Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

Ing. Peter Vojtek

PRÍSPEVOK K RELAČNEJ KLASIFIKÁCII S VYUŽITÍM PREDPOKLADU HOMOFÍLIE

Dizertačná práca

Študijný program: Programové systémy
Pracovisko: Ústav informatiky a informačných technológií
Školiteľka: prof. Ing. Mária Bieliková, PhD.
Bratislava, 2010

Autor:
Ing. Peter Vojtek
Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 3
842 16 Bratislava
Slovensko

Školiteľ:
prof. Ing. Mária Bieliková, PhD.

Oponenti:
doc. Ing. Ján Paralič, PhD. (Technická univerzita, Košice)
doc. RNDr. Lubomír Popelínský, PhD. (Masarykova univerzita, Brno)

Kľúčové slová:
klasifikácia, relácie, homofília, sociálne siete
výmena informácií, grafy, relačné učenie
kolektívne usudzovanie

ACM klasifikácia:
H.2.8 [Database Applications] Data mining
G.2.2 [Graph Theory] Graph algorithms
I.2.6 [Artificial Intelligence] Learning
I.5.1 [Pattern Recognition] Models

ABSTRAKT

Práca sa venuje novej paradigme dolovania v dátach, relačnej klasifikácii, ktorá vychádza z atribútovo-viazanej klasifikácie. Relačná klasifikácia zahŕňa skupinu metód, ktoré pri klasifikácii okrem atribútov inštancií zohľadňujú aj vzťahy medzi nimi a vďaka tejto dodatočnej informácií sú schopné zatriedťovať kvalitnejšie. Pre takúto formu údajov sa využíva reprezentácia matematickým grafom. Jednotlivé relačné metódy sa ďalej delia do viacerých podtried podľa toho, ako integrujú relačnú zložku dát do tvorby klasifikačného modelu.

V grafoch, ktoré zachytávajú spoločenské väzby, je prítomný jav, nazývaný homofília. Ide o sociologický jav, ktorý určuje, že vrcholy, ktoré sú spojené reláciou (hranou) sú si navzájom podobnejšie (v zmysle atribútov, triedy) než vrcholy, ktoré hranou prepojené nie sú. Súčasné relačné klasifikačné metódy vo svojej konštrukcii implicitne predpokladajú prítomnosť homofílie (čo označujeme ako predpoklad homofílie), zároveň však tieto klasifikátory nevedia zohľadniť meniacu sa mieru homofílie v grafe v prospech kvality klasifikácie.

Práca sa venuje klasifikácii relačných klasifikátorov a analyzuje dôsledky, ktoré predpoklad homofílie prináša pre jednotlivé podtriedy klasifikátorov. Ďalej definuje homofíliu a spôsoby ako ju merať. Na základe týchto znalostí sú navrhnuté dve nové metódy. Prvá spadá do podtriedy priamych relačných metód a za pomoci lokálneho ohodnocovania grafu mení funkciu susednosti, druhá patrí medzi metódy s kolektívnym usudzovaním a aplikuje moderovanie výmeny informácií. Obe metódy sú schopné na základe výpočtu homofílie priniesť vyššiu kvalitu zatriedenia než doterajšie relačné metódy. Prínos oboch metód v zvýšení kvality klasifikácie je experimentálne overený pomocou rozsiahlych dátových vzoriek. Moderovanie výmeny informácií je použité pri klasifikácii vedeckých publikácií z projektu MAPEKUS. Pri overení klasifikácie za pomoci lokálneho ohodnocovania grafu je použitá dátová vzorka foaf.sk (sociálna sieť Obchodného registra SR).

ABSTRACT

This work is focused on relational classification, an emerging paradigm of data mining based on attribute-based classification. Relational classification is a set of methods which employ relations between instances in a dataset as well as their attributes. Due to this feature relational methods provide higher quality of classification in networked datasets (i.e. data represented via mathematical graph). Relational methods are classified into several branches according to their varying capability to integrate relations into the classifier model.

Homophily is a phenomenon present in graphs capturing real-world data, e.g. social connections between humans. Homophily is defined as following: related (neighbouring) vertices are more likely to share similarities (e.g., the same class, attributes) as non-related instances. Current relational classifiers implicitly require homophily to be present in a graph (so called homophily assumption), however these methods are unable to determine the homophily of each node and take benefit of this information.

This work is at first dedicated to classification of relational classifiers. Next, impact of homophily assumption to particular branches of relational classifiers is analyzes and then homophily measures are defined. Two new relational methods are designed. The first classifier belongs to simple relational methods and employs local graph ranking in order to redefine neighbourhood function, the second method belongs to collective inference branch of methods and applies information interchange moderation between the classified vertices. Both methods are capable to increase the quality of class assignment in networked data due to their capability to employ and measure homophily in a graph.

The contribution was experimentally evaluated using large-scale datasets. In evaluation of information exchange moderation based classifier MAPEKUS dataset was employed. Local graph ranking based classifier was evaluated using foaf.sk dataset (social network of Slovak companies).

Obsah

Kapitola 1

Úvod

Pri riešení mnohých problémov a úloh sa vedome i nevedome využíva klasifikácia. Keď lekár určuje diagnózu pacienta, klasifikuje jeho ochorenie podľa príznakov. Ak sa mladý človek rozhoduje, na ktorú vysokú školu sa prihlási, klasifikuje vzdelávacie inštitúcie podľa toho, čo si o nej prečítal a čo sa dopočul. Rovnako, keď do emailovej schránky príde nová pošta, emailový klient rozhoduje na základe obsahu emailu, či správu označí ako nevyžiadanú, tzv. *spam*. Takýto pohľad na klasifikáciu je atribútovo-orientovaný, triedu klasifikovaného subjektu (inštancie) určujeme na základe jeho čŕt, vlastností, atribútov. Automatická, strojová klasifikácia údajov pomáha rýchlo a efektívne zatriedťovať veľké objemy dát a predstavuje spôsob na organizáciu našich znalostí.

S rozmachom hypertextových dokumentov, obzvlášť s príchodom webu, sa postupne zistilo, že atribútový pohľad na klasifikovanie sveta nie je vždy dostatočný. Klasifikovanie webových stránok sa využíva pri vytváraní katalógov webových stránok, identifikácii jazyka, v ktorom je webová stránka napísaná alebo pri určení portálov, ktoré sú pre používateľov rizikové, lebo vnášajú do webového prehliadača škodlivý softvér. Pri klasifikačných úlohách tohto charakteru sú popri atribútoch zásadné aj explicitné prepojenia – relácie, napríklad hypertextové odkazy medzi webovými stránkami alebo väzby medzi členmi sociálnej siete. Snaha využiť aj takýto druh informácií viedla k vytvoreniu novej paradigmy klasifikácie – k vzniku relačnej klasifikácie a k pohľadu na dáta ako na matematický graf. K

rýchlemu a úspešnému rozvoju relačnej klasifikácie prispieva aj nenáročné a priamočiare získanie relácií zo zdrojových dát a absentujúce problémy s rôznorodosťou a diskretizáciou, ktoré sú bežné pri hodnotách atribútov.

V tejto práci sa venujeme relačným metódam klasifikácie a ich súčinnosti s tzv. predpokladom homofílie, z ktorého tieto metódy vo svojej podstate implicitne vychádzajú. Homofília je jav pôvodne spozorovaný sociológmi a vyskytuje sa najmä v grafoch, ktorých štruktúra[1] je ovplyvnená ľudskou činnosťou a naším nazeraním na svet. Homofília je jav opísaný takto – inštancie, ktoré sú navzájom prepojené (vzťahom) sú si podobné[2] s vyššou pravdepodobnosťou než inštancie, ktoré prepojené nie sú. Najčastejšie tento fenomén pozorujeme v medziľudských vzťahoch, odkiaľ sa prenáša aj do grafov, v ktorých ľudia ako vrcholy grafu priamo nevystupujú. Napríklad, ak spomínané webové stránky klasifikujeme podľa toho, či sú o športe, alebo nie (binárna klasifikácia), relačný prístup predpokladá, že hypertextové odkazy vedúce z webových stránok o športe smerujú na stránky, ktoré sú tiež o športe s vyššou pravdepodobnosťou, než je náhodné rozdelenie.

Schopnosť relačných klasifikátorov správne určiť triedu je založená práve na tom, že predpokladajú homofíliu v dátach. Ak by inštancie boli prepojené s rozdelením nezávislým od triedy klasifikácie (najjednoduchší prípad je rovnomerné náhodné rozdelenie), klasifikátor nedokáže určiť triedu a dosiahne rovnaké výsledky ako náhodný generátor. V tejto práci sa venujeme dôsledkom, ktoré takéto *naivné* očakávanie homofílie v návrhu klasifikačnej metódy prináša.

Existujúce relačné klasifikačné metódy nevedia využiť vo svoj prospech rôznu úroveň homofílie, ktorá je odlišná pre jednotlivé vrcholy, pretože je závislá na vlastnej príslušnosti vrcholu k triede rovnako ako na okolí vrcholu. Dôsledkom sú dva javy, ktoré môžu zapríčiniť zníženie kvality klasifikácie:

- vrcholy medzi sebou zdieľajú informáciu o príslušnosti k triede bez ohľadu na kvalitu tejto informácie,
- spôsob, akým sa získava množina susedných vrcholov nie je dostatočne

[1] Štruktúra grafu – spôsob, akým sa vytvárajú hrany medzi inštanciami.

[2] Podobnosť je v našom prípade príslušnosť k triede.

flexibilný a nezohľadňuje odlišnosti v štruktúre grafu.

Pri uvedomelom zohľadnení homofílie sa obom negatívnym javom dá predísť, teda vieme zamedziť výmene bezcenných a mätúcich informácií medzi vrcholmi počas relačnej klasifikácie a tiež vieme pre každý vrchol získať také susedstvo, ktoré napriek meniacej sa lokálnej základnej štruktúre grafu poskytuje klasifikátoru dostatočnú a hodnotnú množinu okolitých vrcholov. V oboch prípadoch tak vieme predísť zníženiu kvality klasifikácie.

1.1 Ciele práce

Význam javu homofílie v relačnej klasifikácii bol spozorovaný, ale nebol doteraz skúmaný dostatočne do hĺbky. Hlavným cieľom našej práce je:

> Navrhnúť a experimentálne overiť relačnú klasifikačnú metódu, ktorý bude informovane využívať meniacu sa homofíliu v grafe vo svoj prospech. To znamená zamerať sa na vyššiu robustnosť metódy v zmysle zohľadnenia meniacej sa homofílie v dátach, taktiež vytvoriť prístup schopný zohľadniť štruktúru grafu na základe analýzy spôsobov, akými si pri relačnej klasifikácii susediace inštancie zvyčajne vymieňajú informácie.

Uvedený cieľ vyžaduje, aby sme identifikovali významné črty v architektúre, v ktorých sa medzi sebou líšia relačné klasifikačné metódy, vďaka čomu budeme schopní determinovať vhodný postup, v ktorom metóda bude určovať homofíliu a tiež vykonávať samotnú klasifikáciu. V prvom rade teda potrebujeme vytvoriť *klasifikáciu klasifikačných metód*, v ktorej sa zameriame najmä na relačné prístupy a tieto zatriedime do podtried, pričom rozšírime doterajší stav poznania v notácii zápisu významných čŕt jednotlivých prístupov.

Na základe takejto klasifikácie relačných metód sa zameriame na zvolené skupiny metód a v týchto zavedieme analýzu homofílie v klasifikovanom grafe. Tento cieľ vyžaduje podrobne preskúmať doterajšie poznatky o prepojení homofílie a klasifikácie. Na ich základe je potrebné definovať pojem

homofília špecificky pre relačný pohľad na klasifikáciu a vytvorit' metriku pre kvantifikovanie hodnoty homofílie. Našim cieľom je tiež ukázat', že vhodne zvolená miera homofílie predstavuje ekvivalent používaných mier na určenie kvality klasifikátora.

Súvisiacim cieľom je experimentálne overenie novovytvorenej klasifikačnej metódy. Je potrebné identifikovat' vhodné dátové vzorky, v ktorých bude možné rigorózne porovnat' rôzne klasifikačné metódy s našou metódou, v zmysle porovnania kvality výsledku klasifikácie medzi týmito metódami. Našou ambíciou je aj prispiet' do procesu vytvárania dátových vzoriek, vhodných na takéto experimenty a dat' ich k dispozícii komunite.

1.2 Hypotézy práce

Na základe určených cieľov vieme stanovit' hypotézy, ktorých potvrdenie alebo vyvrátenie bude základným prínosom našej práce.

Hypotéza č. 1 znie:

> Zdieľat' pri relačnej klasifikácii menej informácií (z pohľadu prepojených vrcholov) je prospešné pre výslednú kvalitu zatriedenia.

Vrchol je v procese relačnej klasifikácie pričlenený k triedam klasifikácie s meniacou sa pravdepodobnost'ou (príslušnost' vrcholu k triede je neostrá) a táto je rôzne distribuovaná. Na základe tvaru distribúcie vieme získat' informáciu o kvalite príslušnosti vrcholu k triede (ako dobre daný vrchol reprezentuje určitú triedu). Ak vrchol triedu nereprezentuje primeraným spôsobom, teda neposkytuje dostatočne dobre vyhranenú informáciu susediacim vrcholom, tieto vrcholy môžu byt' nevhodne ovplyvnené.

Inšpirujúcou pre nás bola jedna z prvých prác v relačnej klasifikácii [Chakrabarti *et al.*, 1998], kde priveľké množstvo informácií, o ktoré sa obohacovali pôvodné vlastné atribúty inštancie, viedlo k zníženiu kvality klasifikácie. V hypotéze č. 1 predpokladáme, že ak zohľadníme homofíliu ako spôsob ohodnotenia tvaru distribúcie príslušnosti vrcholu ku triede a na základe tohto údaja budeme *moderovat'* výmenu informácií v grafe (budeme

pre vrchol povoľovať alebo zamietať šíriť informáciu o sebe do okolia), zvýšime tým kvalitu klasifikácie.

Hypotéza č. 2 znie:

> Po rozšírení bežnej funkcie susedstva vrcholu tak, aby bola zohľadnená lokálna štruktúra grafu, sa zvýši kvalita klasifikácie, pretože sa zvýši homofília.

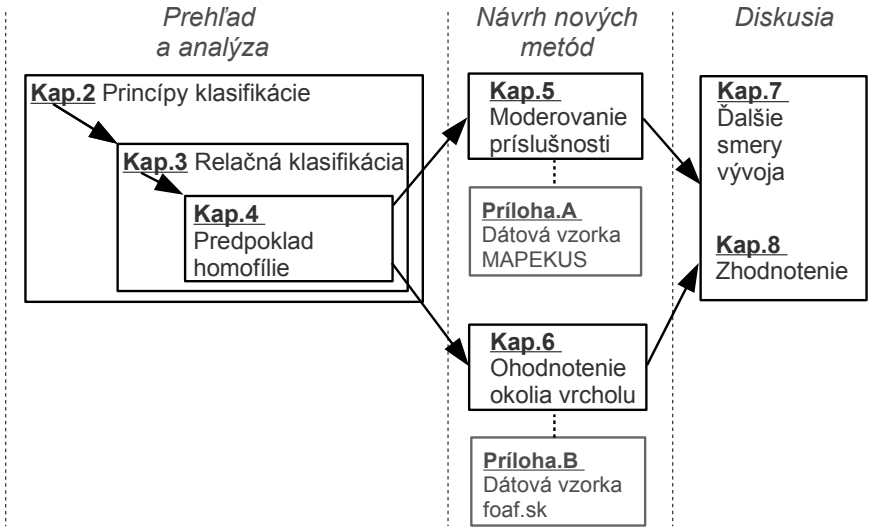
Homofília sa mení tak, ako sa mení náš spôsob nazerania na to, čo je to susedstvo vrcholu v grafe. Doterajšie relačné klasifikačné metódy používajú základnú funkciu susedstva – priame susedstvo, kde pre zvolený vrchol patria do množiny jeho susedov práve tie vrcholy, ktoré sú s ním spojené hranou. Náš predpoklad je, že ak použijeme funkciu susedstva založenú na lokálnom ohodnocovaní grafu, tento prístup zohľadní štruktúru grafu, zvýši homofíliu vrcholu, a tým sa zníži chybová miera klasifikátora.

Obe hypotézy spája zámer overiť, či je zmysluplné hlbšie analyzovať homofíliu grafu počas relačnej klasifikácie.

1.3 Štruktúra práce

V kapitole 2 sa venujeme princípom klasifikácie a analyzujeme ich z pohľadu dolovania v dátach. Kapitola 3 približuje jednotlivé paradigmy klasifikácie a aj samotné metódy ktoré k nim prislúchajú. V kontexte grafovej reprezentácie dát pri klasifikácii sa v kapitole 4 venujeme predpokladu homofílie a jeho dôsledkom na priebeh klasifikácie .

V kapitole 5 uvádzame nami navrhnutú klasifikačnú metódu, ktorá zvyšuje robustnosť relačnej klasifikácie moderovaním výmeny informácií v grafe a potvrdzuje prvú hypotézu. Dátová vzorka použitá pri vyhodnotení tejto metódy je opísaná v prílohe A. V kapitole 6 skúmame druhú hypotézu za pomoci uvedenia nami navrhnutej klasifikačnej metódy, ktorá vychádza z lokálneho ohodnocovania v grafe a ovplyvňuje mieru závislosti relačnej klasifikácie od homofílie. Dátová vzorka použitá pri experimentálnom overení prístupu je analyzovaná v prílohe B.



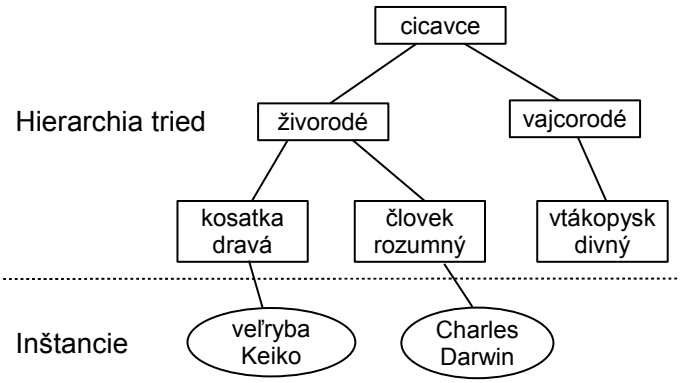
Obrázok 1.1: Logická nadväznosť kapitol práce.

V kapitole 7 diskutujeme o otvorených problémoch a ďalších možných smeroch vývoja relačnej klasifikácie. Obsahom kapitoly 8 je zhrnutie práce. Logická nadväznosť a previazanosť kapitol je na obr. 1.1.

Kapitola 2

Princípy klasifikácie

Klasifikácia (zatrieďovanie, triedenie) je proces priraďovania tried inštanciám. Príkladom je zaraďovanie živých organizmov do skupín na základe biologickej taxonómie (obr. 2.1). Inštanciou tu chápeme konkrétny subjekt z reálneho prostredia, napríklad veľrybu Keiko známu z filmu *Free Willy*.



Obrázok 2.1: Výsek taxonómie živočíchov.

Ďalším prípadom použitia rozsiahlej klasifikácie je medzinárodné desatinné triedenie[1] a medzinárodná klasifikácia ochorení[2]. Uvedené klasifikácie vznikali skôr než prvý elektronický počítač a v danej dobe boli ľudia

[1]Medzinárodné desatinné triedenie `http://www.snk.sk/?MDT`

[2]Medzinárodná klasifikácia chorôb `http://www.who.int/classifications/icd/en/`

tí, ktorí jednotlivé inštancie zatrieďovali na základe expertných znalostí (napr. zoológ, knihovník, lekár).

Nás zaujímajú tie klasifikačné úlohy, ktoré možno riešiť *strojovo*, automatizovane. Rozmach strojovej klasifikácie informácií nastáva v druhej polovici dvadsiateho storočia a spája sa s dolovaním v dátach. Dolovanie v dátach je disciplína, ktorá vychádza zo strojového učenia v rámci umelej inteligencie. Zahŕňa mnohé prístupy na extrakciu a odvodenie znalostí z veľkého množstva dát [Han & Kamber, 2006].

V priebehu uplynulých desaťročí vzniklo mnoho metód vhodných na riešenie rôznych klasifikačných problémov (pozri kap. 3). Medzi klasifikačné úlohy riešené strojovo, s ktorými sa denne stretávame, patria úlohy jednoduchého mechanického rázu i značne sofistikované úlohy vyžadujúce softvérové riešenie a znalosti expertného systému. Ako príklady uvádzame:

- triedenie materiálu, napr. preosievanie piesku rôznej granularity, triediace stroje na ovocie,
- automat na lístky na MHD, ktorý rozpoznáva, aké mince doň boli vhodené,
- rozoznanie typu pivovej fľaše v automate na fľaše,
- výstupná kontrola súčiastok, napr. či vyrobená žiarovka naozaj svieti,
- určenie autora textu, napr. rozpoznanie, či študentská práca nie je plagiátom,
- identifikácia jazyka textu, napr. súčasné textové procesory na základe tejto funkcionality umožňujú automatickú kontrolu pravopisu,
- určenie, či sa na obraze nachádza ľudská tvár[3],
- rozpoznanie rukou písaných písmen, čo je užitočné pri strojovom spracovaní formulárov (daňové priznanie) alebo v zariadeniach s dotykovou obrazovkou,
- porovnávanie odtlačku prsta – autentifikačná metóda dostupná na mnohých súčasných notebookoch,
- spracovanie hovoreného jazyka, napr. rozpoznanie hlások v reči umožňuje strojovo zaznamenať diktované údaje,

[3]Táto funkcionalita je v súčasnosti s obľubou integrovaná do fotoaparátov a prináša automatické ostrenie na tvár(e) na snímke `http://www.nikon.com/about/news/2005/0216_06.htm` [cit. 2009-09-24].

- stanovenie rizika poistnej udalosti pre poistenca, napr. výška povinného poistenia motorového vozidla sa určuje aj podľa doterajších údajov o klientovi ako vek, pohlavie, rodinný stav, nehodovosť držiteľa motorového vozidla,
- filtrovanie nevyžiadanej elektronickej pošty,
- stanovenie diagnózy pacienta na základe príznakov ako tep srdca, krvný tlak, krvný obraz.

Vyčerpávajúci prehľad o tom, čo všetko možno klasifikovať v kognitívnych vedách, prináša práca [Cohen & Lefebvre, 2005]. Nami uvedené príklady spájajú výrazy ako filtrovanie, identifikácia, predikcia, rozpoznanie vzorov. Spája ich tiež atribútovo orientovaný pohľad na inštancie, ktoré sú predmetom zatrieďovania. Automat na lístky rozpoznáva mincu podľa jej tvaru, veľkosti a hmotnosti. To aké mince boli vhodené pred aktuálnou mincou je z hľadiska určenia jej nominálnej hodnoty nepodstatné.

