK, E. *Data Mining: Practical Machina Learning Tools and Techniques.* San Francisco: Morgan Kaufman Publishers, 2005. 526 s. ISBN 0-12-088407-0.
[39] YOUSEF, R., HINDI, el K. Locating Center Points for Radial Basis Function Networks Using Instance Reduction Techniques. In *Proceedings of World Academy of Science, Engineering and Technology,* 2005, Vol. 4, s. 213-216. ISSN 1307-6884.
[40] ŽELEZNÝ, F., KLÉMA, J., ŠTĚPÁNKOVÁ, O. Strojové učení v dobývání znalostí. In: MAŘÍK V. et al.: *Umělá inteligence 4*. 1. vyd. Praha: Academia, 2003. s. 355 - 406. ISBN 80-200-1044-0.

**doc. Ing. Jiří Křupka, CSc.**
Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémováho inženýrství a informatiky
Studentská 84, 532 10 Pardubice
Jiri.Krupka@upce.cz

**Ing. Miloslava Kašparová, Ph.D.**
Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémováho inženýrství a informatiky
Studentská 84, 532 10 Pardubice
Miloslava.Kasparova@upce.cz

**Ing. Pavel Jirava, Ph.D.**
Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémováho inženýrství a informatiky
Studentská 84, 532 10 Pardubice
Pavel.Jirava@upce.cz

Doručeno redakci: 20. 1. 2009
Recenzováno: 9. 3. 2009, 25. 5. 2009
Schváleno k publikování: 23. 6. 2010

***ABSTRACT***

***QUALITY OF LIFE MODELLING BASED ON DECISION TREES***

**Jiří Křupka, Miloslava Kašparová, Pavel Jirava**

*This paper presents one of the possibilities of the decision theory that can be used in the modelling of the quality of life in a given city in the Czech Republic. Real data sets of citizen questioners for the city of Chrudim were analysed, pre-processed, and used in the model. This model is defined as classification model and it uses algorithms used in decision trees.*

*A decision tree is a predictive model which can be used to represent both classifiers and regression models. In operations research, a decision trees refers to a hierarchical model of decisions and their consequences. When decision tree are used as a classification tasks, it is more appropriately referred to as a classification tree. Classification trees are used to classify an object, or an instance, to a predefined set of classes based on their attribute's value. These trees are frequently used in applied fields such as: finance, marketing, engineering, and medicine. They are useful as an exploratory technique. There are various top-down decision trees inducers such as ID3, C4.5, and C&RT. In our case we used C5.0, C&RT, and CHAID algorithm for the modelling of the quality of life in the previously mentioned city. Similar results were achieved through out our research when using the mentioned algorithms. However, the best results were achieved when using the C5.0 algorithm.*

*Finally, we summarized the presented problems, and compared its accuracy in classification on the basis of decision trees, with classification results based on the probabilistic neural network and the radial basis function neural network.*

***Key Words:*** *Regional management, Quality, Classification, Decision Trees.*

***JEL Classification:*** *C44, C63, L38.*