Existujú však klasifikačné úlohy, kde je okolie klasifikovanej inštancie nemenej dôležité ako jej vlastné črty. Takéto úlohy nazývame relačne orientované a patria medzi ne napríklad tieto témy:

- šírenie epidémie – na určitom území je časť populácie infikovaná chorobou, ktorá sa prenáša bežným kontaktom (napr. pandémie chrípky: r.1918 španielska chrípka, r.2003 H5H1 [Yu-Chia *et al.*, 2006]). Na základe siete kontaktov medzi ľuďmi je relačný klasifikačný model schopný predikovať ďalšie šírenie epidémie [Galstyan & Cohen, 2006],
- daňové podvody – spoločnosť, ktorá je usvedčená z daňového podvodu vytvára predpoklad, že s ňou personálne previazané spoločnosti môžu byť rovnako zneužité na daňové úniky. Spoločnosti a ľudia sú v nej prepojení napr. cez reláciu *jeKonateľ*. Klasifikačný model je schopný predikovať riziko daňovej kriminality. Obdobný príklad s burzovými špekuláciami je uvedený v [Neville, 2006],
- webové stránky – klasifikácia webových stránok obohatená o informáciu o hypertextových prepojeniach, prípadne aj o návštevníkoch webových sídiel, ktorí sú charakteristickí svojimi vzormi správania a spôsobmi prehľadávania hyperpriestoru (podrobnejšie je táto doména analyzovaná v úvode kapitoly 3).

V tab. 2.1 uvádzame relačný aj atribútový pohľad na vybrané domény.

Tabuľka 2.1: Príklady klasifikovaných domén.

Inštancia	*Atribúty*	*Triedy*	*Relácie*
minca	hmotnosť priemer	1 euro 50 centov	predošlá minca
webová stránka	text URL	o športe o vede	hypertext návštevníci
klient poisťovne	vek typ vozidla nehodovosť	havaruje nehavaruje	rodina klienta
textový dokument	slová textu distribúcia písmen	anglický jazyk slovenský jazyk	poradie slov

Automatizované zatrieďovanie v relačných úlohách si vyžaduje odlišný prístup ako pri čisto atribútovo-orientovanom zatrieďovaní. Pochopiteľne, v oblasti postihnutej nákazlivou chorobou sa možno pokúsiť identifikovať nakazené osoby len na základe symptómov, ale ak by klasifikátor dokázal využiť aj informáciu, či sa osoba dostala do kontaktu s inými nakazenými a kedy, kvalite výsledku by to určite neuškodilo, skôr naopak. Relačné informácie síce vieme transformovať na atribúty inštancie, napríklad inštancia Osoba má n atribútov zodpovedajúcich každej osobe, s ktorou je prepojená v sociálnej sieti. Takto však získame meniaci sa počet atribútov pre jednotlivé inštancie (napr. *Ján* má šesť priateľov v sieti Facebook, kým *Mária* ich má sto). Problém je v tom, že väčšina atribútových klasifikačných metód vyžaduje zhodný počet atribútov pre všetky inštancie a relačná klasifikácia v tomto ponímaní predstavuje inú paradigmu.

2.1 Významné udalosti v klasifikácii

V tejto časti uvádzame vybrané udalosti v priebehu histórie ľudstva, ktoré ovplyvnili naše znalosti o klasifikácii.

- *4. storočie p.n.l.* Písaná teória klasifikácie začína ešte pred kresťanským letopočtom Aristotelovým dielom Kategórie[4]. Toto dielo sa

[4]Aristotel, Kategórie. Preklad E. M. Edghill
`http://etext.library.adelaide.edu.au/a/aristotle/categories/` [cit. 2009-09-13]

považuje za zásadné pre vznik tzv. klasickej teórie klasifikácie, ktorej najvýznamnejšou črtou je predpoklad disjunkcie tried prislúchajúcich k rovnakej úrovni. Na príklade z obr. 2.3 (na s. 14) tento predpoklad znamená, že ak je inštancia stoličkou, nemôže byť zároveň stolom.

- *18. storočie.* Carl von Linné zavádza do biológie klasifikáciu organizmov – taxonómiu[5]. Reverend T. Bayes študuje štatistiku podmienených udalostí, na základe ktorej neskôr vzniká Bayesov teorém[6]. Ten tvorí podstatu naivnej Bayesovej klasifikačnej metódy [Zhang, 2004].
- *19. storočie.* H. Hollerith zdokonaľuje mechanické spracovanie diernych štítkov, ktoré sú úspešne použité pri sčítaní ľudu (klasifikácii obyvateľstva) koncom 19. storočia[7].
- *60. roky 20. storočia.* S príchodom elektronickej éry sa začalo uvažovať o zautomatizovaní procesu klasifikácie. Jednou z prvých známych publikácií je prelomový príspevok [Maron, 1961], v ktorom sa uvádza metóda automatického indexovania (dobové pomenovanie klasifikácie) založená na Bayesovom teoréme. Na túto prácu nadväzuje [Borko & Bernick, 1963], kde sa uvádza experiment s automatizovanou klasifikáciou dokumentov do zvolených tried na základe indexových slov. Autori výsledky považujú za dostatočné a konštatujú, že automatická klasifikácia dokumentov je možná. Medzi ďalšie články z daného obdobia patrí [Doyle, 1965], kde sa autor venuje otázke, či je strojová klasifikácia vhodnou metódou na štatistickú analýzu textu.
- *70. a 80. roky 20. storočia.* V. Vapnik publikuje klasifikačnú kernelovú metódu Support Vector Machines ([Drucker *et al.*, 1999], [Vapnik, 1982]), ktorá dodnes patrí medzi veľmi obľúbené. R. Quinlan zverejňuje C4.5, rýchla metóda založený na rozhodovacích stromoch [Quinlan, 1993].
- *90. roky 20. storočia.* Jedna z prvých relačných metód. Prístup pracuje s hypertextovými dokumentami a v oblasti relačnej klasifiká-

[5] Zoznam pôvodných publikácií von Linného je dostupný na `http://huntbot.andrew.cmu.edu/HIBD/Departments/Library/LinnaeanDiss.shtml` [cit. 2009-09-16]

[6] Záznam z korešpondencie sa nachádza na `http://www.stat.ucla.edu/history/essay.pdf` [cit. 2009-09-16].

[7] `http://www.rhd.uit.no/census/ft1900e.html` [cit. 2009-09-16].

cie je často citovaný [Chakrabarti *et al.*, 1998].

2.2 Klasifikácia, kategorizácia, konceptualizácia

Doteraz sme hovorili výhradne o pojme *klasifikácia*. V literatúre venovanej tejto oblasti dolovania v dátach sa však často stretávame s používaním pojmov *kategorizácia* a *klasifikácia*, akoby boli synonymá (napr. [Ganti *et al.*, 2008, Chakrabarti *et al.*, 1998, Cai & Hofmann, 2003]).

Pojem kategorizácia je podľa [Jacob, 2004] určený takto:

> Kategorizácia je proces rozdeľovania sveta na skupiny entít, ktorých prvky sú si navzájom podobné.

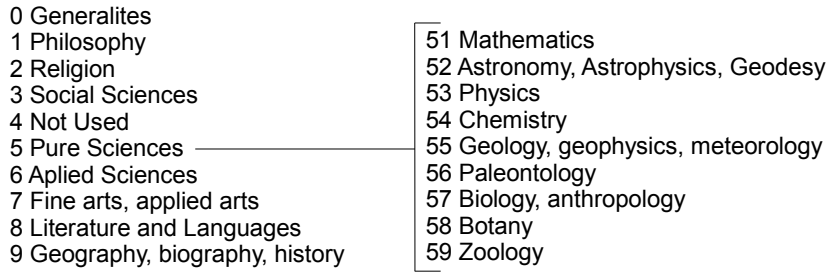
Kategórie vznikajú tak ako prehľadávame a skúmame entity (inštancie) nejakého sveta a atribúty týchto inštancií. Ak sa počet preskúmaných inštancií v doméne zväčšuje, máme čoraz viac informácií o kategóriách, teda sme schopní čoraz presnejšie a formálnejšie vymedziť *priestor* kategórií, čiže vzťahy a rozdiely medzi nimi. Pôvodne hrubo ohraničené kategórie sa postupne stávajú triedami, sú dobre definované a doménovo špecifické. Triedy, ktoré vzniknú, podporujú čoraz presnejšie zachytenie expertnej znalosti a umožňujú zdieľať znalosti o doméne na vysokej formálnej úrovni. Zároveň ale strácajú pôvodnú adaptívnosť v zmysle schopnosti prispôsobiť sa novoobjaveným netradičným inštanciám, ktoré narúšajú existujúcu kategorizáciu/klasifikáciu.

Pojem klasifikácia je podľa [Jacob, 2004] vymedzený takto:

> Klasifikácia ako proces zahŕňa systematické zatriedenie každej entity do práve jednej triedy v rámci systému vzájomne výlučných a neprekrývajúcich sa tried.

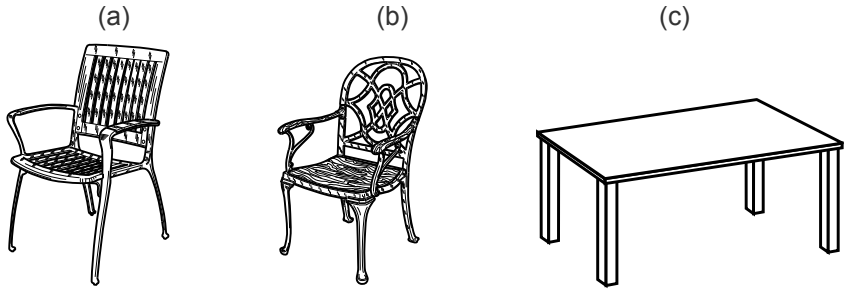
Výlučnosť tried je základom klasickej teórie klasifikácie, založenej na vnímaní sveta tak, ako ho načrtol Aristoteles v knihe Kategórie. Tradičnými príkladmi klasifikácií, ktoré sa vyvíjali dlhé roky a ich vývoj stále nie je ukončený, je napr. vývojová taxonómia živých tvorov (obr. 2.1). Ďalší príklad klasifikácie je už spomínané knižničné Medzinárodné desatinné triedenie

(MDT). Štruktúra MDT je rozširovateľná a pridanie novej triedy do schémy nenarúša existujúce triedy a na rozdiel od vývojovej taxonómie umožňuje priradenie viacerých tried jednej inštancii. Na obr. 2.2 je prvá úroveň a časť druhej úrovne MDT.



Obrázok 2.2: Výsek prvých dvoch úrovní Medzinárodného desatinného triedenia publikácií.

Ak na proces klasifikácie nazeráme zdola nahor, tak v reálnom svete, v ktorom sú len inštancie, tieto zoskupujeme do čoraz abstraktnejších tried a dostávame sa ku konceptualizácii poznania sveta. Schopnosť rozlišovať koncepty (napríklad identifikovať, že inštancie na obr. 2.3(a) a 2.3(b) sú stoličky, kým 2.3(c) je stôl) je pre ľudské bytosti kľúčová a vďaka nej je nám ľuďom umožnené zovšeobecňovať a voliť mieru granularity pri nazeraní na komplexný svet okolo nás. Z pohľadu klasifikácie je pre nás pojem koncept a trieda totožný.

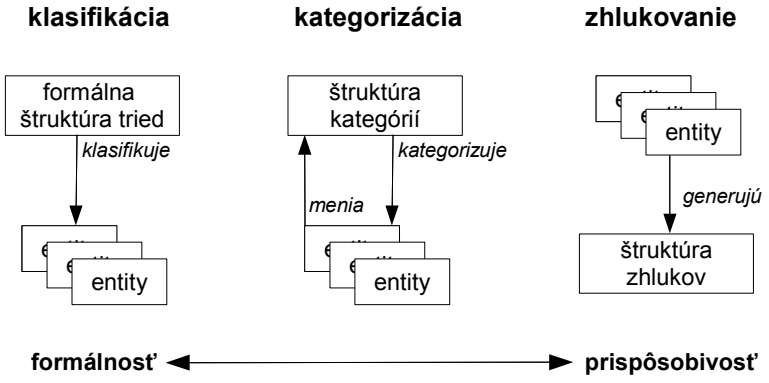


Obrázok 2.3: Tri inštancie zodpovedajúce konceptom *stôl* a *stolička* v oblasti nábytku (ilustrácie prevzaté `http://www.clker.com/`).

Pre doplnenie pojmov vyskytujúcich sa v našej oblasti je vhodné vymedziť pojem *zhlukovanie*. Podľa [Paralič, 2003]:

"Zhlukovanie je identifikácia skupín podobných objektov."

Zhluk zvyčajne nemá pomenovanie, zhluky vznikajú na základe učenia bez učiteľa – ide o deskriptívnu (popisnú) metódu. Rozdiel medzi zhlukovaním, klasifikáciou a kategorizáciu je na obr. 2.4.



Obrázok 2.4: Porovnanie klasifikácie, kategorizácie a zhlukovania.

Špeciálnym pojmom je fazetová klasifikácia. Ide o aplikovanie viacerých klasifikácií nad jedným svetom inštancií. Jednotlivé klasifikácie zod-

povedajú aspektom (dimenziám), v ktorých je pre nás vhodné na svet inštancií nazerať a nazývame ich fazety [Vickery, 2008, Gnoli *et al.*, 2006, Denton, 2003]. Fazetová klasifikácia je obľúbený spôsob navigácie v informačnom priestore. Napríklad ak hľadáme objektívy na fotoaparát, ponúknuté nám môžu byť fazety: ohnisková vzdialenosť, clonové číslo, výrobca, hmotnosť, počet optických elementov.

Z praktických dôvodov sa v tejto práci budeme v nasledujúcich častiach vyjadrovať o prístupe na organizáciu znalostí ako o klasifikácii, aj keď môže ísť o kategorizáciu.

2.3 Klasifikácia ako metóda dolovania v dátach

Pri strojovom spracovaní údajov a dolovaní v dátach chápeme klasifikáciu ako proces systematického zaraďovania inštancií do tried.

Podľa [Preisach & Schmidt-Thieme, 2006] je proces a cieľ klasifikácie špecifikovaný takto:

Majme množinu všetkých inštancií $x \in X$ a atribúty priradené inštanciám $a : X \to A$. Zápis $a(x)$ vyjadruje vektor atribútov pre inštanciu x. Množina X má dve podmnožiny:

- množinu $X_{tr} \subseteq X$ nazývame trénovacou množinou, každá inštancia má priradenú triedu $c : X_{tr} \to C$,
- množinu $X_{tst} \subseteq X$ nazývame testovacou množinou, inštancie majú takisto priradenú triedu $c : X_{tst} \to C$, ktorú využijeme pri vyhodnotení kvality klasifikátora.

Cieľom klasifikácie je vytvoriť atribútovo-viazaný model ($\hat{c}$), ktorý každej inštancii z množiny X_{tst} priradí triedu na základe znalosti o atribútoch získaných z trénovacej množiny X_{tr}. Formálne:

$$\hat{c} : A \to C \tag{2.1}$$

Z praktického hľadiska sa snažíme optimalizovať ($\hat{c}$) tak, aby zodpovedala relácii (c), pre čo najväčšie množstvo inštancií $x \in X_{tst}$. Kvalitu tohto procesu vyjadrujeme chybovou mierou (2.2), ktorá vyjadruje podiel

počtu inštancií, v ktorých sa rozhodnutie klasifikačného modelu nezhoduje s realitou.

$$err_{X_{tst}}(\hat{c} \circ a, c) := \frac{|\{x \in X_{tst} | \hat{c}(a(x)) \neq c(x)\}|}{|X_{tst}|} \tag{2.2}$$

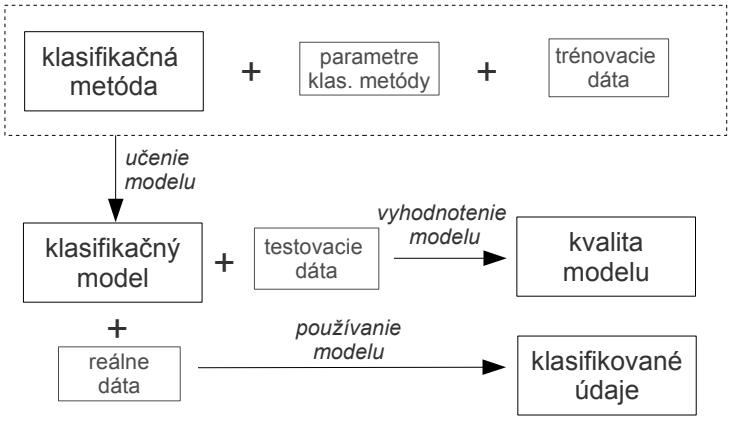
Model klasifikačnej metódy a učenie s učiteľom

Klasifikácia je v rámci dolovania v dátach zaradená medzi prediktívne metódy, kde využívame anotované údaje na predikciu a organizovanie novopríchodzích údajov, ide o tzv. učenie s učiteľom ([Kotsiantis *et al.*, 2006], [Giudici, 2003]). Druhým významným smerom sú deskriptívne metódy, ktoré organizujú údaje do nepomenovaných zoskupení, napr. zhlukovanie.

Pri učení s učiteľom vytvárame (učíme) tzv. klasifikačný model. Tento predstavuje zhmotnenie klasifikačnej metódy[8] nad zvolenou dátovou vzorkou, teda zachytenie (zovšeobecnenie) atribútov trénovacej množiny určitým spôsobom tak, aby čo najlepšie reprezentovali a determinovali triedy prítomné nad inštanciami. To, ktoré charakteristické vlastnosti inštancií trénovacej množiny sú zachytené a ktoré sú ignorované závisí od zvolenej klasifikačnej metódy a nastavenia jej parametrov. Na obr. 2.5 je tento proces znázornený.

Klasifikačný model môžeme používať na zatrieďovanie inštancií. Zvyčajne nás zaujíma, ako presne model predikuje triedu, čo vieme vyhodnotiť pomocou testovacej množiny inštancií, pri ktorých objektívne vieme ich príslušnosť k triede.

[8] Výraz *klasifikátor* predstavuje synonymum pojmu *klasifikačný model*.



Obrázok 2.5: Zvyčajný priebeh klasifikácie ako učenia s učiteľom.

Rozhodovacia tabuľka

Proces klasifikácie (2.1) zodpovedá napĺňaniu tzv. rozhodovacej tabuľky (tab. 2.2) [Sebastiani, 2002], kde sa prirad'ujú hodnoty 0 alebo 1 každému vstupu v tabuľke.

Tabuľka 2.2: Rozhodovacia tabuľka.

	x_1	...	x_j	...	x_m
c_1	a_{11}	...	a_{1j}	...	a_{1n}
...	...	...	...	...	...
c_i	a_{i1}	...	a_{ij}	...	a_{in}
...	...	...	...	...	...
c_m	a_{m1}	...	a_{mj}	...	a_{mn}

$C = \{c_1, \ldots, c_m\}$ je množina preddefinovaných tried a $X = \{x_1, \ldots, x_m\}$ je množina inštancií, ktoré treba klasifikovat'. Hodnota 1 pre a_{ij} znamená, že x_j patrí do triedy c_i, hodnota 0 znamená, že do tejto triedy nepatrí. Tradičnej, aristotelovsky chápanej klasifikácii tu zodpovedá predpoklad, že súčet v stĺpci nemôže presiahnut' hodnotu 1.

Pri zohľadnení trénovacej a testovacej množiny X_{tr} a X_{tst} sa klasifikátor vyhodnotí pomocou tzv. správnej rozhodovacej tabuľky (tab. 2.3).

Tabuľka 2.3: Správna rozhodovacia tabuľka.

	Trénovacia množina			Testovacia množina		
	x_1	...	x_g	x_{g+1}	...	x_s
c_1	ca_{11}	...	ca_{1g}	$ca_{1(g+1)}$	...	ca_{1s}
...	...	...	...	...	...	...
c_i	ca_{i1}	...	ca_{ig}	$ca_{i(g+1)}$	...	ca_{is}
...	...	...	...	...	...	...
c_m	ca_{m1}	...	ca_{mg}	$ca_{m(g+1)}$	...	ca_{ms}

Hodnota 1 pre ca_{ij} v tab. 2.3 znamená, že inštancia x_j bola objektívne (expertom) zaradená do triedy c_i (pozitívny príklad), hodnota 0 znamená, že do tejto triedy nepatrí (negatívny príklad).

2.4 Vyhodnocovanie klasifikácie

Ak vieme, ako inštancie objektívne prislúchajú triedam a máme výsledok, ako boli zatriedené klasifikátorom (teda porovnanie rozhodovacích tabuliek generovaných c a $\hat{c}$), na ich základe vieme naplniť tzv. *tabuľky podmienenosti* (angl. contingency table) [Lewis, 1991].

Klasifikácia sa skladá z n binárnych rozhodnutí a každé má práve jednu správnu odpoveď – *Áno* (1) alebo *Nie* (0). Výsledok takýchto n rozhodnutí podľa tab. 2.4 vyjadruje počet rozhodnutí daného typu. Napríklad a (v tab. 2.4) je počet prípadov, kedy sa klasifikátor rozhodol pre *Áno* a bolo to správne rozhodnutie.

Tabuľka 2.4: Tabuľka podmienenosti pre množinu binárnych rozhodnutí.

	Áno je správne	*Nie* je správne	
Rozhodnutie *Áno*	a	b	$a+b$
Rozhodnutie *Nie*	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d=n$

Z tab. 2.4 vyplývajú tieto miery efektívnosti klasifikácie[9]:

$$Recall = \frac{a}{(a+c)} .100\% \tag{2.3}$$

$$Precision = \frac{a}{(a+b)} .100\% \tag{2.4}$$

$$Fallout = \frac{b}{(b+d)} .100\% \tag{2.5}$$

$$Accuracy = \frac{a+c}{(a+b+c+d)} .100\% \tag{2.6}$$

$$F_1 = \frac{2a}{(2a+b+c)} .100\% \tag{2.7}$$

Recall (úplnosť, návratnosť) určuje percento zo všetkých relevantných dokumentov, pri ktorých sa klasifikačný model rozhodol správne. *Precision* (presnosť) určuje percento dokumentov, ktoré boli zaradené správne voči počtu dokumentov, ktoré klasifikačný model označil za správne. *Fallout* je percento všetkých nerelevantných získaných dokumentov a *Accuracy* (správnosť) určuje podiel správne posúdených dokumentov.

Chybová miera uvedená v (2.2) má s mierou *Accuracy* tento vzťah:

$$err_{X_{tst}}(\hat{c} \circ a, c) = 1.0 - Accuracy \tag{2.8}$$

F_1 kombinuje *Precision* a *Recall* pomocou harmonického priemeru [Yang & Liu, 1999]. F_1 je špeciálny prípad F_α miery, definovanej takto:

$$F_\alpha = (1+\beta^2)\frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{2.9}$$

Voľba miery pre vyhodnotenie kvality klasifikátora závisí od konkrétnej situácie. Môžeme tiež sledovať závislosti zvolených mier pre zvolený klasifikátor, najčastejšie priebeh závislosti *Precision* a *Recall*, nazývaný *Area under curve* [Davis & Goadrich, 2006].

Ak klasifikácia obsahuje viac ako dve triedy, tabuľku podmienenosti

[9]Napriek existujúcim prekladom týchto mier do slovenského jazyka uvádzame kvôli prehľadnosti ich anglický názov.

napĺňame s ohľadom na zvolenú triedu (rozhodnutie *Áno*) voči všetkým ostatným triedam (rozhodnutie *Nie*). Ak takto vytvoríme rozhodovacie tabuľky pre všetky triedy a výsledky zvolenej miery odvodenej z tabuliek spriemerujeme, získame makropriemer (angl. macro-averaging). Ak naopak postupne hodnotami napĺňame iba jednu tabuľku podmieneností a zvolenú mieru počítame z nej, získame mikropriemer (angl. micro-averaging) [Sebastiani, 2002].

2.5 Zvýšenie relevancie vyhodnotenia

Pri určení kvality klasifikačného modelu sledujeme miery efektívnosti modelu. Pre zvýšenie relevantnosti tohto výsledku je zmysluplné *pomôct'* procesu vytvárania trénovacej a testovacej množiny tak, aby sa zredukoval bias pseudonáhodného výberu inštancií, najmä ak je množina všetkých inštancií príliš malá.

2.5.1 Krížová validácia

Často používanou metódou na štatistické zvýšenie presnosti vyhodnocovania je krížová validácia (angl. cross-validation) [Isaksson *et al.*, 2008] a [Kohavi, 1995]. Pri *k-fold* krížovej validácii je dátová vzorka X rozdelená na k neprekrývajúcich sa podmnožín $X_1, X_2, \ldots, X_k$ približne rovnakej veľkosti. Pre každé $X_j, j \in \{1, 2, \ldots, k\}$, zoberieme $X_j = X_{tst}$ ako testovaciu množinu a $X \backslash X_j = X_{tr}$ ako trénovaciu množinu. Výhodou je postupné použitie všetkých inštancií aj v trénovacej aj testovacej fáze klasifikácie.

Špeciálnym prípadom *k-fold* krížovej validácie je *leave-one-out* krížová validácia. Platí, že veľkost' k je zhodná s počtom inštancií, $k = |X|$, čo znamená, že trénovacia množina obsahuje všetky inštancie okrem jednej, ktorú používame na testovanie. Toto riešenie je vhodné pri veľmi malých množinách inštancií [Han & Kamber, 2006].

Ďalší spôsob výberu inštancií, *bootstrapping*, vznikol s cieľom lepšie aproximovat' výber v reálnom svete [Efron & Tibshirani, 1995]. Výber inštancií do množiny X_{tr} z X prebieha tak, že $|X|$ krát sa z množiny X vyberie inštancia. Výber sa deje s opakovaním, takže pravdepodobnost', že inštancia v_i nebude v množine X_{tr} po dokončení výberu [Kohavi, 1995] je takáto:

$$\forall v_i \in X : p(v_i \notin X_{tr}) = \left(1 - \frac{1}{|X|}\right)^{|X|} \approx e^{-1} \approx 0.368 \qquad (2.10)$$

2.5.2 Vrstvenie

Náhodné rozdelenie dát môže spôsobiť, že niektorá trieda nebude zastúpená v trénovacej množine a klasifikátor ju nebude schopný rozoznať. Pri vrstvení (angl. stratification) [Keller, 2001, Witten & Frank, 1999] je cieľom zachovať v trénovacej a testovacej množine distribúciu inštancií tried zhodnú s distribúciou v celej dátovej vzorke.

2.5.3 Validačná množina

Ak využívame klasifikačnú metódu s meniteľnými parametrami (napr. veľkosť vektora slov pri klasifikácii dokumentov a rôzna váha zložiek vektora, ale šírka okolia k pri kNN metóde), snažíme sa nastaviť parametre tak, aby klasifikačný m čo najlepšie zovšeobecňoval inštancie trénovacej množiny. Ak je parametrom veľkosť vektora slov, vytvára sa pre každú zvolenú hodnotu parametra model na základe trénovacej množiny. Určenie váh zložiek vektora sa potom deje za pomoci *validačnej množiny* [Gutierrez-Osuna, 2001], na základe ktorej sa zvolí najvhodnejší model a ten sa vyhodnotí pomocou testovacej množiny. Tieto tri množiny sú disjunktné – ak by sme vyhodnocovali konečnú kvalitu modelu na validačnej množine a nie na testovacej, vniesli by sme do výsledku bias.

Algoritmus trénovania a vyhodnocovania klasifikácie je pri použití validačnej množiny takýto [Gutierrez-Osuna, 2001]:

1. dátová vzorka X sa rozdelí na X_{tr}, X_{val} a X_{tst}, platí $X_{tr} \cap X_{val} \cap X_{tst} = \emptyset$,
2. zvolí sa n verzií klasifikačnej metódy,
3. pre každú verziu (ktorej zodpovedá model) sa natrénujú na základe X_{tr} modely $\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_n$,
4. každý model $\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_n$ sa vyhodnotí pomocou X_{val},
5. zvolia sa parametre najlepšieho modelu z predošlého kroku a natrénuje sa model $\hat{c}_{best}$ pomocou množiny $X_{tr} \cup X_{val}$,
6. kvalita klasifikácie modelu $\hat{c}_{best}$ sa vyhodnotí pomocou X_{tst}.

2.5.4 Predikčná sila klasifikátora

Ak klasifikátor určí pre inštanciu triedu, je vhodné vedieť, nakoľko jednoznačné je toto rozhodnutie a o koľko nižšiu pravdepodobnosť má druhá trieda v poradí. Napríklad, ak priradenie druhej triedy v poradí je len o málo pravdepodobnejšie než výsledok víťaznej triedy, môžeme zvažovať, či radšej inštancii nepriradiť žiadnu triedu, pretože riziko omylu klasifikátora je vysoké. Riešením je vypočítať pre každú zatriedenú inštanciu predikčnú silu modelu [Keller *et al.*, 2000] a ak táto hodnota nepresiahne zvolený prah, inštancii triedu nepriradíme. Výpočet predikčnej sily možno získať viacerými spôsobmi, napr. cez rozdiel logaritmov pravdepodobností víťaznej (c_{win}) a druhej víťaznej triedy (c_{2nd}):

$$trieda(v_i) = c_{win} \Leftrightarrow \log p(c_{win}|v_i) - \log p(c_{2nd}|v_i) > T \tag{2.11}$$

kde T je stanovený prah predikčnej sily.

V tejto kapitole sme sa venovali klasifikácii zo všeobecného hľadiska. Definovali sme klasifikáciu a vymedzili sme tento pojem voči príbuzným oblastiam a pojmom v dolovaní v dátach. Uviedli sme, ako vyhodnotiť výsledky zatriedenia a aké sú prístupy vedúce k zvýšeniu relevantnosti a nezávislosti výsledkov.

Pohybovali sme sa na úrovni aplikovateľnej prakticky na všetky klasifikačné metódy, čisto atribútové i relačné (aj keď definícia procesu klasifikácie v časti 2.3 je určená len pre atribútovo-viazané metódy). Samotné metódy sme však zatiaľ vôbec neuvádzali, ich prehľad prináša nasledujúca kapitola.

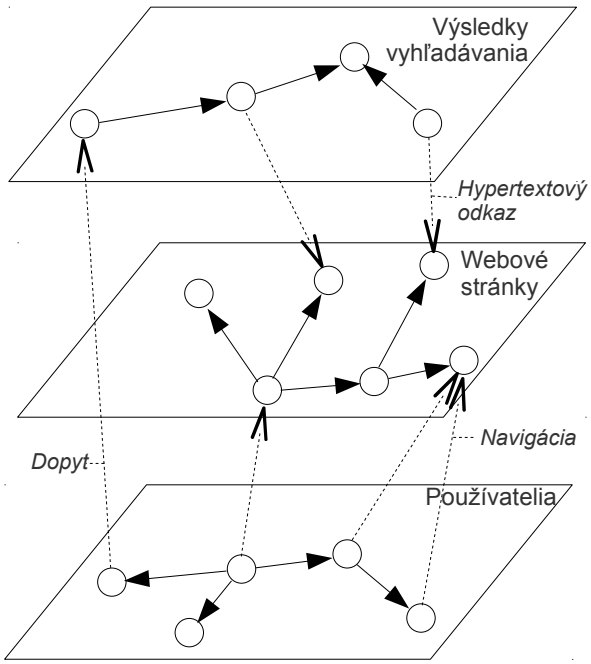
Kapitola 3

Relačná klasifikácia

S rozmachom hypertextu a zvlášť s príchodom webových technológií sa začali objavovať klasifikačné úlohy (napr. zatrieďovanie webových stránok do klasifikačnej hierarchie), v ktorých sa výraznejšie ukázalo, že dátová vzorka obsahuje aj údaje, ktoré doteraz známe metódy nevyužívali. Webové stránky možno zatrieďovať len na základe ich obsahu (html značky, text či obrázky), ale vyvstala otázka, či nie je možné vytvoriť metódu, ktorá vie zohľadniť napr. hypertextové odkazy, ktoré odkazujú na iné klasifikované webové stránky (t.j. inštancie) [Chakrabarti *et al.*, 1998] a vďaka tomu poskytnúť lepšie výsledky.

Doména webových stránok je ešte bohatšia, obr. 3.1 ukazuje rôznorodosť relácií medzi troma typmi inštancií predstavujúcich zvyčajný scenár navigácie používateľov na webe, kedy títo najprv zadávajú do vyhľadávača kľúčové slová, na základe ktorých získajú zoznam relevantných odkazov, ktoré potom navštívia. Vidíme tu tri druhy inter-relácií (*hypertextový odkaz*, *navigácia* a *dopyt* a tri druhy nepomenovaných intra-relácií (*výsledok vyhľadávania*→*výsledok vyhľadávania*, *webová stránka*→*webová stránka* a *používateľ*→*používateľ*).

Pre uvedené klasifikačné úlohy začali vznikať relačné klasifikačné metódy priamo zohľadňujúce vzťahy medzi klasifikovanými inštanciami. Keďže metódy sa od seba zásadne odlišovali, v prácach [Macskassy & Provost, 2007, Jensen *et al.*, 2004] bol zavedený prístup ku *klasifikácii* klasifikačných metód na základe ich schopnosti abstrahovať relačnú zložku.



Obrázok 3.1: Rôzne typy inštancií a vzťahov medzi nimi na webe (prevzaté z [Xue *et al.*, 2006]).

Pre relačnú klasifikáciu sú charakteristické tieto črty[1]:

- prepojené dáta – medzi inštanciami sú prítomné explicitné väzby, zhmotňujúce relácie. Vo väčšine dát sú prítomné skryté vzťahy, napr. medzi webovými stránkami by sme vedeli vytvoriť reláciu na základe podobnosti farby pozadia. Takéto implicitné väzby vygenerované na základe podobnosti však vieme vytvoriť takmer pre akýkoľvek

[1]Pri vytváraní zoznamu sme sa zčasti inšpirovali prácou[Macskassy & Provost, 2007].

atribút a neraz bez pridania informačnej hodnoty. Pre účely relačnej klasifikácie sú preto významné relácie, ktoré sú v dátovej vzorke prítomné explicitne.

- klasifikácia v rámci prepojení (angl. within-network classification) – inštancie, ktorých trieda je v rámci klasifikačnej úlohy známa, sú prepojené s inštanciami, ktorých triedu zatiaľ nevieme. Učenie relačnej časti klasifikačného modelu je založené na prenose informácie o atribútoch, či triede inštancie k susedom cez tieto prepojenia.
- inštančne-orientovaný prístup (angl. node-centric) – metódy pracujú s jednotlivými inštanciami (ich atribútmi a okolím) postupne v čase. Na rozdiel od kernelových metód sa teda nevytvára n-rozmerný priestor v ktorom by boli inštancie rozdeľované do tried generovaním rezov priestoru.

3.1 Používaná notácia a označenie

Na uvedenom príklade relačnej domény návštevníkov webového sídla (obr. 3.1) sme zaviedli viacero pojmov, ktoré budeme využívať v nasledujúcich častiach:

- *inštancia*[2] je dobre ucelená informačná jednotka, napr. používateľ `Ján` je inštancia,
- inštancie prislúchajú typom, napr. `www.fiit.sk` je inštancia typu `webová stránka`, `Ján` a `Mária` sú inštancie typu `používateľ`,
- typ inštancie má priradené *atribúty*[3], napr. typ `používateľ` má atribút `vek`,
- inštancia nadobúda pre každý atribút svojho typu jeho hodnotu, napr. používateľ `Ján` má atribút `vek` naplnený hodnotou `25 rokov`,
- typy inštancií môžu byť prepojené cez *relácie*, napr. typ inštancie `používateľ` je spojený s typom inštancie `webová stránka` cez reláciu `bolaNavštívená`. Relácie rozdeľujeme takto:

[2] V literatúre sa možno stretnúť s ekvivalentnými pojmami *entita* a *objekt*. Pri grafovej reprezentácii dát hovoríme aj o *vrchole* (angl. node, vertex).

[3] Tiež *vlastnosti* (angl. features).

  - inter-relácia je vzt'ah medzi inštanciami rovnakého typu,
  - intra-relácia je vzt'ah medzi inštanciami rôzneho typu,
- relácie sú zhmotnené *vzt'ahmi* medzi inštanciami, ak využívame grafovú reprezentáciu dát, hovoríme o *hranách*. Napr. inštancie `Mária` a `www.fiit.sk` sú spojené hranou spadajúcou do relácie `navigácia`,
- ak uvažujeme o grafovej reprezentácii dát, *hrana* môže byt' *váhovaná* alebo *neváhovaná*. Váha hrany v sebe zvyčajne nesie takýto význam [Preisach & Schmidt-Thieme, 2006]:
  - váha zodpovedá korelácii medzi prepojenými inštanciami,
  - váha je úmerná počtu inštancií, ktoré majú spoločné hranou prepojené inštancie.

Uvedené pojmy sumarizuje tab. 3.1.

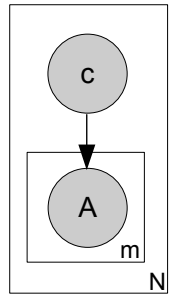
Tabuľka 3.1: Abstraktná a inštančná úroveň používaných pojmov.

Abstraktná úroveň	*Inštančná úroveň*	*Grafová reprezentácia inštancií*
typ inštancie	inštancia	vrchol
atribút	hodnota atribútu	–
relácia	vzt'ah	hrana

3.2 Klasifikácia klasifikačných metód

Z hľadiska porovnávania prístupov ku klasifikácii je významný článok [Jensen *et al.*, 2004], ktorý zavádza grafickú notáciu pre odlíšenie rôznych tried klasifikačných metód podľa spôsobu, akým pristupujú k informáciám nachádzajúcim sa v okolí klasifikovaných inštancií. Ide o prístup orientovaný na vrcholy grafu (angl. node-centric).

Na obr. 3.2 je uvedená notácia pre všeobecný atribútovo-viazaný prístup. Každá z N inštancií má m vlastných atribútov (atribút je A v šedom kruhu), pričom príslušnost' inštancie k triede (označené ako c v šedom kruhu) závisí iba od jednotlivých vlastných atribútov (šípka vedúca z c do A). Tento zápis zodpovedá atribútovo-viazanému modelu, ktorý je vyjadrený pomocou Bayesovej siete [Friedman *et al.*, 1997] a nesie v sebe predpoklad nezávislosti atribútov.



Obrázok 3.2: Jensenova schéma atribútovo-viazanej klasifikácie.

V nasledujúcich častiach sa venujeme trom triedam metód:
atribútovo-viazané prístupy, relačné metódy a prístupy využívajúce kolektívne usudzovanie.

3.2.1 Atribútovo-viazané metódy

Do tejto skupiny patria všetky prístupy, ktoré na základe atribútov inštancie a modelu predikujú triedu inštancie. Jensenov diagram pre tento prístup je na obr. 3.2. Uvádzame niekoľko charakteristických atribútovo-viazaných metód:

- *k-najbližších susedov* (kNN) [Liu, 2006]: metóda využíva rozmiestnenie inštancií v n-rozmernom priestore, každá dimenzia zodpovedá jednému atribútu. Klasifikovaná inštancia svoju triedu určuje na základe k-najbližších inštancií so známou triedou, vzájomná vzdialenosť sa v stavovom priestore zvyčajne vypočíta pomocou euklidovskej vzdialenosti. Základný kNN prístup je zaujímavý tým, že nemá trénovaciu fázu – vytvorenie modelu je tu reprezentované rozmiestnením inštancií do stavového priestoru.
- *rozhodovacie stromy* [Ling *et al.*, 2004]: na základe zvolenej miery korelácie sa zoradia atribúty podľa toho, ako výrazne rozdeľujú svet inštancií medzi triedy (napr. miera informačný zisk (angl. information gain [Yang & Pedersen, 1997])). Potom sa zostrojí rozhodovací strom, v ktorom vrcholy predstavujú atribúty, zoradené od koreňa k listom na základe ich schopnosti determinovať triedu.

- *Support Vector Machines* (SVM) [Drucker *et al.*, 1999]: metóda na základe atribútov inštancií vytvorí ich stavový priestor, v ktorom vytvára (viacrozmerné) roviny, ktoré čo najlepšie oddeľujú inštancie rôznych tried. Trieda neklasifikovanej inštancie sa potom určí na základe toho, do ktorého podpriestoru patrí.
- *neurónové siete* [Han & Kamber, 2006]: široká trieda metód na rozoznávanie vzorov, zložitejší a zdĺhavejší priebeh učenia vyvažuje schopnosť rozsiahlejšieho zovšeobecnenia naučených vzorov.
- *metóda na základe Bayesovho teorému* [McCallum & Nigam, 1998]: vypočítava podmienené pravdepodobnosti vyjadrujúce závislosť jednotlivých atribútov od tried. Pomocou predpokladu nezávislosti atribútov potom umožňuje vyčísliť závislosť tried od každého atribútu a sčítaním týchto závislostí získame distribúciu príslušnosti k triedam pre inštanciu reprezentovanú atribútmi.
- *Bayesove siete* [Pearl, 1998]: klasifikačné metódy založené na Bayesových sieťach [Sacha, 1999] umožňujú modelovať podmienené závislosti medzi atribútmi inštancie (zovšeobecňujú prístup založený na Bayesovom teoréme, ktorý predpokladá nezávislosť atribútov).

Pre atribútovo-viazané metódy je charakteristické, že predpokladajú nezávislosť atribútov, napríklad že vek osoby nekoreluje s jej farbou vlasov. Nazeranie na inštanciu je *ploché*, inštancia je determinovaná svojimi atribútmi a priamo nie je vytvorená informácia o tom, aké iné inštancie sa v danom svete nachádzajú a či je medzi nimi vzťah.

Obdobná situácia je aj pri kernelových metódach (SVM) [Li *et al.*, 2007] a metódach založených na blízkosti susedstva (kNN) [Kwon & Lee, 2000]. Na prvý pohľad sa môže zdať, že ide o relačné metódy, pretože využívajú napr. výpočet vzdialenosti medzi inštanciami (kNN). Samotný stavový priestor, v ktorom sa inštancie nachádzajú je však ohraničený v dimenziách, ktoré sú generované hodnotami atribútov a jednotlivé inštancie sú v tomto smere o svojom susedstve rovnako *neinformované* ako pri iných atribútovo-viazaných metódach (napr. rozhodovacie stromy [Ling *et al.*, 2004]). Vzťahy medzi inštanciami sú implicitné, vypočítané na základe podobnosti atribútov a explicitné vzťahy (relácie), ak vôbec sú v dátovej vzorke zachytené, klasifikátor nevie analyzovať a využiť.

Klasifikácia textu

Uvedené metódy sú vhodné pre domény, v ktorých možno všetky inštancie reprezentovať rovnakým počtom atribútov, napr. inštancie typu *Osoba* s atribútmi *Meno* a *Vek*. Situácia je o niečo zložitejšia v doménach, kde má každá inštancia rozdielny počet základných atribútov. Zvyčajne ide o klasifikáciu textových dokumentov. Obvykle sa text reprezentuje pomocou vektora slov (angl. bag of words) a tento sa prípadne ešte váhuje pomocou tf-idf [Salton & Buckley, 1987]. Keď je počiatočný vektor slov priveľký (dokumentov je veľa a slov, ktoré sa vyskytujú vo viacerých z nich je málo), hrozí, že následne použitý klasifikátor bude pomalý, pretože stavový priestor, ktorý sa konštruuje, je priveľký. Riešením je redukovať dimenzionalitu priemetom stavového priestoru do menej rozmerného podpriestoru, obľúbenou metódou je Principal Component Analysis [Zhang *et al.*, 2007, Chin *et al.*, 2006].

Špeciálny prípad klasifikácie textu predstavuje identifikácia jazyka dokumentu [Vojtek, 2006]. Pri tejto úlohe sa osvedčilo rozdeľovať text dokumentu na menšie jednotky, ako sú slová. Pri využití reťazcov znakov možno aplikovať metódy založené na n-gram analýze [Cavnar & Trenkle, 1994] a Markovových reťazcoch [Teahan, 2000].

Atribúty inštancií

Vlastné a odvodené atribúty predstavujú východisko pre atribútovo-viazané metódy. Zaužívaným spôsobom ako triediť atribúty je rozdeliť ich na *kvalitatívne* a *kvantitatívne* [Giudici, 2003].

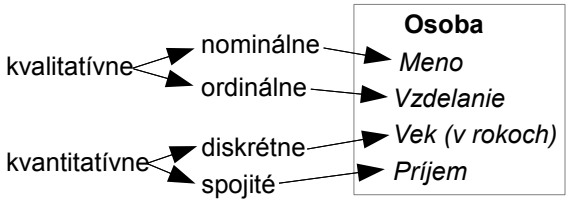
Kvalitatívne atribúty[4] zvyčajne predstavujú prídavné mená, napr. atribút `pohlavie` (hodnoty `mužské`, `ženské`), poštové smerovacie číslo, či krajské a okresné delenie Slovenskej republiky. Ak možno hodnoty kvalitatívneho atribútu zoradiť, hovoríme o *ordinálnom* atribúte (napr. stupeň vzdelania). Nezoraditeľné kvalitatívne atribúty, napr. farba vlasov osoby, nazývame

[4]Presnejšie by bolo hovoriť o *hodnotách atribútov*, ale pre zvýšenie zrozumiteľnosti budeme uvádzať aj pre tento prípad výraz *atribút*.

nominálne[56].

Pri kvalitatívnych nominálnych atribútoch vieme zaviesť reláciu zhody resp. odlišnosti $(=, \neq)$, ordinálne atribúty sú bohatšie o relácie zoradenia $(<, >, \leq, \geq)$. Za špeciálnu podskupinu ordinálnych atribútov považujeme tzv. *cyklické* atribúty[7], napr. dni v týždni.

Kvantitatívne atribúty nadobúdajú číselné hodnoty a rozdeľujú sa na *diskrétne* a *spojité*. Na rozdiel od kvalitatívnych atribútov môžeme s kvantitatívnymi atribútmi vykonávať aritmetické operácie (napr. sčítanie, podiel). Klasifikácia typov atribútov je zobrazená na obr. 3.3.



Obrázok 3.3: Rôzne typy atribútov na príklade inštancie `Osoba`.

Rozdelenie atribútov na kvalitatívne a kvantitatívne vychádza z dolovania v dátach, kde nám toto rozdelenie pomáha určiť vhodnú (či skôr použiteľnú) metódu dolovania v dátach. Napr. na predikciu ordinálneho spojitého atribútu je vhodná lineárna regresia, kým pri predikcii binárneho diskrétneho atribútu sa používa logistická regresia [Giudici, 2003].

Pri zohľadnení vzťahov medzi inštanciami je zmysluplné rozdeľovať atribúty na vlastné a odvodené. Odvodené atribúty rozdeľujeme na pravdepodobnostné a priame (angl. probabilistic and fixed).

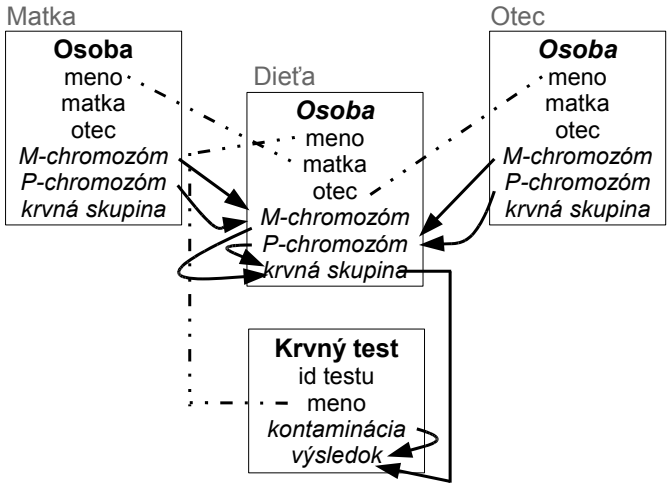
Na obr. 3.4 je príklad jednoduchého genetického modelu. Krvná skupina

[5]Pri delení kvantitatívnych atribútov na ordinálne/nominálne je určujúci aj kontext. Napríklad v súčasnosti je v našej spoločnosti legislatívne zakotvené nazerať na atribút `rasa` (inštancia `osoba`) ako na nominálny atribút, ale napr. v období Tretej ríše Norimberské zákony o občianstve a rase podporovali pohľad na atribút `rasa` v ordinálnom nazeraní (`http://frank.mtsu.edu/~baustin/nurmlaw2.html` [cit. 2009-09-02]).

[6]Číselná hodnota atribútu nemusí vždy nevyhnutne znamenať, že ide o ordinálny atribút – napr. knižničné Medzinárodné desatinné triedenie označuje triedy aj číselne, ale ide o nominálny atribút.

[7]Vid' `http://msdn.microsoft.com/en-us/library/ms174572.aspx` [cit.2009-10-07].

osoby je tu určená jedným génom. `M-chromozóm` je kópia matkinho chromozómu obsahujúceho daný gén (neprerušovaná čiara medzi `Matka.M-chromozóm` a `Dieťa.M-chromozóm` naznačuje pravdepodobnostnú väzbu). Prerušovaná čiara medzi `Matka.meno` a `Dieťa.matka` naznačuje priamy vzťah. Atribúty sú v súlade s týmito reláciami naznačené normálnym písmom (priamy typ `Dieťa.meno`) alebo kurzívou (pravdepodobnostný typ `Dieťa.krvná_skupina`).

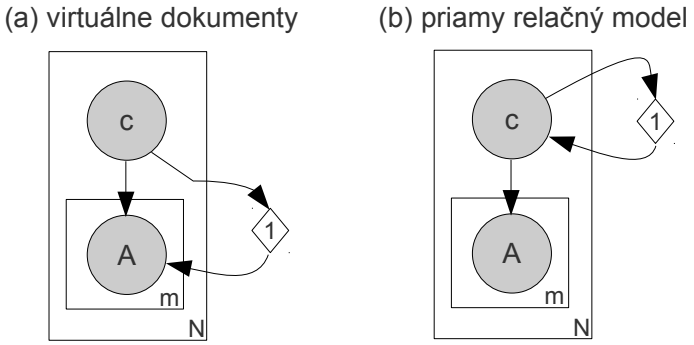


Obrázok 3.4: Príklad prepojených inštancií s priamymi a pravdepodobnostnými atribútmi a reláciami (prevzaté z [Friedman *et al.*, 1999]).

Atribúty môžu byť kontextovo závislé a nezávislé. Napríklad, denný úhrn zrážok rovný `5mm` môže byť klasifikovaný ako extrémne vysoký v suchom období roka v púštnej oblasti, ale zároveň môže byť považovaný za veľmi nízky počas monzúnového obdobia na tom istom mieste. Potom tvrdenie, že *úhrn zrážok je vysoký*, dáva zmysel len v kontexte so zemepisnou polohou a ročným obdobím [Jacob, 2004].

3.2.2 Virtuálne dokumenty

Ak medzi atribúty inštancie zaradíme aj atribúty susedov, hovoríme o takejto inštancii ako o *virtuálnom dokumente*. Medzi metódy z tejto triedy patrí napr. prístup uvedený v prácach [Chakrabarti *et al.*, 1998]
a [Slattery & Mitchell, 2000], kde sa pri klasifikácii webových stránok berie do úvahy nielen vlastný text inštancie, ale aj texty webových stránok ktoré sú s klasifikovanou stránkou prepojené cez hypertextové odkazy. Zodpovedajúci Jensenov diagram je zobrazený na obr. 3.5(a). Prepojenie na pravej strane (číslovka 1 v kosoštvorci) vyjadruje závislosť triedy vrcholu na atribútoch jeho priamych susedov.



Obrázok 3.5: Schéma pre virtuálne dokumenty a priamy relačný model.

Existujú viaceré spôsoby *absorbcie* atribútov susedných inštancií. Najjednoduchšia možnosť je spojit' obsah vlastných a susedných atribútov bez toho, aby boli odlíšené (klasifikátor v takomto prípade nie je informovaný o tom, ktorá hodnota atribútu bola pôvodne vlastná a ktorá je prevzatá). Ďalšou možnosť'ou je označkovať prevzaté atribúty a dať tak klasifikátoru informáciu o ich odlišnom pôvode (zvyčajne s cieľ'om priradiť im inú váhu).

Spôsoby samotnej klasifikácie sú rôzne, v práci [Chakrabarti *et al.*, 1998] sa využíva Bayesov klasifikátor, v práci [Lu & Getoor, 2003] logistická regresia. Výhodou takéhoto prístupu je, že po úprave atribútov môžeme použiť' bežný atribútovo-viazaný klasifikátor.

3.2.3 Priame relačné metódy

Reprezentácia priamych relačných metód [8] je podľa Jensenovej notácie naznačená na obr. 3.5(b). Príslušnosť inštancie k triede je odvodená nielen z hodnôt vlastných atribútov, ale zohľadňujú sa triedy ku ktorým prislúchajú susedné inštancie (číslo uvedené v kosoštvorci vyjadruje podľa Jensena šírku okolia, ktorá sa zohľadňuje).

V nasledovných častiach uvádzame vybrané relačné prístupy. Keďže uvažujeme reprezentáciu dát grafom, budeme používať aj označenie z tab. 3.1. Pri prehľade metód nám bola nápomocná práca [Macskassy & Provost, 2007].

V ďalších častiach budeme notáciou $\hat{c}_a$ označovať atribútovo-viazané klasifikačné metódy a $\hat{c}_r$ priame relačné metódy. Pre úplnosť ešte doplňme niekoľko definícií[9].

Definícia 3.1: Graf. Graf je dvojica množín $G = (V, E)$ taká, že $E \subseteq [V]^2$, t.j. prvky v množine E sú dvojprvkové podmnožiny množiny V. Prvky množiny V nazývame *vrcholy* grafu G, prvky množiny E nazývame *hrany* grafu G.

Definícia 3.2: Množina susedných vrcholov. V grafe $G(V, E)$, kde V je množina všetkých vrcholov grafu a E množina všetkých jeho hrán označuje V_k množinu takých vrcholov, z ktorých každý je s vrcholom $v_k \in V$ spojený hranou.

Definícia 3.3: Príslušnosť k triede. Hodnota $p(c_m|v_k)$ vyjadruje pravdepodobnosť, s akou vrchol v_k prislúcha k triede c_m. Túto pravdepodobnosť nazývame príslušnosť k triede.

Po uvedení definície príslušnosti ku triede je očividný rozdiel medzi virtuálnymi dokumentami a priamymi relačnými metódami. V prvom prípade cez relácie prenášame atribúty, kým v prípade priamych relačných metód inštancie zdieľajú príslušnosť ku triede.

[8] V práci [Macskassy & Provost, 2007] sa táto trieda metód označuje ako *relačný model* (angl. relational model), ale keďže zvyčajne sa v literatúre klasifikačným modelom myslí realizácia nad konkrétnymi dátami, používame výraz *metóda*.

[9] Definície z teórie grafov preberáme z práce [Diestel, 2005]

Jednoduchý relačný klasifikátor (Simple Relational Classifier)

Metóda SRC (Simple Relational Classifier) [Macskassy & Provost, 2003] predikuje triedu c_m danej inštancie v_k podľa tohto vzorca:

$$p(c_m|v_k) = \frac{1}{Z} \sum_{v_j \in V_k | class(v_j)=c_m} w(v_k, v_j) \tag{3.1}$$

kde V_k je množina susedných vrcholov, $Z = \sum_{v_j \in V_k} w(v_k, v_j)$ normalizuje výsledok, $w(v_k, v_j)$ je váha hrany medzi vrcholmi v_j a v_k, a c_m je trieda z množiny všetkých tried C.

Podmienená pravdepodobnosť $p(c_m|v_k)$ vyjadruje zložku vektora podľa definície 3.3. Jej celkový tvar je pre vrchol v_k a triedy $C = \{c_1, c_2, \ldots, c_m\}$ takýto:

$$\boldsymbol{p}(v_k, C) = \begin{bmatrix} p(c_1|v_k) \\ p(c_2|v_k) \\ \ldots \\ p(c_2|v_k) \end{bmatrix}$$

Keďže pri relačnom modeli uvažujeme najmä o príslušnosti k triede od susedných vrcholov, korektnejšie by bolo odhad triedy c_m pre vrchol v_k zapisovať $p(c_m|V_k)$ namiesto $p(c_m|v_k)$. V literatúre sa však môžeme stretnúť najmä s druhým spôsobom zápisu, preto ho budeme používať'[10].

Metóda SRC predstavuje najjednoduchšiu možnú realizáciu výmeny príslušnosti ku triede medzi vrcholmi. Do vlastnej príslušnosti vrcholu ku triede premieta pomerné váhované zastúpenie tried susedných vrcholov. Je očividné, že na to, aby metóda fungovala korektne, je nevyhnutné, aby v grafe prevládala homofília, t.j. aby triedy do ktorých patria dvojice vrcholov spojených hranou korelovali (viac v časti 4.1.2). Táto súvislosť s homofíliu sa týka aj ostatných priamych relačných metód.

[10] Zápis $p(c_m|v_k)$ je celkovo zjednodušujúci, pretože v jazyku podmienenej pravdepodobnosti by sme tento výraz mohli interpretovať slovami *Aká je pravdepodobnosť, že ak nastal jav v_k, nastane jav c_m?*

Relačná zložka IRC metódy (Iterative Reinforcement Categorization)

Metóda IRC (Iterative Reinforcement Categorization) [Xue *et al.*, 2006] sa podobá predošlej metóde, ale umožňuje aplikovať rôzne váhy pre vrcholy prislúchajúce k trénovacej a testovacej množine. Výpočet príslušnosti k triede je takýto:

$$p(c_m|v_k) = \lambda_1 p(c_m|v_k) + \lambda_2 \frac{\sum_{v_z \in V_k \cap X_{tr}} w(v_k, v_z) p(c_m|v_z)}{|V_k \cap X_{tr}|} + \\ + \lambda_3 \frac{\sum_{x_z \in V_k \cap X_{tst}} w(v_k, v_z) p(c_m|v_z)}{|V_k \cap X_{tst}|} \tag{3.2}$$

kde $V_k \cap X_{tr}$ je množina susedných vrcholov k vrcholu v_k, ktoré prislúchajú k trénovacej množine (podobne pre $V_k \cap X_{tst}$). λ_1, λ_2 a λ_3 sú váhy, ktorými môžeme nastaviť vplyv vlastnej zložky, trénovacej a testovacej množiny na výslednú príslušnosť vrcholu (v_k) k triede.

Relačná zložka REC metódy (Relational Ensemble Classification)

Metóda REC [Preisach & Schmidt-Thieme, 2006] (Relational Ensemble Classification) sa na úrovni relačnej zložky klasifikácie podobá na metódu IRC. Pri tvorbe vektora príslušnosti k triede pre vrchol v_k je informácia zo susedných vrcholov absorbovaná týmto spôsobom:

$$p(c_m|v_k) = \left(\prod_{v_z \in V_k} P(c_m|v_z)^{w(v_k, v_z)} \right)^{1/\sum_{v_z \in V_k} w(v_k, v_z)} \tag{3.3}$$

čiže sa aplikuje geometrický priemer, na rozdiel od predošlých dvoch metód, kde sa používa aritmetický priemer.

CDRN (Class-distribution Relational Neighbor Classifier)

Metóda CDRN [Macskassy & Provost, 2007] (Class-distribution Relational Neighbor Classifier) spresňuje metódu SRC o zohľadnenie lokálnej distribúcie tried v susedstve vrcholu. Zavádza sa tzv. *vektor tried vrcholu* a *referenčný vektor triedy.*

Pre vrchol v_k označujeme $\boldsymbol{CV}(v_k)$ vektor tried, jeho zložka pre triedu c_m je určená takto:

$$\boldsymbol{CV}(v_k)_m = \sum_{v_z \in V_k \cap X_{tr}, trieda(v_k)=c_m} w(v_k, v_z) \tag{3.4}$$

Vektor triedy určuje distribúciu, s akou susedia vrcholu v_k prislúchajú k jednotlivých triedam. Na základe vektorov tried všetkých trénovacích vrcholov v grafe definujeme referenčný vektor triedy. Pre triedu c_m je určený takto:

$$\boldsymbol{RV}(c_m) = \frac{1}{|\{v_k \in V_k \cap X_{tr}, trieda(v_k) = c_m\}|} \sum_{v_z \in V_k \cap X_{tr}, trieda(v_z)=c_m} \boldsymbol{CV}(v_z) \tag{3.5}$$

Referenčný vektor je teda vektor, ktorý aritmetickým priemerom združuje hodnoty jednotlivých vektorov tried $\boldsymbol{CV}(c_k)$ pre všetky vrcholy z trénovacej množiny, $v_k \in X_{tr}$. Jedna zložka vektora $\boldsymbol{RV}(c_m)$ tak vyjadruje, akých susedov majú vrcholy, ktoré patria do triedy c_m.

Referenčný vektor vytvára predpoklad, ktorý potom aplikujeme na vrcholy z testovacej množiny. Pravdepodobnost', že vrchol v_k patrí do triedy c_m je podľa metódy CDRN takáto:

$$p(c_m|v_k) = sim(\boldsymbol{CV}(v_k), \boldsymbol{RV}(c_m)) \tag{3.6}$$

kde $sim(a, b)$ je funkcia podobnosti vektorov, normalizovaná na interval $\langle 0, 1\rangle$, napr. kosínová podobnost'. Výpočet teda porovnáva pre vrchol v_k a triedu c_m distribúciu triedy v jeho susedstve voči *globálnej* distribúcii triedy v grafe. Víťazná trieda pre vrchol v_k je tá, pre ktorú sa lokálna distribúcia najlepšie zhoduje s globálnou.

Metóda nBC (Network-only Bayes Classifier)

Metóda nBC (Network-only Bayes Classifier) [Macskassy & Provost, 2007] vychádza z práce [Chakrabarti *et al.*, 1998]. Pri výpočte príslušnosti vrcholu v_k k triede c_m sa využíva Bayesov teorém:

$$p(c_m|v_k) = \frac{p(v_k|c_m)p(c_m)}{p(v_k)} \tag{3.7}$$

to znamená, že $p(c_m)$ je pomer počtu vrcholov ktoré patria do triedy c_m voči všetkým vrcholom, $p(v_k) = \frac{1}{|V|}$ a podmienená pravdepodobnosť $p(v_k|c_m)$ je určené ako:

$$p(v_k|c_m) = \frac{1}{Z} \prod_{v_j \in V_k} p(trieda(v_j) = c_n|c_m)^{w(v_j,c_k)} \tag{3.8}$$

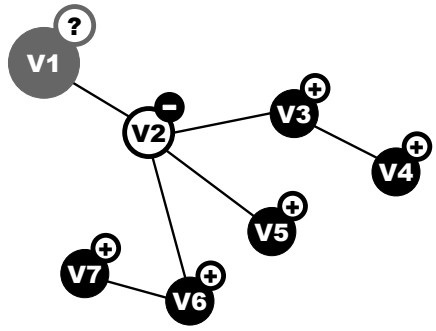
kde Z je normalizačná konštanta.

Uvedené priame relačné metódy spája ich jednoduchosť a priamočiarosť. Zároveň sa metódy *odosobňujú* od atribútov inštancií a manipulujú iba s príslušnosťou ku triede. Pri striktnom pohľade už vlastne vôbec nejde o učenie s učiteľom, pretože metódy nemajú trénovaciu a testovaciu fázu. Prakticky sa priama relačná metóda kombinuje s atribútovou metódou, ktorá zabezpečí vytvorenie príslušnosti ku triede z atribútov (viac v časti 3.4). Pre korektné a zmysluplné zatriedenie je pre priame relačné metódy kľúčové, aby bola v klasifikovanom grafe prítomná homofília, teda aby pravdepodobnosť s akou majú dva vrcholy rovnakú triedu, korešpondovala s existenciou hrany medzi týmito vrcholmi.

3.2.4 Kolektívne usudzovanie

Atribútové klasifikačné prístupy predstavujú spôsob, ako vlastné atribúty inštancie transformovať na vektor príslušnosti k triede. Priame relačné metódy nám umožňujú nazerať na prepojené inštancie a zdieľať medzi sebou príslušnosť k triede. Toto zdieľanie však zodpovedá Markovovskému reťazcu rádu 1, čiže vrchol cez relačný model *vidí* iba svoje priame okolie (vrcholy priamo spojené hranou). Veľakrát je užitočné absorbovať informáciu aj z vzdialenejších vrcholov. Jedným z dôvodov je situácia na obr. 3.6.

V grafe vrcholy prislúchajú k dvom triedam $C = \{c_+, c_-\}$. Vrchol v_1 patrí do testovacej množiny a jeho triedu chceme určiť pomocou relačného modelu, ostatné vrcholy majú triedy známe. Ak by sme aplikovali ktorúkoľvek priamu relačnú metódu, stane sa, že vrcholu v_1 bude priradená trieda c_-. Keďže však zvyšok grafu obsahuje už len vrcholy z triedy c_+, môžeme sa domnievať, že vrchol v_2 bol nesprávne označený a takisto mal prislúchať k triede c_+.

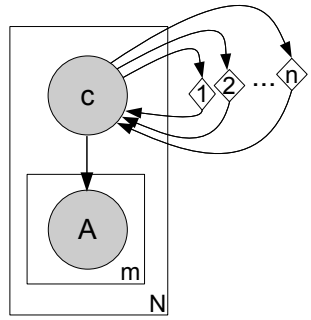


Obrázok 3.6: Príklad grafu, v ktorom priame relačné metódy nemusia byť dostatočné.

Existuje viacero spôsobov ako premostiť informáciu o príslušnosti k triede od nepriamo susediacich vrcholov k vrcholu v_1. V Jensenovej notácii im zodpovedá diagram na obr. 3.7. Tento diagram sa v pôvodnej notácii nenachádza, ide o nami navrhnuté rozšírenie. Jednotlivé hodnoty od 1 po n v kosoštvorcoch zodpovedajú vrcholom z danej vzdialenosti (myslené dĺžkou cesty v grafe, pozri def. 4.1).

Prvým spôsobom ako preniesť informáciu aj zo vzdialenejších vrcholov je priamo nazerať aj na *susedných susedov*, čiže na vrcholy, ktoré sú od vrcholu v_k vzdialené cez dve hrany. Takéto rozšírenie je navrhované v metóde REC pre tie vrcholy, ktoré majú príliš nízky stupeň vlastného susedstva [Preisach & Schmidt-Thieme, 2006], podobný návrh je aj v prácach [Gallagher *et al.*, 2008, Vojtek & Bieliková, 2010].

Druhým spôsobom, ako *rozšíriť obzor* vrcholu pri jeho klasifikácii, je tzv. kolektívne usudzovanie (angl. collective inferencing) [Gürel & Kersting, 2005]. V tomto prístupe sa iteratívne nad grafom viacnásobne aplikuje niektorý



Obrázok 3.7: Kolektívne usudzovanie na základe [Jensen *et al.*, 2004].

z priamych relačných klasifikátorov, až kým sa šírenie príslušnosti medzi triedami neustáli. Kolektívne usudzovanie v striktnom zmysle slova teda nie je súbor klasifikačných metód, ale trieda prístupov k optimalizácii prvotne dosiahnutého stavu klasifikovaného grafu. Počet iterácií kolektívneho usudzovania určuje šírku okolia, z ktorej sa k vrcholu môžu dostať informácie. Vzhľadom na charakter metód je však táto informácia primerane *utlmená* tak, aby bola vo vhodnej miere zohľadnená sila homofílie.

Nižšie uvádzané metódy kolektívneho usudzovania nevznikli samostatne, ale autori ich navrhovali komplexne, t.j. aj so zapojením atribútového klasifikátora aj relačnej zložky.

Iteratívna klasifikácia

S metódou s názvom iteratívna klasifikácia sa stretáme v prácach [Lu & Getoor, 2003, Macskassy & Provost, 2007]. Jednotlivé kroky metódy sú uvedené v algoritme 1. Z algoritmu vyplýva, že môže nastať situácia, kedy po ukončení kolektívneho usudzovania niektoré vrcholy z X_{tst} zostanú bez priradenej triedy. To môže byť výhodné, ak nám záleží na presnej klasifikácii a uprednostňujeme situáciu kedy klasifikátor radšej triedu nepriradí, ako by ju priradil s rizikom zvýšenej chyby.

Algoritmus 1 Metóda iteratívnej klasifikácie

Vstupná podmienka: Inštancie $v_1, v_2, \ldots, v_k \in X_{tst}$
Výstupná podmienka: nie je

1: Vrcholy $v_1, v_2, \ldots, v_k$ sa náhodne zoradia do usporiadania O.
2: Pre každý vrchol v poradí $v_i \in O$:
   1. aplikuje sa priama relačná metóda inicializovaná atribútovým klasifikátorom $\hat{c}_r \leftarrow \hat{c}_a$, pričom sa do úvahy berú iba tie susedné vrcholy, ktoré sú z trénovacej množiny (majú určenú triedu), teda $V_i \cap X_{tr}$. Ak $V_i \cap X_{tr} = \emptyset$, vrchol v_i zostáva s neurčenou triedou.
   2. trieda vrcholu v_i sa určí takto: $trieda(v_i) = argmax_{c_j}[p(c_j|v_i)]$,
3: Kroky 1 a 2 sa opakujú, kým $\exists v_i : V_i \cap X_{tr} \neq \emptyset$ (čiže v danej iterácii aspoň jeden vrchol zmenil stav) alebo kým sa nedosiahne 1000 iterácií.

Metóda IRC (Iterative Reinforcement Categorization)

Metóda IRC (Iterative Reinforcement Categorization) [Xue *et al.*, 2006] pozostáva z krokov uvedených v algoritme 2. Postup je zjednodušený pre jeden typ vrcholu. Pôvodný návrh umožňuje klasifikovať viac typov vrcholov naraz. Viactypovému rozšíreniu iteratívnych klasifikačným metód sa v tejto práci pre udržanie prehľadnosti nevenujeme (zaoberá sa ňou práca [Vojtek, 2008]).

Algoritmus 2 Metóda IRC

Vstupná podmienka: Inštancie $v_1, v_2, \ldots, v_k \in X_{tst}$
Výstupná podmienka: Každá inštancia má priradenú triedu, $\forall v_k \in V : trieda(v_k) = c \in C$

1: Na základe vlastných atribútov inštancie sa pomocou atribútovo-viazaného klasifikátora $\hat{c}_a$ inicializuje príslušnosť k triede pre každý vrchol v_k,
2: Príslušnosť k triede sa pre každý vrchol upraví na základe:
   1. vlastnej distribúcie,
   2. distribúcie susedných vrcholov (pozri vzorec (3.2)).
3: Krok 2 sa opakuje až kým rozdiel v zmene distribúcií pre všetky vrcholy medzi dvoma iteráciami neklesne pod určitú medzu,
4: vrchol v_k sa označí triedou, ktorá v príslušnosti k triede dominuje, to znamená $c = argmax_{c_i}[p(c_i|v_k)]$.

Relaxation Labeling

Metóda bola publikovaná v [Chakrabarti *et al.*, 1998], jej kroky sú uvedené v algoritme 3. Metóda je veľmi podobná metóde iteratívnej klasifikácie.

Algoritmus 3 Relaxation Labeling

Vstupná podmienka: Inštancie $v_1, v_2, \ldots, v_k \in X_{tst}$
Výstupná podmienka: Každá inštancia má priradenú triedu $\forall v_k \in V : trieda(v_k) = c \in C$

1: $\forall v_k : X_{tst}$ sa určí vektor príslušnosti k triede $\boldsymbol{p}(v_k, C)$ pomocou $\hat{c}_a$.
2: $\forall v_k : X_{tst}$ sa upraví stav $\boldsymbol{p}(v_k, C)$ na základe $\hat{c}_r$.
3: Krok 2 sa opakuje v počte iterácií $T = 99$. Predikcia $\hat{c}_r$ pre graf v iterácii (t) prebieha na základe stavu grafu z iterácie $(t-1)$, t.j. berú sa do úvahy vektory $\boldsymbol{p}(v_k, C)_{(t)}$

Gibbsovo vzorkovanie

Gibbsovo vzorkovanie bolo zavedené v práci [Geman *et al.*, 1993] ako metóda pre odhadovanie zloženej distribúcie dvoch a viac náhodných premenných. Ak hodnota príslušnosti vrcholu k triede $trieda(v_k) \in C$ je náhodná premenná, potom odhad tejto hodnoty vieme iteratívne vyjadriť takto[11]:

$$trieda(v_k)_{(t+1)} \sim p(trieda(v_k)|trieda(v_1)_{(t)}, trieda(v_2)_{(t)}, \ldots) \qquad (3.9)$$

Kroky metódy sú uvedené v algoritme 4 podľa [Geman *et al.*, 1993, Macskassy & Provost, 2007].

Pri Gibbsovom vzorkovaní i metóde Relaxation Labelling je pozoruhodné, že majú určený pevný, pomerne vysoký počet iterácií. Rozsah a štruktúra grafu teda vôbec nie je zohľadnená v prospech rýchleho ukončenia algoritmu.

[11] Notácia prevzatá z `http://web.mit.edu/~wingated/www/introductions/mcmc-gibbs-intro.pdf` [cit. 2009-11-07]

Algoritmus 4 Gibbsovo vzorkovanie

Vstupná podmienka: Inštancie $v_1, v_2, \ldots, v_k \in X_{tst}$

Výstupná podmienka: Každá inštancia má priradenú triedu $\forall v_k \in V : trieda(v_k) = c \in C$

1: $\forall v_k : X_{tst}$ sa určí vektor príslušnosti k triede $\boldsymbol{p}(v_k, C)$ pomocou $\hat{c}_a$.

2: $\forall v_k : X_{tst}$ sa upraví stav $\boldsymbol{p}(v_k, C)$ na základe $\hat{c}_r$.

3: Vrcholy $v_1, v_2, \ldots, v_k$ sa náhodne zoradia do usporiadania O.

4: Pre každý vrchol v poradí $v_i \in O$ sa vektor príslušnosti k triede $\boldsymbol{p}(v_i, C)$ upraví na základe $\hat{c}_r$ a táto hodnota sa hneď uloží ako aktuálna pre $\boldsymbol{p}(v_i, C)$. Keď teda $\hat{c}_r$ určuje $\boldsymbol{p}(v_i, C)$ v iterácii t, vektory $\boldsymbol{p}(v_1, C), \boldsymbol{p}(v_2, C), \ldots, \boldsymbol{p}(v_{i-1}, C)$ už obsahujú údaje z iterácie t, vektory $\boldsymbol{p}(v_{i+1}, C), \boldsymbol{p}(v_{i+2}, C), \ldots, \boldsymbol{p}(v_k, C)$ pochádzajú ešte z iterácie $(t-1)$.

5: Krok 4 sa zopakuje v 200 iteráciách, priebeh sa nezaznamenáva.

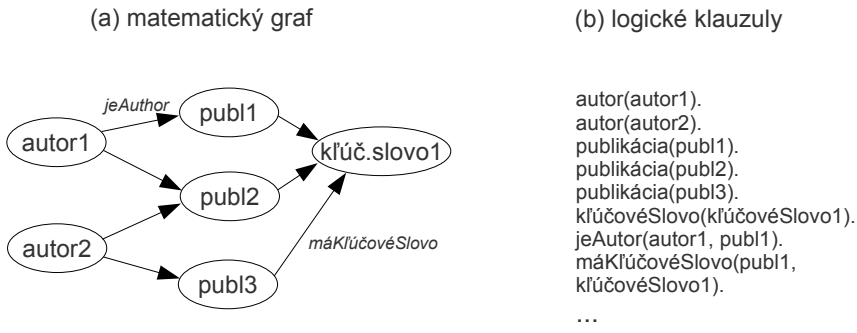
6: Krok 4 sa zopakuje v počte iterácií $T = 2000$, zaznamenáva sa počet priradení každej triedy pre každý vrchol $v_i \in X_{tst}$. Na základe distribúcie počtov priradených *tried* sa určia jednotlivé zložky vektora príslušnosti k triede ako $p(c_m|v_i) = \frac{1}{T} \sum_{t \in T} p(c_m|v_i)_{(t)}$.

3.3 Alternatívna reprezentácia dát

Relačné metódy uvádzané v tejto práci spája zápis dát pomocou matematického grafu. Existujú však aj prístupy založené na reprezentácii dát a klasifikovaní pomocou induktívneho logického programovania (ILP) [Dzeroski & Todorovski, 1995].

Príklad zachytenia rovnakej dátovej vzorky pomocou grafu aj klauzúl logického programovania je zobrazený na obr. 3.8. Typ `autor` má dve inštancie, `autor1` a `autor2`, typ `publikácia` má tri inštancie. Relácia medzi `autorom` a `publikáciou` má názov `jeAutor`. Podobne sú v dátovej vzorke prítomné typy inštancií `kľúčové slovo` a vzťah `máKľúčovéSlovo`. Relačná klasifikačná metóda, ktorá využíva ILP je v práci [Frank *et al.*, 2007].

V práci [Ketkar *et al.*, 2005] sú experimentálne porovnané oba prístupy k reprezentácii dát na základe schopnosti zachytiť štrukturálne rozsiahle alebo sémanticky bohaté koncepty. Uvádza sa, že grafová reprezentácia je vhodnejšia pre štrukturálne zložité koncepty, kým ILP je vhodné pre sémanticky zložitejšie svety.



Obrázok 3.8: Doména publikácií, reprezentácia grafom aj klauzulami.

V našej práci sme sa zamerali na grafovú reprezentáciu, pretože táto oblasť relačnej klasifikácie je bohatšie rozvinutá a tiež dátové vzorky, s ktorými sme uvažovali vykonať experimentálne overenie vznikali ako grafy.

Zmieňované paradigmy reprezentácie dát určujú matematický aparát ktorý máme k dispozícii pri návrhu klasifikačných metód. Pri použití dátovej vzorky v praxi je z dôvodu rýchleho a pohodlného narábania s veľkým objemom dát obvyklé ukladať dáta v entitno-relačnej databáze, čiže v reprezentácii, ktorá pozostáva z množiny tabuliek s entitami a reláciami.

3.4 Zhrnutie metód relačnej klasifikácie

V prehľade relačných metód sme uviedli najvýznamnejšie klasifikačné metódy atribútovo-viazaného prístupu, väčšinu známych priamych relačných metód a všetky nám známe prístupy k kolektívnemu usudzovaniu. Tieto triedy metód sú zvyčajne prepojené, teda prístupy pre kolektívne usudzovanie $\hat{c}_{ci}$ v sebe zahŕňajú priame relačné metódy a tie v sebe zahŕňajú atribútové metódy: $\hat{c}_{ci} \leftarrow \hat{c}_r \leftarrow \hat{c}_a$. Okrem toho sme uviedli niekoľko metód tzv. virtuálnych dokumentov $\hat{c}_{vd}$, ktoré predstavujú doplnkový smer k priamym relačným metódam. Jednotlivé triedy metód možno plnohodnotne použiť v týchto konfiguráciách:

- $\hat{c}_a$: atribútovo-viazaný klasifikátor,
- $\hat{c}_r \leftarrow \hat{c}_a$: priamy relačný klasifikátor inicializovaný $\hat{c}_a$,
- $\hat{c}_{ci} \leftarrow \hat{c}_r \leftarrow \hat{c}_a$: kolektívne usudzovanie obaľujúce priamy relačný klasifikátor inicializovaný atribútovo-viazaným klasifikátorom,
- $\hat{c}_r$: priamy relačný klasifikátor bez schopnosti absorbovať informáciu z atribútov,
- $\hat{c}_{ci} \leftarrow \hat{c}_r$: kolektívne usudzovanie nad priamym relačným klasifikátorom bez schopnosti absorbovať informáciu z atribútov[12],
- $\hat{c}_{vd}$: použitie virtuálnych dokumentov.

3.4.1 Porovnanie jednotlivých prístupov

Uvedené triedy metód relačnej klasifikácie sú porovnané z hľadiska parametrov dátovej vzorky v tab. 3.2. V riadkoch sú uvedené tri zvyčajne sa vyskytujúce archetypy dátových vzoriek: inštancie s atribútmi bez explicitne uvedených relácií (napr. tabuľka zákazníkov v relačnej databáze), inštancie s atribútmi aj reláciami (graf s vrcholmi, ktoré majú aj atribúty) a graf s inštanciami bez atribútov.

Tabuľka 3.2: Porovnanie aplikovateľnosti tried klasifikačných metód podľa typu dátovej vzorky (*áno* znamená, že metóda je aplikovateľná).

	$\hat{c}_a$	$\hat{c}_r$	$\hat{c}_{ci}$	$\hat{c}_{vd}$
iba atribúty	áno	nie	nie	áno
atribúty, relácie	áno	áno	áno	áno
iba relácie	nie	áno	áno	nie

V tab. 3.3 uvádzame porovnanie jednotlivých metód kolektívneho usudzovania na základe zvolených vlastností. Zaručenie triedy znamená, že metóda určite priradí každému klasifikovanému vrcholu triedu, pevná zastavovacia podmienka znamená, že klasifikátor je zastavený najneskôr po určitom počte iterácií, bez ohľadu na priebeh. Ucelenosť iterácie značí, že relačná metóda počas iterácie (t) berie do úvahy stav z iterácie ($t-1$).

[12]Tento a predošlý prípad sú zaujímavé skôr teoreticky, napr. pre generované grafy.

Tabuľka 3.3: Porovnanie prístupov k kolektívnemu usudzovaniu.

	IRC	*Iter. klas.*	*Gibbs.*	*Relax. Label.*
$\forall v_i \in X_{tst}$ zaručuje triedu	áno	nie	áno	áno
pevná zast. podmienka	nie	áno	áno	áno
ucelenost' iterácie	áno	áno	nie	áno

Pre priame relačné metódy ($\hat{c}_r$) je charakteristické, že pri odhade príslušnosti konkrétneho vrcholu k triede nahliadajú na okolie vrcholu, a práve susediace vrcholy ovplyvňujú jeho vlastnú príslušnosť k triede. Tento poznatok je až príliš očividný a ľahko ho považovať za banálny – má však zásadné dôsledky na mechanizmus klasifikácie, pretože vnáša predpoklad homofílie, ktorému sa venujeme v kap. 4. Keďže metódy kolektívneho usudzovania v sebe obsahujú priame relačné metódy, závislosť od homofílie sa prenáša aj do nich.

Uvedený prístup ku *klasifikácii* klasifikačných metód prináša porovnanie metód na základe ich schopnosti zohľadniť a abstrahovať relačnú zložku dátovej vzorky. Pri kolektívnom usudzovaní je v oboch prehľadových prácach uvedený zjednodušený prípad, kedy je príslušnosť vrcholu k triede generovaná príslušnosťou susedných vrcholov a tvarom grafu, to znamená, že klasifikujeme na základe jednej náhodnej premennej (pre každý vrchol). Prakticky tento predpoklad určuje, že uvedené metódy sú vhodné pre grafy kde klasifikujeme len z jednej množiny tried. Pritom však existujú dátové vzorky (a klasifikačné problémy), v ktorých sa vyskytuje viacero klasifikačných premenných naraz. Napríklad v doméne vyhľadávania na webe (obr. 3.1) máme v grafe až tri typy vrcholov, pričom každý typ môže mať inú množinu klasifikačných tried. V takom prípade webové stránky zatrieďujeme podľa obsahu (triedy *šport*, *veda*) a používateľov zatrieďujeme podľa správania (triedy *náchylný kliknúť na kontextovú reklamu*, *ignoruje kontextovú reklamu*). Môžu nastať dve situácie:

- každý typ inštancie má svoju klasifikáciu. Metódy na priamu klasifikáciu takýchto dátových vzoriek sú analyzované v prácach [Macskassy & Provost, 2007] (kap. 3.5.4) a [Vojtek, 2008] (kap. 5),
- ak sa nad jedným typom inštancií aplikuje viacero klasifikácií, ide o fazetovú klasifikáciu. Pre tento prípad nám nie sú známe žiadne re-

> lačné metódy, ktoré by boli schopné využit' existenciu viacerých klasifikácií v prospech výsledku. Z technického hľadiska nie je problém použit' ktorúkoľvek metódu pre každú klasifikáciu (obdoba základného riešenia viactriedneho priradenia, pozri časť 7.1).

Z prehľadu a porovnania klasifikačných metód v tejto kapitole vidíme, že existuje pomerne veľa metód. Voľba správnej metódy pre určitú klasifikačnú úlohu práve preto nie je jednoduchá a zvyčajne ani jednoznačná. Je vhodné mat' na pamäti metodológiu CRISP-DM ([Fayyad *et al.*, 1996], [Shearer, 2000]), podľa ktorej sa najprv treba snažit' porozumiet' doméne a dátam a následne vychádzat' z tejto znalosti pri voľbe klasifikačnej metódy. Opačný, *idylický* postup, teda hľadat' univerzálnu klasifikačnú metódu vhodnú pre takmer každú úlohu totiž nie je v súlade s tzv. *No Free Lunch Theorem* [Wolpert & Macready, 1997]:

> Ak algoritmus vynikajúco rieši určitú podmnožinu všetkých úloh, potom nevyhnutne poskytuje slabé výsledky pri zvyšných úlohách.

V tejto kapitole sme v prehľade existujúcich relačných metód viackrát upozornili na silný predpoklad, vďaka ktorému relačné prístupy vôbec môžu poskytovat' korektné výsledky pri klasifikácii. Ide o implicitný predpoklad homofílie v grafe, ktorý je natoľko očividný a všadeprítomný, že pri návrhu metód sa bez pochýb predpokladalo, že klasifikované grafy sú homofílne. V skutočnosti to však tak byt' nemusí, homofília grafu sa pre jednotlivé vrcholy mení a pre niektoré vrcholy, či dokonca celé grafy môže byt' úplne nevýrazná. V ďalšej kapitole sa preto venujeme homofílii v grafe vo všeobecnosti a na základe tohto poznania definujeme homofíliu pre klasifikované grafy (vyznačujúce sa tým, že vrcholy majú priradenú príslušnost' ku triede).

Kapitola 4

Predpoklad homofílie a jeho dôsledky

Matematické grafy predstavujú užitočný mechanizmus na zaznamenanie previazanosti objektov vo svete okolo nás. Ak uvažujeme o grafoch, ktoré zachytávajú sociálne väzby v spoločenstve jedincov, zistíme, že v týchto grafoch je všadeprítomný jav nazývaný homofília. Sociológovia pod pojmom homofília rozumejú takýto úkaz [Mcpherson *et al.*, 2001]:

> Ľudia, ktorí sú navzájom prepojení (vzťahom) sú si podobní s vyššou pravdepodobnosťou ako ľudia, ktorí prepojení nie sú.

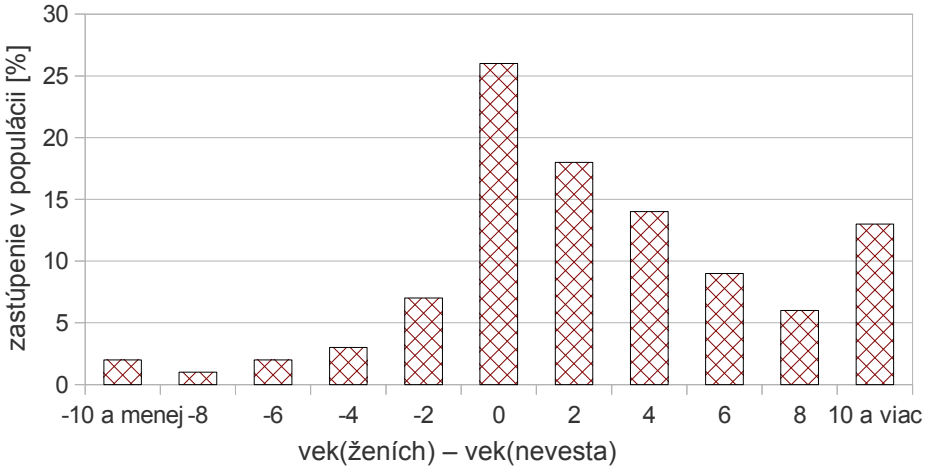
V našej spoločnosti sú homofílne tendencie prítomné pri mnohých typoch väzieb, uvádzame najvýznamnejšie z nich (inšpiráciu sme čerpali z prác [Thelwall, 2009, Mcpherson *et al.*, 2001]).

Vek

Vek je významný činiteľ pri vytváraní väzieb medzi osobami. So zvyšujúcim sa rozdielom vo veku významne klesá pravdepodobnosť vzniku väzby. Jeden z najznámejších príkladov vytvárania väzieb medzi osobami rovnakého veku je manželstvo. Distribúcia rozdielu vo veku medzi mužom a ženou je znázornená na obr. 4.1, na osi x je naznačený rozdiel medzi vekom ženícha a vekom nevesty v Nórsku v roku 2002[1]. Vidíme, že väčšina

[1] Zdroj Statistics Norway `http://www.ssb.no/english/magazine/art-2005-01-31-01-en.html` [cit. 2009-10-12].

dvojíc vstupuje do manželstva v rovnakom veku a pravdepodobnosť vzniku prepojenia klesá s rozdielom veku (okrajové hodnoty majú vyššiu pravdepodobnosť, pretože agregujú širšie časové intervaly).



Obrázok 4.1: Distribúcia rozdielu vo veku nevesty a ženícha v nórskej populácii v roku 2002.

Rasa a národnosť

Rasová príslušnosť a národnosť výrazne rozdeľujú ľudí po celom svete. Napriek vzrastajúcej globalizácii[2] nášho spoločenstva ľudia preferujú styk s ľuďmi rovnakej východiskovej kultúrnej pozície. Vo veľkých mestách vyspelých krajín (napr. Londýn) nastáva situácia, že v urbanizovanom priestore žijú na pomerne malej ploche takmer všetky ľudské rasy a národnosti, ale vytvárajú silne ohraničené zhluky. Dôsledkom je vznik pomerne dobre ohraničených *národných štvrtí*, napr. čínske štvrte v amerických mestách (New York, San Francisco).

[2]Globalizáciu v našom ponímaní chápeme ako zvyšovanie celosvetovej miery prepojenia medzi ľuďmi a nimi spravovanými statkami.

Náboženstvo a geografická poloha

Medziľudské vzťahy sú výrazne ovplyvnené náboženským presvedčením jednotlivca, v minulosti bola táto črta ešte výraznejšia. Táto tendencia je primárne pozorovateľná pri odlišných náboženstvách, kde vznik prepojení častokrát ovplyvňuje aj geografická poloha, v dôsledku čoho došlo napríklad v roku 1947 k oddeleniu moslimského Pakistanu od vtedajšej prevažne hinduistickej Britskej Indie. Podobné črty vykazujú aj denominácie určitého náboženstva, napr. prevaha šítskej vetvy Islamu v Iráne (89% z celej populácie) voči prevahe sunnitskej vetvy v Alžírsku (99% z celej populácie), alebo prevaha rímsko-katolíckej denominácie v Poľsku (89.8%) voči ortodoxnej vetve v Bielorusku (80%)[3].

Popri vyššie menovaných väzbách sa v spoločnosti prejavujú mnohé iné druhy prepojení, v ktorých je možné pozorovať homofílne tendencie, napr. vzdelanie, pohlavie, pracovná pozícia, správanie jedinca či materinský jazyk. Prítomnosť homofílie v uvedených väzbách je súčtom osobného výberu jednotlivca a spoločenských obmedzení, druhý vplyv často prevažuje nad osobnou voľbou. Segregácia mužov a žien je v moslimských krajinách, v porovnaní so súčasnou európskou civilizáciou, privedená do extrému ako dôsledok dôsledného dodržiavania právneho systému šaría (t.j. náboženská väzba). Národnostná homofília je mnohým ľuďom vnútená, keďže žijú v krajinách, ktoré sú málo otvorené svetu (napr. Bhután alebo Kórejská ľudovodemokratická republika).

Uvedené atribúty majú na život spoločnosti rôzne dôsledky. Segregácia na základe veku nemá zďaleka také negatívne dôsledky ako národnostná diferenciácia. Kým v prvom prípade hrozí napr. starším ľuďom strata kontaktu s novými trendami, či typické nepochopenie medzi rodičmi a ich deťmi v puberte, v prípade národnosti a vierovyznania sú dôsledky ďalekosiahlejšie a vedú k ozbrojeným konfliktom (napr. dlhodobé napätie na Blízkom východe medzi arabskými štátmi a Izraelom).

[3]Podľa CIA Factbook `https://www.cia.gov/library/publications/the-world-factbook/` [cit. 2009-10-01].

Homofília však má aj pozitívne dôsledky, jej zvýšenie môže odvrátiť vojnový konflikt, ako bulvárne uvádza tzv. *teória zlatých oblúkov*[4]:

> Žiadne dve krajiny, v ktorých vznikne pobočka rýchleho občerstvenia McDonald's, potom už medzi sebou nevedú vojnu.

Reťazec rýchleho občerstvenia predstavuje metaforu na výrazné zdieľanie spoločných hodnôt.

Napriek definícii na začiatku kapitoly a uvedeným príkladom nie je homofília v grafoch obmedzená na ľudské spoločenstvo. Ako sme v úvode uviedli, homofílne tendencie sa prejavujú v takmer akomkoľvek spoločenstve, ktoré sa skladá z jedincov vyvíjajúcich sa v čase a súperiacich o obmedzené zdroje. Homofíliu môžeme chápať ako prejav vývojovej stratégie. Vzhľadom na všadeprítomnosť tohto javu je však náročné oddeliť príčiny (dôvody vzniku) od následku, keďže jednotlivé typy väzieb medzi jedincami nie sú nezávislé. Napríklad náboženstvo a národnosť sú v niektorých spoločenstvách výrazne korelované, čo možno pozorovať na už spomínanej šítskej vetve islamu medzi iránskymi občanmi perzskej národnosti.

4.1 Miera homofílie

Samotný jav homofílie a jej dôsledky je jednoduché sledovať a brať do úvahy najmä tam, kde sa stretávame s dátami nesúcimi v sebe explicitné prepojenia, väzby (pozri diskusiu o explicitných a implicitných väzbách v kap. 3). Takéto dáta sú zvyčajne vhodne reprezentovateľné matematickým grafom, z čoho vyplýva možnosť zamerať sa pri analyzovaní homofílie na susednosť vrcholov. Cieľom tejto časti je definovať pojem *homofília v grafe* a určiť vhodnú metriku na meranie homofílie.

V dôsledku nášho záujmu o homofíliu v kontexte relačnej klasifikácie je pre nás výhodné nazerať na graf vrcholovo-orientovaným spôsobom, keďže pri analýze (spracovaní) grafu je klasifikačná metóda navrhnutá tak, že postupne sa prechádza cez vrcholy v grafe a analyzuje sa okolie zvoleného vrcholu. Väčšina relačných klasifikátorov spracováva graf práve takýmto vrcholovo-orientovaných spôsobom (pozri prehľad v kap. 3).

[4] Angl. Golden Arches Theory [Friedman, 1999].

4.1.1 Vrchol grafu a jeho okolie

Existuje mnoho spôsobov, akými sa môže generovať množina susedných vrcholov V_k na základe zvoleného vrchola v_k (definícia 3.2 v kap. 3.2.3), voľba generujúcej funkcie susedstva záleží od prípadu použitia. Rôzne prístupy na určenie toho, ktoré vrcholy spadajú do susedstva grafu dávajú rozličné pohľady na homofíliu toho istého grafu. Uvádzame niekoľko základných pohľadov na susedstvo v grafe.

Priame jednoduché susedstvo

Najjednoduchší prípad susedstva definujeme pomocou *cesty* v grafe[5].

Definícia 4.1: Cesta. Cesta je neprázdny graf $P = (V, E)$ v tvare $V = \{v_0, v_1, \ldots, v_k\}$, $E = v_0v_1, v_1v_2, \ldots, v_{k-1}v_k$ kde v_i sú všetky rôzne. Vrcholy v_0 a v_k sú spojené cestou P. Počet hrán cesty určuje *dĺžku cesty*, ktorú označujeme ako P^k.

Priame jednoduché susedstvo vrcholu v grafe tvorí množina $V_k = \{v_i | P(V, E), V = \{v_i, v_k\}, E = v_iv_k\}$, t.j. V_k zahŕňa len tie vrcholy, ktoré sú s vrcholom v_k priamo spojené hranou (cestou dĺžky P^1). Vo väčšine prípadov práce s grafom sa stretávame s týmto druhom susedstva.

Susedstvo n-tého stupňa

Zovšeobecnením priameho jednoduchého susedstva je susedstvo n-tého stupňa:

$$V_k = \{v_i | P(V, E), v_i \in V, v_k \in V, E = v_iv_0, v_0, v_1, \ldots, v_{k-1}v_k, |V| < n\} \tag{4.1}$$

Ide o množinu vrcholov, ktoré sú od v_k vzdialené cestami dĺžok $P^1, \ldots, P^n$.

Susedstvá na mriežke

Ak uvažujeme o dvojrozmernom priestore ako o mriežke a vrcholy sú v nej pravidelne rozmiestnené (analógia s bunkami celulárneho automatu), mriežka generuje rôzne druhy okolia. Najznámejšie typy mriežkového okolia sú von Neumannove a Moorove okolie (bližšie sa im venujeme v časti 7.3).

[5] Definície z teórie grafov preberáme z práce [Diestel, 2005].

Susedstvo nad existujúcim grafom

Ak zoberieme existujúci neohodnotený graf $G(V, E)$ v ktorom sú vrcholy už prepojené priamym jednoduchým susedstvom, ohodnocovacia funkcia $f : V \times V \to \mathbb{R}$ nám vracia mieru podobnosti medzi ľubovoľnými dvoma vrcholmi, zohľadňujúc existujúcu štruktúru prepojení. Aplikovaním lokálnej ohodnocovacej funkcie tak získame pre zvolený vrchol $v_k \in V$ jeho širšie okolie, ktoré ku každému vrcholu asociuje váhu $w \in \mathbb{R}$. Nastáva tu situácia, kedy pomocou existujúceho základného okolia vrcholu získavame širšie okolie. Podrobnejšie sa lokálnym ohodnocovacím algoritmom venujeme v kap. 6.

4.1.2 Formalizácia homofílie

Nad daným grafom a zvoleným prístupom k susednosti medzi vrcholmi môžeme uvažovať o pravdepodobnosti, s akou sú si podobné susediace vrcholy. Ak uvážime sociologickú *definíciu* homofílie uvedenú na začiatku kapitoly:

> Ľudia, ktorí sú navzájom prepojení (vzťahom) sú si podobní s vyššou pravdepodobnosťou než ľudia, ktorí prepojení nie sú.

uvedenú definíciu môžeme v kontexte relačnej klasifikácie zovšeobecniť a formalizovať takto:

Definícia 4.2: Homofília. V grafe $G(V, E)$ má každý vrchol priradenú distribúciu príslušnosti k triede $p(c \in C|v \in V)$, kde C je množina tried klasifikácie. Hovoríme, že podobnosť distribúcií $p(c \in C|v_i)$ a $p(c \in C|v_j)$ je úmerná pravdepodobnosti existencie hrany $v_i v_j \in E$ medzi vrcholmi $v_i, v_j \in V$.

Ak máme napr. binárnu klasifikáciu s triedami $C = \{c_+, c_-\}$, a tri vrcholy s takouto distribúciu príslušnosti k triede:

- $p(c_+|v_1) = 1.0$, $p(c_-|v_1) = 0.0$,
- $p(c_+|v_2) = 1.0$, $p(c_-|v_2) = 0.0$,
- $p(c_+|v_3) = 0.0$, $p(c_-|v_3) = 1.0$,

potom pravdepodobnosť existencie hrany medzi vrcholmi v_1 a v_2 je vyššia než medzi vrcholmi v_1 a v_3, pretože vrcholy v_1 a v_2 patria do rovnakej triedy na rozdiel od vrcholu v_3. V procese klasifikácie sa často stretávame s tým, že distribúcia príslušnosti k triede je neostrá, t.j. $p(c \in C|v \in V) \in \mathbb{R}$, zvyčajne je normovaná na intervale $\langle 0, 1\rangle$.

Miera homofílie

Určiť mieru homofílie medzi dvoma vrcholmi na základe definície 4.2 znamená porovnať distribúciu príslušnosti k triede medzi týmito dvoma vrcholmi. Ak by sme sa obmedzili len na binárnu klasifikáciu, dostatočným spôsobom ako určiť rozdiel medzi dvoma distribúciami príslušnosti k triede vrcholov v_i a v_j je štvorec rozdielu medzi hodnotami distribúcie:

$$homophily(v_i, v_j) = \sqrt{[p(c_+|v_i) - p(c_+|v_j)]^2 + [p(c_-|v_i) - p(c_-|v_j)]^2} \tag{4.2}$$

Prakticky sa však stretávame s väčším množstvo tried, zovšeobecnene možno vzorec (4.2) zapísať ako:

$$homophily(v_i, v_j) = \sqrt{\sum_{c \in C} [p(c|v_i) - p(c|v_j)]^2} \tag{4.3}$$

Potom združená homofília vrcholu v_k v kontexte jeho okolia je priemer hodnôt z (4.3):

$$homophily_n(v_k) = \frac{1}{|V_k|} \sum_{v_j \in V_k} homophily(v_k, v_j) \tag{4.4}$$

Tento spôsob výpočtu – $homophily_n$ – nazývame vrcholovo-orientovaný (n ako node), kedy existuje ústredný vzťah centrálneho vrcholu voči všetkým vrcholom v susedstve.

Druhý pohľad na homofíliu je množinovo orientovaný, kedy na vrchol v_k a množinu jeho susedov V_k nazeráme bez ohľadu na *výnimočnosť* vrcholu v_k, čiže počítame homofíliu ako $homophily_s(V_k \cup \{v_k\})$ (s ako set). Tu môžeme agregovať distribúciu príslušnosti k triede pre každý vrchol z množiny $V_k \cup \{v_k\}$ napr. za pomoci entropie v (4.5) [Shannon *et al.*, 1998].

$$homophily_s(V_k \cup \{v_k\}) = 1.0 + \sum_{v_i \in V_k \cup \{v_k\}, c \in C} p(c|v_i) \log_{base} p(c|v_i) \quad (4.5)$$

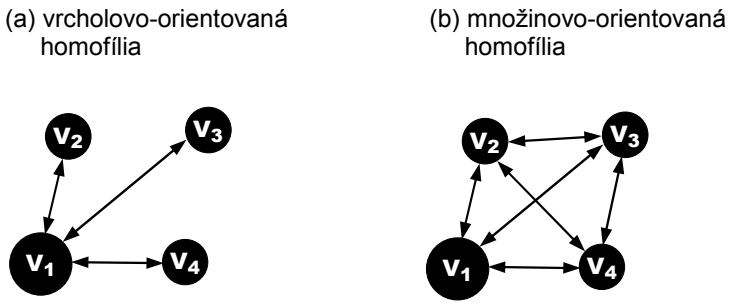
Špeciálny prípad $homophily_s$ predstavuje táto hodnota pre jediný vrchol:

$$homophily_s(v_k) = 1.0 + \sum_{c \in C} p(c|v_k) \log_{base} p(c|v_k) \quad (4.6)$$

kde hodnota $homophily_s(v_k)$ vyjadruje mieru súdržnosti distribúcie príslušnosti vrcholu k triede. Čím výraznejšie vrchol prislúcha k jednej triede (napr. $p(c_+|v_k) = 0.99, p(c_-|v_k) = 0.01$), tým viac sa jeho hodnota $homophily_s$ blíži k 1.0 (ak berieme logaritmus so základom 2). Naopak, čím menej určitá je trieda vrcholu (najhorší prípad $p(c_+|v_k) = p(c_-|v_k) = 0.5$), tým hlbšie klesá hodnota $homophily_s(v_k)$ k 0.0.

Porovnanie vrcholovo- a množinovo-orientovanej homofílie je takéto:

- $homophily_n$ dáva odpoveď na otázku:
  Ako sa na zvolený vrchol podobajú jeho susedia? (obr. 4.2(a)),
- $homophily_s$ dáva odpoveď na otázku:
  Ako sa na seba vzájomne podobajú vrcholy z danej množiny?(obr. 4.2(b)).



Obrázok 4.2: Dva spôsoby nazerania na homofíliu.

Ak by sme teda pre každý vrchol z množiny vrcholov $V = v_1, \ldots, v_m$ vypočítali $homophily_n$ a tieto hodnoty spriemerovali, dostaneme rovnakú informáciu, ako keď vypočítame $homophily_n(V)$ (numericky sa tieto dve hodnoty nezhodujú kvôli rozdielnemu škálovaniu a priebehu funkcie).

4.2 Stav poznania v prepojení homofílie a klasifikácie

V tejto časti vychádzame z dvoch prác, v ktorých sa autori venovali prepojeniu homofílie s klasifikáciou. Ide o jediné dve nám známe práce s touto tematikou, čo naznačuje, aká malá pozornosť sa doteraz venovala homofílii v relačnej klasifikácii.

4.2.1 Relačná autokorelácia

V práci [Jensen & Neville, 2002], ktorá sa primárne venuje analýze závislosti medzi mierou prepojení v dátovej vzorke a jej vplyvom na vyhodnocovanie výsledkov klasifikácie, autori definujú metriku na porovnanie dvoch množín vrcholov na základe hodnôt zvolených atribútov týchto vrcholov. Miera nazvaná *relačná korelácia* vyjadruje podobnosť medzi dvoma množinami vrcholov X a Y na úrovni atribútu f (pre vrcholy z X) a g (vrcholy z Y) a pre množinu ciest P spájajúcu X a Y:

Definícia 4.3: Relačná korelácia $C(X, f, P, Y, g)$ je korelácia medzi všetkými pármi $(f(x), g(y))$, kde $x \in X, y \in Y$ a $p(x, y) \in P$. ▶

Na základe definície cesty v grafe (def. 4.1) $p(x, y)$ predstavuje $E = xy$, teda cesty s dĺžkou P^1. Pre výpočet korelácie autori uvádzajú možnosť použiť bežné miery ako informačný zisk, χ-kvadrát či Pearsonov koeficient [Clarke & Cooke, 1998].

Uvedená definícia je pomerne všeobecná, v práci [Jensen & Neville, 2002] sa uvádza ešte ďalšia miera, ktorá predstavuje špeciálny prípad relačnej korelácie.

Definícia 4.4: Relačná autokorelácia C' je $C'(X, f, P) \equiv C(X, f, P, Y, g)$, kde $\forall p(x_i, x_j) \in P, x_i \neq x_j$. ▶

Relačná autokorelácia vyjadruje podobnosť medzi inštanciami z jednej množiny pri zohľadnení iba jedného atribútu. Z pohľadu homofílie v kolektívnom usudzovaní tento náhľad možno reprezentovať takto: atribút f je

príslušnosť vrcholu k triede, a X je množina vrcholov, ktoré sú prepojené hranami určenými v P.

Menšia komplikácia výpočtu autokorelácie príslušnosti k triede pre zvolenú množinu vrcholov nastáva, keď máme viac ako dve triedy. V prípade binárnej klasifikácie je postačujúce príslušnosť k triede vyjadrovať jednou hodnotou (pretože $p(c_+|v) = 1.0 - p(c_-|v)$) a môžeme použiť niektorú zo základných korelačných metód. Pri troch a viac triedach je potrebné uvažovať o korelačnej matici $n - 1$ náhodných premenných (n je počet tried).

Z pohľadu na homofíliu uvedenom v kap. 4.1.2 ide o analógiu k množinovo-orientovanému prístupu $homophily_s$.

4.2.2 Homofília v generovaných grafoch

Druhá práca, ktorá sa venuje homofílii a zároveň klasifikácii v grafoch je výskum [Jackson, 2008]. Skúma sa tu vzťah medzi homofíliou grafu a jeho typickými charakteristikami ako priemer grafu, klasterizačný koeficient, atď. Napriek tomu, že v práci sa autor vôbec nezmieňuje o klasifikácii, zavádza zatrieďovanie vrcholov do tzv. *skupín* resp. *typov*, čo zodpovedá bežnej predstave klasifikácie.

Autor stanovuje klasifikáciu ako rozdelenie množiny vrcholov $V = \{v_1, v_2, \ldots, v_n\}$ do k podmnožín $V_1, \ldots, V_k$, a predpokladá, že vrcholy, ktoré sú si príbuzné (angl. same characteristics) sú v jednej skupine. Na základe uvedených príkladov ako vek, vzdelanie, či zamestnanie to zodpovedá našej predstave podobnosti na základe atribútov vrcholov. V nasledujúcich odstavcoch budeme o týchto skupinách a typoch hovoriť ako o triedach a množinou V_k rozumieme triedu c_k.

Homofília je v práci [Jackson, 2008] zavedená pomocou tzv. zoznamu stupňov vrcholov (angl. degree sequence) a tzv. relatívnej náklonnosti medzi triedami (angl. relative proclivity). Zoznam stupňov vrcholov je postupnosť $\{d_1, \ldots, d_n\}$, kde d_i zodpovedá stupňu vrchola v_i [Chung *et al.*, 2002].

$$D_k = \sum_{i \in V_k} d_i \tag{4.7}$$

zodpovedá súčtu stupňov vrcholov z triedy c_k.

Relatívna náklonnosť medzi dvoma triedami c_i a c_j označovaná ako $h_{c_i c_j} > 0$ vyjadruje mieru, s akou vrcholy patriace k jednej a druhej triede vytvárajú hrany medzi týmito dvoma triedami[6].

Platí, že $\sum_{c_j} D_{c_j} h_{c_i c_j} = D$ pre každú triedu c_i. Potom hrana medzi vrcholmi $v_e \in c_i$ a $v_f \in c_j$ vzniká s pravdepodobnosťou podľa (4.8).

$$\frac{h_{c_i c_j} d_i d_j}{D} \leq 1.0 \tag{4.8}$$

Homofília je potom definovaná takto:

Definícia 4.5: Homofília podľa Jacksona. Majme dvojice tried c_i, c_j tak, že $c_i \neq c_j$. Ak platí $h_{c_i c_i} > h_{c_i c_j}$, potom hovoríme, že v rámci triedy c_i je v grafe prítomná homofília. ▶

Predpoklad nerovnosti pre relatívnu náklonnosť ($h_{c_i,c_i} > h_{c_i,c_j}$) znamená, že pravdepodobnosť vzniku hrany je vyššia medzi vrcholmi rovnakej triedy, ako medzi vrcholmi rôznych tried. Ak $h_{c_i,c_j} = 1$ pre všetky c_i, c_j, potom triedy nedávajú o grafe žiadnu dodatočnú informáciu.

V kontexte nami definovanej homofílie (def. 4.2) homofília podľa Jacksona zodpovedá množinovo-orientovanému prístupu $homophily_s$. Autor priamo neuvádza spôsob ako homofíliu vyčísliť.

4.3 Doplňujúce poznatky o homofílii

V tejto časti zhŕňame dodatočné poznatky a úvahy o homofílii v grafe. Uvádzame tu niekoľko teoretických vlastností uvedených mier, diskutujeme o ich obmedzeniach a uvádzame aj diskusiu o heterofílii.

4.3.1 Homofília ako jedna z príčin korelácie v sociálnej sieti

V práci [Anagnostopoulos *et al.*, 2008] sú analyzované príčiny, ktoré spôsobujú koreláciu medzi správaním používateľa a jeho zaradením v sociálnej

[6] V práci [Jackson, 2008] nie je stanovené, či má parameter h horné ohraničenie, Ak nepresiahne hodnotu 1.0, môžeme hovorit' o pravdepodobnosti vzniku hrany.

sieti (atribútmi, príslušnosťou ku skupine). Boli identifikované tri príčiny:

- *vplyv* (angl. influence), kde je správanie používateľa ovplyvnené správaním jeho priateľov, napr. pri kúpa knihy z dôvodu, že si ju kúpil aj kamarát.
- *(mätúce) prostredie* (angl. confounding environment) – vonkajšie vplyvy ktoré sa premietajú do sveta sociálnej siete, napr. kúpa priatelia v sociálnej siete si obaja kúpia knihu o určitom meste z dôvodu, že obaja v tomto meste žijú.
- *homofília* – vychádza z [Mcpherson *et al.*, 2001], od predošlej príčiny ju odlišuje to, že homofília zahŕňa iba nadväzovanie priateľstva medzi používateľmi, kým vplyv *prostredie* agreguje nadväzovanie priateľstva a ostatné akcie používateľov (napr. kúpa knihy).

V práci [Anagnostopoulos *et al.*, 2008] je ďalej homofília zovšeobecnená spolu s prostredím do jedného modelu. Práca sa venuje najmä dynamike v sociálnej sieti, konkrétne schopnosti siete reagovať na prvotný impulz.

4.3.2 Ekvivalencia pre $homophily_s$

V tejto časti uvádzame náš postreh ohľadom existencie zhodných úrovní množinovo-orientovanej homofílie v grafe.

Definícia 4.6: Najkratšia vzdialenosť medzi vrcholmi.
Majme dva vrcholy v_i a v_j v grafe G, medzi ktorými je k ciest $P_k(V,E)|V = v_i, v_j, E = v_i, \ldots, v_j$. Potom $d_G(v_i, v_j)$ označuje najkratšiu vzdialenosť medzi vrcholmi v_i a v_j v grafe G, teda $d_G(v_i, v_j)$ zodpovedá ceste $P_k = argmin_m[P_k^m]$. ▶

Definícia 4.7: Priemer grafu, označovaný $rad\, G$ je najväčšia vzdialenosť medzi ľubovoľnými dvoma vrcholmi v_i, v_j, takže
$rad\, G = argmax_{d_G(v_i,v_j)}[G(V,E), v_i, v_j \in V]$ ▶

Je očividné, že $homophily_s(\{v_i\} \cup V_i, c)$ je zhodná pre všetky vrcholy v_i, v_j, pre ktoré $\{v_i\} \cup V_i = \{v_j\} \cup V_j$, čomu zodpovedajú všetky grafy s priemerom $rad\, G = 1$.

4.3.3 Obmedzenia uvedených mier

V nami uvedenej definícii homofílie (def. 4.2) rovnako ako v nadväzujúcich mierach predpokladáme, že všetky vlastnosti vrcholu sú *zapuzdrené* práve do príslušnosti k triede, čo zodpovedá predstave sveta klasifikačných metód s kolektívnym usudzovaním alebo priamemu relačnému prístupu. Ak by sme chceli v rámci homofílie klasifikovaného grafu pokryť tie skupiny klasifikačných metód, ktoré pracujú aj s vlastnými atribútmi vrcholov, nevyhnutne by sme museli porovnávať hodnoty atribútov. Tu je ťažké predstaviť si všeobecný spôsob ako vypočítať homofíliu, jednak rôzne atribúty vrcholu vplývajú odlišným spôsobom na jeho cieľovú triedu, navyše mieru na porovnanie dvoch rôznych hodnôt nominálneho atribútu možno stanoviť častokrát len empiricky (napr. ako číselne porovnať hodnoty `stredoškolské` a `vysokoškolské` atribútu `vzdelanie`, pozri diskusiu o atribútoch v kap. 3.2.1).

4.3.4 Vznik väzby medzi vrcholmi

Ak sa vrátime k pôvodnej sociologickej definícii:

> Ľudia, ktorí sú navzájom prepojení (vzťahom), sú si podobní s vyššou pravdepodobnosťou ako ľudia, ktorí prepojení nie sú.

Táto intuitívne dáva zmysel aj keď zameníme príčinu a dôsledok:

> Medzi ľuďmi, ktorí sú si podobní nastáva vzťah(kontakt) s vyššou pravdepodobnosťou ako pri ľuďoch, ktorí sú si odlišní.

Dôvodom, pre ktorý je pre nás výhodnejšie používať prvú formuláciu, je stav, v akom sa k nám dáta dostávajú pri klasifikácii. Ak by sme uvažovali o druhej formulácii, vidíme, že sa hovorí o *vzniku* vzťahu a homofília je jav, ktorý tieto vzťahy *generuje*. V našom prípade však už vzťahy (hrany) existujú. V dátovej vzorke, ktorá je určená na klasifikáciu, sú už pevne stanovené. V čase, keď hrany vznikali (napr. počas pridávania nových členov predstavenstva do vedenia spoločnosti zachytených v dátovej vzorke foaf.sk v prílohe B) bola skutočne prítomná sociologická homofília, kým

v čase použitia nami stanovených mier už len môžeme konštatovať, nakoľko kohézne väzby vznikli. Predikovať pomocou definovaných mier homofílie pravdepodobnosť, s akou budú vznikať v dátovej vzorke nové väzby, je síce možné (a aj štatisticky vyhodnotiteľné, ak je zachytená aj časová následnosť vzniku väzieb), ale nie je to zámerom tejto práce, pretože aj keď ide o relačnú klasifikáciu, predikuje sa vždy atribút (trieda) a nie relácia.

4.3.5 Heterofília – opak homofílie

Homofília je všadeprítomný spoločenský jav. Ak na vrcholy grafu, v ktorom tento jav sledujeme, nazeráme ako na súperiacich jedincov, z hľadiska jednotlivca je spájanie sa s vrcholmi tak, aby boli preferované homofílne väzby, pomerne dobrá stratégia. Napríklad, ak sa mladý človek rozhodne, že bude robiť tú istú prácu ako jeho rodičia, je pravdepodobné, že vďaka informáciám a kontaktom, ktoré od rodičov získa, dokáže dosiahnuť primeranú životnú úroveň s menším rizikom ako keby sa rozhodol pracovať v oblasti, v ktorej mu nemá kto poskytnúť cenné informácie znižujúce riziko.

Uvedená stratégia je konzervatívna, pretože znižuje pravdepodobnosť, že si jedinec osvojí inovatívny prístup (napr. sa ako prvý v meste naučí robiť isté remeslo a vďaka tomu sa jeho životná úroveň podstatne zvýši). Sú aj iné stratégie, v práci [Rogers *et al.*, 2003] sa uvádza, že heterofílne vzťahy môžu priniesť rýchlejšie osvojovanie inovácií.

V tejto kapitole sme sa v úvode venovali homofílii z hľadiska sociológie a na základe tohto poznania sme definovali homofíliu v kontexte klasifikácie nad grafmi. Ukázali sme, že homofília sa v klasifikovaných dátových vzorkách vyskytuje, ale *naivne* predpokladať jej prítomnosť nie je vhodné a bezpečné riešenie.

V nadväzujúcich dvoch kapitolách je naším cieľom prepojiť znalosti o konštrukcii relačných klasifikačných metód s poznaním ako merať homofíliu v klasifikovanom grafe. V nadväznosti na hypotézy z časti 1.2 je naším cieľom navrhnúť takú metódu, ktorá bude robustná v zmysle zohľadnenia homofílie vrcholov grafu a dokáže jej meniacu sa úroveň využiť v prospech výsledku klasifikácie.

Kapitola 5

Návrh metódy moderovania príslušnosti ku triede

V tejto kapitole sa venujeme prvej nami stanovenej hypotéze z kap.1.2:

> Zdieľat' pri relačnej klasifikácii menej informácií (z pohľadu prepojených vrcholov) je prospešné pre výslednú kvalitu zatriedenia.

V kontexte relačnej klasifikácie je vhodné zamerat' sa na tú skupinu metód, v ktorej je zdieľanie informácií najvýraznejšie – kolektívne usudzovanie. Pri tejto skupine metód má klasifikátor, z pohľadu informácií o inštancii, k dispozícii len jej príslušnost' k triede. Hypotéza potom znie:

> Ak počas iteratívnej výmeny informácií v klasifikátore využívajúcom kolektívne usudzovanie budeme uprednostňovat' inštancie s lepšie vyhranenou príslušnost'ou k triede (teda moderovat' výmenu informácií), zvýšime tým kvalitu klasifikácie.

Uvedená hypotéza je ešte stále pomerne široká, dôvodom je naša snaha navrhnút' metódu na overenie hypotézy tak, aby mohla byt' zapojená do ľubovoľného už existujúcej klasifikačnej metódy typu $\hat{c}_{ci}$. Experiment, ktorý má overit' hypotézu, je navrhnutý takto: pre zvolený graf inštancií a vzt'ahov medzi nimi porovnávame tri rôzne klasifikačné prístupy, každý s rovnakými dátami a zvolenými triedami:

- ($\hat{c}_a$) atribútovo-viazané zatrieďovanie,
- ($\hat{c}_{ci}$) relačný prístup s kolektívnym usudzovaním[1],
- ($\hat{c}_{ci-m}$) nový relačný prístup s kolektívnym usudzovaním a moderovaním výmeny informácií.

Očakávame, že ak je zvolená dátová vzorka dostatočne previazaná (relačná), výsledok klasifikácie (v zmysle jej správnosti) bude najkvalitnejší v prístupe $\hat{c}_{ci-m}$ a prístup $\hat{c}_{ci}$ bude presnejší než $\hat{c}_a$. Voľba konkrétnych metód $\hat{c}_a$ a $\hat{c}_{ci}$ je v časti 5.2, pretože v zmysle metodológie CRISP-DM nadväzuje na analýzu dátovej vzorky použitej pri experimente.

5.1 $\hat{c}_{ci-m}$: moderovanie výmeny informácií

Pri výmene informácií zavádzame tzv. moderovanie ako prístup na obmedzenie množstva zdieľaných údajov medzi inštanciami pri kolektívnom usudzovaní. Využijeme tu $homophily_s(v_k)$, ktorý nám podľa vzorca 4.6 určuje ako výrazne inštancia v príslušnosti k triede preferuje určitú triedu.

Ak uvažujeme o binárnej klasifikácii v iterácii t (v ľubovoľnej klasifikačnej metóde s kolektívnym usudzovaním), kde $p(c_+|v_i) = (c_-|v_i) = 0.5$, vidíme, že príslušnosť vrcholu v_i k obom pólom klasifikácie je rovnaká, jej hodnota $homophily_s(v_i) = 0.0$ je najnižšia možná. V kontexte hypotézy je naším cieľom túto inštanciu dočasne *odstavit'* z procesu odovzdávania jej príslušnosti k triede susedom, zároveň však inštancii ponechať schopnosť absorbovať informáciu od susedov tak, aby sa v čo najmenšom počte nasledujúcich iterácií hodnota $homophily_s(v_i)$ zvýšila.

Opačný prípad predstavuje inštancia s príslušnosťou k triede s hodnotami $p(c_+|v_j) = 1.0$, $(c_-|v_j) = 0.0$, ktorá má najvyššiu možnú hodnotu $homophily_s(v_j) = 1.0$. Pri tejto entite naopak určite chceme zachovať jej schopnosť odovzdávať svoju informáciu. Väčšina vrcholov sa obvykle nachádza niekde medzi najnižšou a najvyššou možnou hodnotou ich homofílie. Tu sa experimentálne pokúšame zistiť, aká je vhodná hodnota $homophily_s(v_j)$, pri ktorej ešte ponecháme inštanciám možnosť zdieľať príslušnosť k triede. Túto hodnotu nazývame *úroveň moderácie*.

[1] Presnejšie ide o $\hat{c}_{ci} \leftarrow \hat{c}_r \leftarrow \hat{c}_a$.

Samotnú metódu na moderovanie informácií vieme formálne zapísať ako rozšírenie existujúcej metódy kolektívneho usudzovania. Na základe analýzy v časti 5.2 sa použije metóda IRC. Pseudokód tejto metódy z časti 3.2.4 je potom pri rozšírení na $\hat{c}_{ci-m}$ uvedený v algoritme 5.

Algoritmus 5 Moderovanie príslušnosti k triede v metóde IRC

Vstupná podmienka: Inštancie $v_1, v_2, \ldots, v_k \in X_{tst}$, zvolený prah homofílie $homophily_s$.

Výstupná podmienka: Každá inštancia má priradenú triedu $\forall v_k \in V : trieda(v_k) = c \in C$.

1: Na základe vlastných atribútov inštancie sa pomocou atribútového klasifikátora $\hat{c}_a$ inicializuje príslušnosť k triede pre každý vrchol v_k.

2: Pre každý vrchol:
   1. príslušnosť k triede sa upraví na základe vlastnej distribúcie,
   2. príslušnosť k triede sa upraví podľa výpočtu relačnej zložky metódy IRC, pričom vo výpočte akceptujeme iba tie susedné vrcholy, ktorých hodnota $homophily_s$ v iterácii presahuje zvolený prah.

3: Krok 2 sa opakuje až kým rozdiel v zmene distribúcií pre všetky vrcholy medzi dvoma iteráciami neklesne pod určitú hranicu.

4: Vrchol v_k sa označí triedou, ktorá v príslušnosti k triede dominuje, teda $c = argmax_{c_i}[p(c_i|v_k)]$.

5.2 Dátová vzorka a zvolené klasifikačné metódy

Dátová vzorka zvolená pre náš experiment pochádza z projektu MAPEKUS[2] (Modeling and Acquisition, Processing anf Employing Knowledge About User Activities in the Internet Hyperspace, [Frivolt *et al.*, 2008]), na vzniku ktorej sa podieľal aj autor práce. Ide o údaje z portálu vedeckých publikácií ACM[3].

Získaný graf má tri typy inštancií:

[2]Viac informácií o projekte MAPEKUS v prílohe A.

[3]Association for Computing Machinery: `http://www.acm.org/dl`.

- klasifikovaný typ `publikácia`,
- dva doplňujúce typy inštancií: `autor` a `kľúčové slovo`.

Medzi uvedenými inštanciami sú dva typy hrán vyjadrené inter-reláciami: `jeAutor` a `máKľúčovéSlovo`. Hrany sú neorientované, t.j. informácie môžu hranou plynúť oboma smermi, vďaka čomu je riešenie všeobecnejšie.

Váha každej hrany je nastavená bezrozdielne na $w(v_i, v_j) = 1.0$. Hrany reprezentované intra-reláciami v tomto experimente neuvažujeme. Graf s uvedenými vlastnosťami sme poloautomaticky extrahovali z dátovej vzorky MAPEKUS, ktorá samotná obsahuje omnoho viac relácií, tried, atribútov a typov inštancií. Rozmery nášho grafu sú: 4 000 inštancií publikácií, 7 600 kľúčových slov a 9 700 autorov, spolu 21 300 vrcholov a 35 000 hrán.

Na základe vlastností dátovej vzorky sme ako atribútovo viazanú metódu $\hat{c}_a$ zvolili prístup založený na Bayesovom teoréme, pretože na zatrieďovanie inštancií `publikácií` použijeme vektor slov vytvorený z textu abstraktu (tento je v anglickom jazyku). Tento prístup patrí medzi najčastejšie používané na klasifikáciu textu [McCallum & Nigam, 1998, Mitchell, 1997]. Abstrakt publikácie sa rozdelí na slová (tokenizuje), tieto sa transformujú na základný tvar pomocou Porterovej metódy [Porter, 1980] a odstránia sa stop-slová.

IRC metódu (uvedená v kap. 3.2.3) sme zvolili ako $\hat{c}_{ci}$ a tiež ako základ pre $\hat{c}_{ci-m}$, pretože metóda IRC dáva možnosť meniť vplyv jednotlivých zložiek prepočtu príslušnosti inštancie k triede v každej iterácii. Konkrétne, relačná zložka IRC metódy obsahuje parametre λ_1, λ_2 a λ_3, ktoré vplývajú na relačnú zložku metódy IRC takto:

$$p(c_m|v_k) = \lambda_1 p(c_m|v_k) + \lambda_2 \frac{\sum_{v_z \in V_k \cap X_{tr}} w(v_k, v_z) p(c_m|v_z)}{|V_k \cap X_{tr}|} + \\ + \lambda_3 \frac{\sum_{x_z \in V_k \cap X_{tst}} w(v_k, v_z) p(c_m|v_z)}{|V_k \cap X_{tst}|} \tag{5.1}$$

kde λ_1 určuje váhu vlastnej príslušnosti k triede z predošlej iterácie,

λ_2 a λ_3 určujú význam susedov z trénovacej a testovacej množiny. Ďalšou výhodou metódy IRC je jej zaručená konvergentnosť.

5.3 Triedy klasifikácie a spôsob vyhodnotenia

Hlavnému typu `publikácia` je priradená jedna alebo viac tried klasifikácie ACM[4]. Zvolený prístup neumožňuje priamu viactriednu klasifikáciu [Ghamrawi & McCallum, 2005] (viac informácií o viactriednej klasifikácii je v kap. 7.1), preto je úloha rozdelená na n binárnych klasifikácií, kde pre každú zvolenú triedu a inštanciu sa vyhodnocuje binárna príslušnosť, ktorá je buď pozitívna (c_+) alebo negatívna (c_-).

V experimente porovnávame jednotlivé klasifikačné metódy na základe tzv. *zisku správnosti* (angl. accuracy gain). Správnosť sme zvolili z dôvodu schopnosti zachytiť správnosť klasifikácie rovnako pre pravdivo pozitívne aj pravdivo negatívne inštancie.

Zisk správnosti vyjadruje rozdiel medzi správnosťou atribútovo viazaného prístupu $\hat{c}_a$ a relačného klasifikátora ($\hat{c}_{ci}$ a $\hat{c}_{ci-m}$) na rovnakej dátovej vzorke X, t.j. sledujeme vývoj dvoch indikátorov:

- $accuracy\_gain(\hat{c}_{ci}, \hat{c}_a, X) = accuracy(\hat{c}_{ci}, X) - accuracy(\hat{c}_a, X)$;
- $accuracy\_gain(\hat{c}_{ci-m}, \hat{c}_a, X) = accuracy(\hat{c}_{ci-m}, X) - accuracy(\hat{c}_a, X)$;

Pomer inštancií v trénovacej a testovacej množine je v pomere 1 : 1. Uvádzané výsledky sú priemerované z 200 behov.

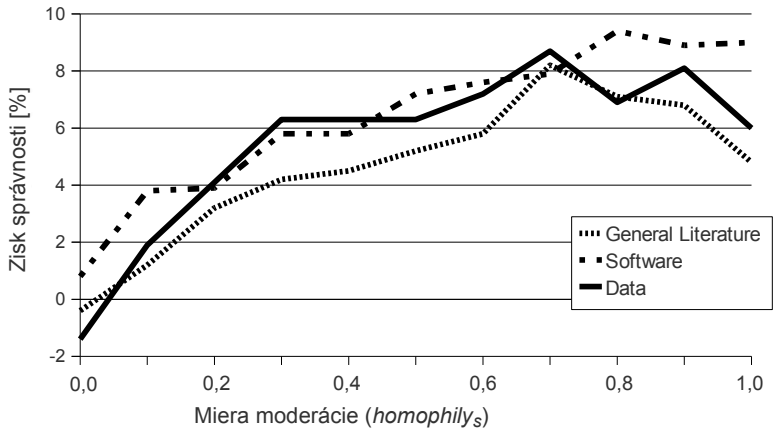
5.4 Vplyv moderácie na zisk správnosti

Úroveň moderácie sme zaviedli s cieľom zvýšiť správnosť klasifikátora. Preto sme vykonali sériu experimentov, kde sme menili prah homofílie v intervale $\langle 0.0, 1.0 \rangle$. Hodnota 0.0 zodpovedá pôvodnej metóde IRC ($\hat{c}_{ci}$). Zvyšovaním úrovne moderácie obmedzujeme výmenu informácií v grafe,

[4]Klasifikačný systém ACM: `http://www.acm.org/class/`.

aktívne sa na relačnej klasifikácii môže podieľať čoraz menej inštancií. Nastavenie prahu na $homophily_s = 1.0$ znamená, že v grafe sú aktívne iba inštancie s výhradne pozitívnou alebo výhradne negatívnou príslušnosťou, teda prakticky iba inštancie z trénovacej množiny X_{tr}. Len tieto majú úplne pozitívne ($p(c_+|v_i) = 1.0$ a $p(c_-|v_i) = 0.0$) alebo úplne negatívne ($p(c_+|v_i) = 0.0$, $p(c_-|v_i) = 1.0$) distribuovanú príslušnosť k triede.

Experiment sme vykonali s tromi rôznymi prvostupňovými triedami ACM: `General literature`, `Software` a `Data`. Parametre metódy IRC λ_1, λ_2 and λ_3 sme nastavili zhodne na $\frac{1}{3}$, čím sme určili rovnakú váhu pre všetky zložky. Obr. 5.1 zobrazuje výsledky experimentu. Os x znázorňuje meniacu sa hodnotu moderácie, os y vyjadruje zodpovedajúci zisk správnosti.



Obrázok 5.1: Vplyv moderácie na zisk správnosti. Rôzne triedy ACM.

Pri všetkých troch skúmaných triedach pozorujeme podobné správanie klasifikátora. So silnejúcou moderáciou vzrastá zisk správnosti. Trend dosiahne maximum, keď je hodnota $homophily_s$ medzi 0.7 a 0.9. Nasledujúci pokles zisku správnosti pri hodnote 1.0 ukazuje, že inštancie z testovacej množiny majú z hľadiska zisku správnosti významnú úlohu (tieto inštancie sú z procesu výmeny informácií vynechané v silne moderovanom prípade $homophily_s = 1.0$).

Uvedený počiatočný experiment ukázal význam moderovania výmeny

informácií v grafe. Pre porovnanie, nemoderovaný relačný klasifikátor (hodnota $homophily_s = 0.0$ na obr. 5.1) dosiahol výrazne slabšie, dokonca až záporné hodnoty zisku správnosti (ide teda o *stratu správnosti*, -1.4% pre triedu `Data`).

5.5 Optimálny pomer vplyvu trénovacej a testovacej množiny

V nasledujúcom experimente analyzujeme účinky parametrov λ na metódu IRC. Z pohľadu každej konkrétnej inštancie ide o tieto parametre:

- λ_1: váha vlastnej príslušnosti k triede (z predošlej iterácie),
- λ_2: váha príslušností k triede tých susedných vrcholov, ktoré sú z trénovacej množiny,
- λ_3: váha príslušností k triede tých susedných vrcholov, ktoré sú z testovacej množiny.

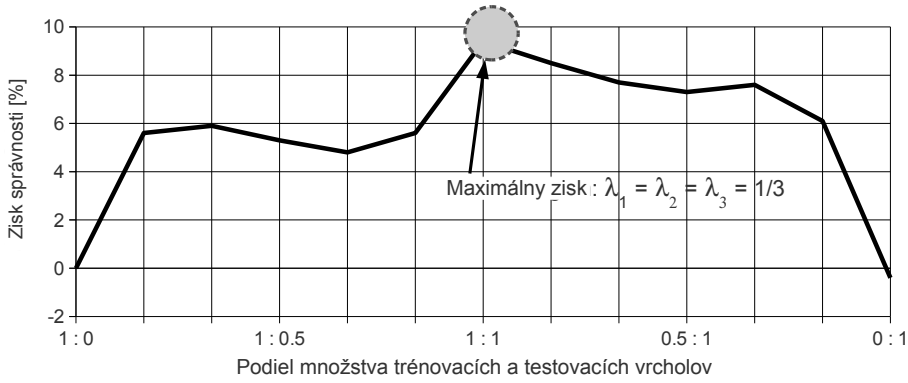
Váhu vlastnej príslušnosti k triede sme stanovili ako konštantnú, $\lambda_1 = \frac{1}{3}$ a v experimente meníme hodnotu λ_2 v intervale medzi 0 a $\frac{2}{3}$. Súčasne hodnotu λ_3 určujeme ako $\lambda_3 = 1.0 - \lambda_1 - \lambda_2$. Pomer vplyvu trénovacej a testovacej množiny sledujeme cez pomer $\lambda_2 : \lambda_3$. Napríklad, ak $\lambda_2 = 0.66$ a $\lambda_3 = 0.0$, pomer je $1 : 0$, teda trénovacia množina má maximálny vplyv a inštancie z testovacej množiny klasifikátor ignoruje.

Na základe uvedenej konfigurácie sme stanovili hypotézu:

> Najnižší a najvyšší pomer vplyvu trénovacej a testovacej množiny ($1 : 0$ alebo $0 : 1$) nedosiahne najvyšší možný zisk správnosti, pretože klasifikátor má k dispozícii príliš obmedzený podgraf, keďže sa buď trénovacie alebo testovacie inštancie úplne ignorujú.

Výsledky experimentu pre triedu ACM `Software` sú uvedené na obr. 5.2, hodnota moderácie je nastavená na $homophily_s = 0.8$. Najvyššia hodnota zisku správnosti je dosiahnutá pre $\lambda_2 = \lambda_3 = \frac{1}{3}$, teda keď je trénovacej aj testovacej množine priradená rovnaká váha. Výsledok potvrdzuje našu

hypotézu, t.j. spodná a horná hranica pomeru váhy množín (1 : 0 and 0 : 1) majú nízky zisk správnosti.



Obrázok 5.2: Hľadanie optimálneho pomeru vplyvu trénovacej a testovacej množiny ($\lambda_2 : \lambda_3$).

Z grafu vidieť, že klasifikačný model prináša vyšší zisk správnosti vtedy, keď majú testovacie vrcholy vyšší vplyv než trénovacie vrcholy. Pomer počtu trénovacích a testovacích vrcholov typu `publikácia` je v grafe rovnaký (1 : 1), avšak v grafe sú ešte vrcholy zvyšných dvoch typov vrcholov (`autor` a `kľúčové slovo`). Tieto dva typy sú významovo priradené ku testovacím údajom, čím robia túto množinu početnejšou. Preto, ak je uprednostnená trénovacia množina (ľavá časť grafu), väčšia časť vrcholov má priradenú menšiu váhu, čiže zdieľanie a šírenie informácií v grafe je viac obmedzené. Inými slovami, ku testovacím vrcholom typu `publikácia`, z ktorých sa vypočítava výsledný zisk správnosti, sa informácia neprešíri z preferovaných trénovacích údajov, pretože medzi nimi môžu tvoriť *bariéru* menej preferované vrcholy typu `autor` a `kľúčové slovo`.

5.6 Kvalitatívny vplyv relácií

V predošlých experimentoch sme využívali celý graf s inštanciami `publikácia`, `autor` a `kľúčové slovo`, prepojenými cez oba typy hrán `jeAutor` a `máKľúčovéSlovo`.

Predpokladáme však, že jednotlivé typy hrán sa podieľajú na výslednom zisku správnosti nerovnakým dielom, keďže vykazujú rôznu mieru homofílie. Preto v nadväzujúcom experimente porovnávame zisk správnosti pre tri rôzne grafy, pričom moderáciu nastavíme na $homophily_s = 0.8$ a klasifikujeme podľa triedy `Software` tieto grafy:

- graf s publikáciami a kľúčovými slovami, prepojenými cez reláciu `máKľúčovéSlovo`,
- graf s publikáciami a autormi, prepojenými cez reláciu `jeAutor`,
- spojený graf s publikáciami, autormi a kľúčovými slovami (`jeAutor`+`máKľúčovéSlovo`).

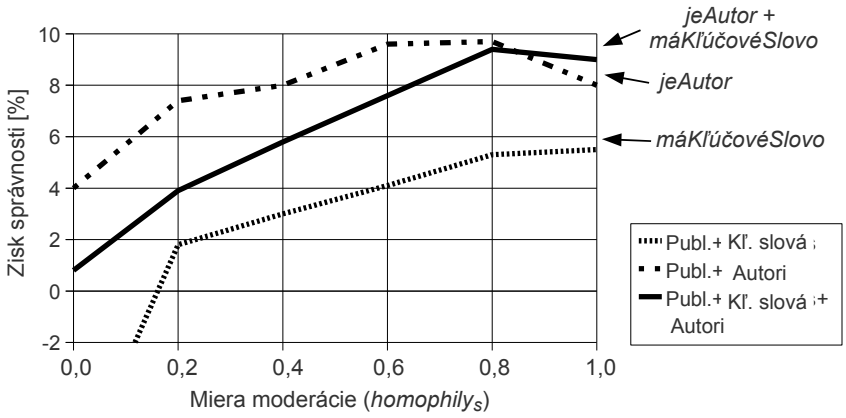
Naša hypotéza je:

> Kvalita výsledku klasifikácie bude najvyššia, keď budú v grafe združené oba typy hrán, keďže vtedy bude mať klasifikátor k dispozícii najviac relevantných informácií.

Výsledky experimentu znázorňuje obr. 5.3. Krivka priebehu zisku správnosti pre $homophily_s \in \langle 0.0, 0.8\rangle$ naznačuje, že naša hypotéza tu nie je správna, pretože zisk správnosti grafu s typom hrany `jeAutor` predčuje graf obsahujúci oba typy hrán (`jeAutor`+`máKľúčovéSlovo`). Hypotéza je splnená len pre graf s typom hrany `máKľúčovéSlovo`, ktorý je z hľadiska zisku správnosti prekonaný oboma grafmi.

Strata správnosti v grafe tvorenom oboma typmi hrán (`jeAutor` + `máKľúčovéSlovo`) je spôsobená ich odlišnou mierou homofílie (v množinovom zmysle). Vzťah `jeAutor` vytvára výraznejšie homofílny vzťah než hrana `máKľúčovéSlovo`, t.j. publikácie od jedného autora prislúchajú do rovnakej triedy s vyššou pravdepodobnosťou než publikácie, ktoré sú viazané na konkrétne kľúčové slovo.

Obr. 5.4(a) znázorňuje situáciu, kde má `autor` (A vnútri kružnice) päť publikácií (P vnútri kružnice), pričom štyri z nich sú pozitívne príklady triedy a jedna je negatívna v zmysle príslušnosti k triede. Uvedené zoskupenie je zjavne silno homofílne a autorovi A zabezpečuje dobre vyhranenú príslušnosť k triede – z pohľadu relačného klasifikátora to môže napr. znamenať, že autorova príslušnosť k triede skonverguje k hodnotám $p(c_+|A) = 0.95$ a $p(c_-|A) = 0.05$.



Obrázok 5.3: Kvalitatívny vplyv relácií na klasifikátor.

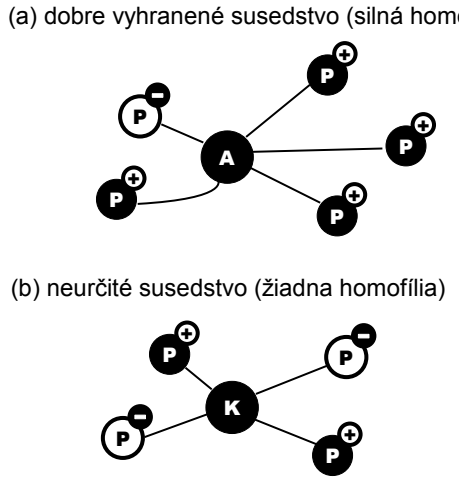
Na druhej strane, obr. 5.4(b) vyjadruje situáciu, kde vrchol `kľúčové slovo` (K vnútri kružnice) má štyri publikácie, dve z nich pozitívne a dve negatívne z hľadiska príslušnosti k triede. Uvedená konštalácia skonverguje do medzistavu, kde príslušnosť sledovaného vrcholu `kľúčové slovo` k triede poskytuje najmenšiu informačnú hodnotu, t.j. $p(c_+|K) = 0.5$ a $p(c_-|K) = 0.5$

Histogram miery homofílie pre typy hrán `jeAutor` a `máKľúčovéSlovo` je na obr. 5.5. Histogram je uvedený z pohľadu autorov a kľúčových slov, keďže oba typy inštancií majú ako susedov inštancie typu `publikácia`. Iba pri type `publikácia` vieme, aká je skutočná príslušnosť vrcholu k triede a tak vieme vyhodnotiť mieru homofílie.

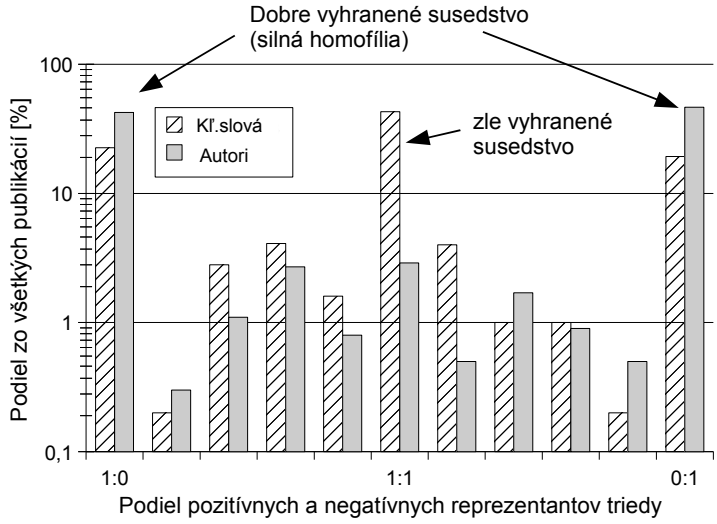
Podiel pozitívnych a negatívnych reprezentantov triedy v rámci susedstva vyjadruje na obr. 5.5 na osi x mieru homofílie. Najsilnejšia homofília je pre podiel $1:0$ alebo $0:1$. Až 89% vrcholov typu `autor` a iba 42% vrcholov typu `kľúčové slovo` má maximálne homofílne susedstvo. Naopak, najviac heterofílne susedstvo (vyjadrené pomerom $1:1$) majú iba 3% inštancií typu `autor` a až 43% inštancií typu `kľúčové slovo`.

Keď sa vrátime k výsledkom experimentu na obr. 5.3, vidíme, že naša pôvodná hypotéza platí[5] pre hodnotu moderácie $homophily_s \in \langle 0.8, 1.0 \rangle$.

[5]Kvalita klasifikácie bude najvyššia, keď sú v grafe prítomné oba typy hrán.



Obrázok 5.4: Príklady rôzne homofílneho susedstva v grafe.



Obrázok 5.5: Distribúcia homofílie pre autorov a kľúčové slová.

Dôvodom je účinok silnej moderácie – väčšina heterofílnych vrcholov je eliminovaných z výmeny informácií v grafe (pre `autorov` aj `kľúčové slová`).

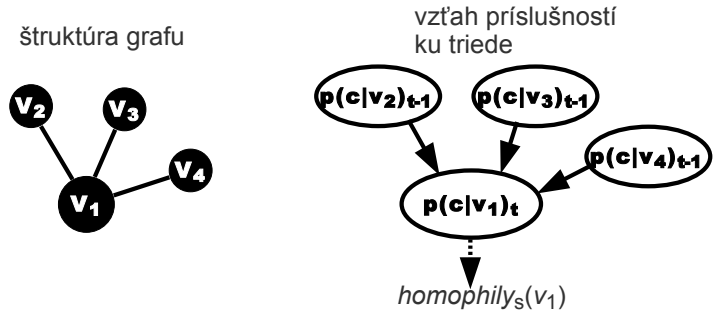
5.7 Diskusia

Ak sa vrátime k prvej hypotéze z časti 1.2:

> Zdieľať pri relačnej klasifikácii menej informácií (z pohľadu prepojených vrcholov) je prospešné pre kvalitu zatriedenia.

túto považujeme za súhlasiacu s dosiahnutými výsledkami. Pri aplikovaní navrhnutého postupu na iné klasifikačné úlohy je potrebné zvážiť podobnosť štruktúry dátovej vzorky a uvedomiť si, že najsilnejšia úroveň moderovania neprináša najlepšie výsledky. Pokiaľ by aplikovanie nami navrhnutej metódy neprinieslo očakávané výsledky, je vhodné zvážiť, či je vôbec potrebné aplikovať na úlohu klasifikátor s kolektívnym usudzovaním ($\hat{c}_{ci}$) a či namiesto neho nestačí použiť priamy relačný klasifikátor ($\hat{c}_r$). Tento prípad použitia ($\hat{c}_r$) môže byť výhodnejší tam, kde sme si istí vysokou mierou homofílie v grafe, vďaka čomu bude prirodzene zabezpečené, že sa v grafe nebude šíriť priveľké množstvo mätúcich informácii.

Uvedený príspevok k stavu poznania považujeme v kontexte homofílie v relačnej klasifikácii za zaujímavý i z dôvodu, že dochádza k *neinformovanému* uprednostneniu homofílnych tendencií v klasifikovanom grafe. Aj keď zvolená miera $homophily_s$ v nami použitom scenári zachytáva *iba* kvalitu vlastnej distribúcie tried vrcholu, napriek tomu jej vplyv počas iteratívneho zdieľania pomáha rýchlo vytvárať podgrafy, v ktorých prevažuje jedna trieda, čiže homofílne orientované zoskupenia vrcholov. Dôvodom vzniku takéhoto usporiadania je iteratívnosť kolektívneho usudzovania.



Obrázok 5.6: Vzťah medzi vrcholmi pri moderovaní výmeny informácií.

Ak uvážime situáciu na obr. 5.6, vidíme, že v určitom iteračnom kroku t počítame hodnotu $homophily_s(v_1)$ iba na základe vektora príslušnosti vrcholu v_1, tento je však ovplyvnený vektormi príslušnosti k triede od susedných vrcholov z iterácie $t-1$.

Naším cieľom je vytvoriť aj takú metódu, ktorá bude zvyšovať mieru homofílie priamo na základe premostenia predpokladu homofílie. Takúto *informovanú* metódu založenú na grafovom ohodnocovacom algoritme šírenie aktivácie (angl. activation spreading [Ceglowski *et al.*, 2003]) uvádzame v ďalšej kapitole.

Kapitola 6

Návrh metódy ohodnotenia okolia vrcholu

V tejto kapitole sa venujeme druhej hypotéze z časti 1.2:
Po rozšírení bežnej funkcie susedstva vrcholu tak, aby bola zohľadnená lokálna štruktúra grafu, sa zvýši kvalita klasifikácie, pretože sa zvýši homofília.

Naším cieľom je aplikovať nad jedným klasifikátorom dva rôzne prístupy k získavania okolia vrcholu (viac v prehľade v kap. 4.1.1) a porovnať, ktoré *okolie* lepšie vplýva na hodnotu homofílie v grafe. Rozhodli sme sa zvoliť priamu klasifikačnú metódu SRC (kap. 3.2.3), pretože ide o najjednoduchšiu priamu relačnú metódu.

S použitím metódy SRC je naším cieľom:

- určiť, ako sa mení miera homofílie v závislosti od zvoleného prístupu na získanie okolia vrcholu, konkrétne porovnať dve metódy: metóda priameho okolia voči navrhnutej metóde s lokálnym ohodnocovaním grafu šírením aktivácie,
- zistiť priamu závislosť medzi mierou homofílie a kvalitou klasifikácie metódy SRC,
- hypotézu experimentálne overiť na dátovej vzorke foaf.sk.

Náš predpoklad je, že miera homofílie okolia vrcholu bude rozdielna pre jednotlivé prístupy k získavaniu okolia, a konkrétne, že navrhnutá metóda s lokálnym ohodnocovaním predčí jednoduché priame okolie, pretože lepšie

vyhladzuje okolie vrcholu a prináša rozsiahlejšie susedstvo aj pri vrcholoch, ktoré majú iba jedného priameho suseda (viac v kap. 6.3).

Algoritmus šírenia aktivácie v grafe je z rodiny metód na lokálne ohodnocovanie okolia vrcholu. Medzi podobné metódy patrí napr. prístup založený na náhodných prechodoch [Tong *et al.*, 2006] (Random Walks with Restart). Z pohľadu globálneho ohodnocovania je príbuznou metódou napr. algoritmus PageRank [Page *et al.*, 1999]. V práci [Suchal, 2008] je uvedený prehľad lokálnych aj globálnych ohodnocovacích algoritmov.

Postup pri ohodnocovaní okolia vrcholu v_k pomocou šírenia aktivácie je takýto:[1]:

```
aktivuj(energia E, vrchol v_k) {
   energia(v_k) = energia(v_k) + E
   E' = E / stupeň vrcholu(v_k)
   ak (E' > T) {
      pre každý vrchol v_j ∈ V_k {
         aktivuj(E', v_j)
      }
   }
}
```

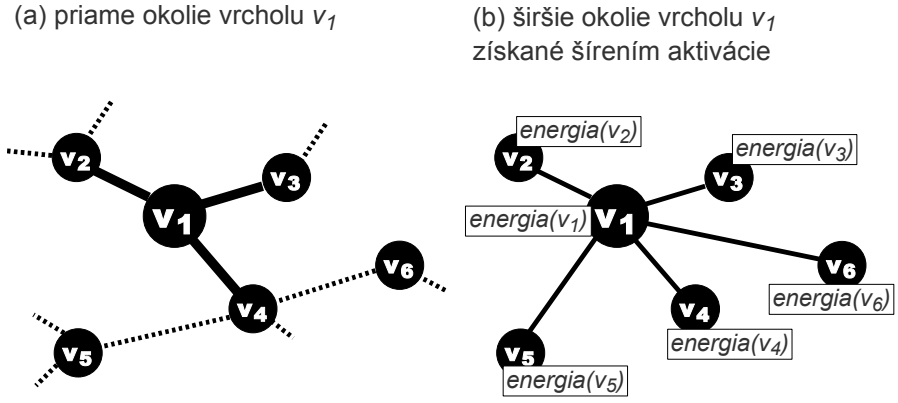
Ide o rekurzívny algoritmus, kde zvolíme počiatočný vrchol, ktorého okolie chceme ohodnotiť, ďalej určíme množstvo energie E, ktoré sa má šíriť a určíme prahovú hodnotu energie T, ktorá zabezpečí rýchlu konvergenciu do ustáleného stavu. $v_j \in V_k$ je množina susedných vrcholov pre aktuálny vrchol v_k (v príklade na obr. 6.1(a) má vrchol v_1 priamych susedov $V_1 = \{v_2, v_3, v_4\}$).

Šírenie aktivácie priraďuje energiu vrcholom, nie hranám. Pre zachovanie súladu s metódou SRC definujeme $w(v_k, v_j)$ takto:

$$w(v_k, v_j) = \frac{energia(v_k)}{energia(v_j)} \tag{6.1}$$

[1] Pseudokód šírenia aktivácie je prevzatý z [Ceglowski *et al.*, 2003] a zjednodušený pre graf s neohodnotenými hranami. Zjednodušenie algoritmu sme zvolili z dôvodu, že sa s takýmto druhom grafov stretávame pri experimente a praktickom využití.

Výsledkom šírenia aktivácie je ohodnotenie okolitých vrcholov inštancie v_1, pričom toto zohľadňuje tvar grafu a vzdialenosť vrcholov v grafe (čím je vrchol vzdialenejší od v_1, tým menej energie sa k nemu prešíri). Obr. 6.1(a) znázorňuje graf s hranami bez váh. Ak by sme ohodnocovali v_1, potom $V_1 = \{v_2, v_3, v_4\}$. Avšak ak aplikujeme šírenie aktivácie (obr. 6.1(b)), dostaneme $V_1 = \{v_2, v_3, v_4, v_5, v_6\}$ spolu s váhami *hrán*.



Obrázok 6.1: Dva druhy susedstva.

6.1 Prepojenie medzi homofíliou a metódou SRC

Ak uvážime ústredný vzorec metódy SRC,

$$p(c_m|v_k) = \frac{1}{W} \sum_{v_j \in V_k | trieda(v_j) = c_m} w(v_k, v_j) \tag{6.2}$$

kde $W = \sum_{v_j \in V_k} w(v_k, v_j)$, čo môžeme prepísať takto:

$$p(c_m|v_k) = \frac{\sum_{v_j \in V_k | trieda(v_j) = c_m} w(v_k, v_j)}{\sum_{v_j \in V_k} w(v_k, v_j)} = \frac{W_{k_{c_m}}}{W_k} \tag{6.3}$$

Je zrejmé že $W_k = \sum_{c_m \in C} W_{k_{c_m}}$ a že tento podiel priamo zodpovedá zložkám výpočtu $homophily_s$ (zložkami myslíme jednotlivé triedy):

$$homophily_s(V_k \cup \{v_k\}) = 1.0 + \sum_{v_i \in V_k \cup \{v_k\}, c \in C} p(c|v_i) \log_{base} p(c|v_i) \quad (6.4)$$

Keď uvažujeme o binárnej klasifikácii s triedami $C = \{c_+, c_-\}$, dostávame $W_k = W_{k_{c_+}} + W_{k_{c_-}}$. Na to, aby sme určili dopad dvoch rôznych metód k získavaniu okolia vrcholu nám potom stačí sledovať pomer $W_{k_{c_+}} : W_k$. Ak $\frac{W_{k_{c_+}}}{W_k} > 0.5$, potom vrchol v_k má podľa metódy SRC priradenú pozitívnu triedu, ak $\frac{W_{k_{c_+}}}{W_k} < 0.5$, potom $trieda(v_k) = c_-$, inak je $trieda(v_k)$ neurčená.

6.2 Podmienky experimentu

Súčasné relačné prístupy využívajú pri zdieľaní informácií medzi vrcholmi priame susedstvo. Naším cieľom je preskúmať, ako sa zmení miera homofílie v pôvodne neváhovanom grafe, ak budeme pri výmene informácií voliť širšie susedstvo získané šírením aktivácie. Náš predpoklad je, že globálne sa v grafe miera homofílie zvýši, pretože klasifikátor má k dispozícii v priemere viac susedných vrcholov, pričom sila ich vplyvu je odstupňovaná na základe vypočítanej váhy hrany.

Ako dátovú vzorku pre experimentálne overenie sme zvolili sociálnu sieť ľudí a firiem z portálu foaf.sk [Suchal & Vojtek, 2009]. Dátová vzorka, na vzniku ktorej sme sa autorsky podieľali, je bližšie opísaná v Prílohe B. Sociálna sieť je bipartitný graf, kde sú ľudia spojení neorientovanými neváhovanými hranami so spoločnosťami, pričom hrana indikuje, že osoba má rolu vo firme (zvyčajne riaditeľ, člen dozornej rady). Graf má 352 000 vrcholov typu `osoba`, 168 000 vrcholov typu `spoločnosť` a 460 000 hrán.

Postup experimentu je takýto:

1. Zvolíme atribút vrcholu, ktorý bude predstavovať predmet klasifikácie. Pre dátovú vzorku sme na základe adresy určili binárny atribút `zBratislavy` podľa toho, či je osoba/spoločnosť z hlavného mesta SR alebo nie.
2. Postupne prechádzame cez všetky vrcholy a analyzujeme ich okolité vrcholy (priame susedstvo voči susedstvu obohatenému šírením aktivácie) s cieľom zistiť, nakoľko hodnota atribútu `zBratislavy` zvoleného vrcholu koreluje s hodnotou tohto atribútu v okolitých vrcholoch – teda počítame hodnotu $homophily_s$ (kap. 4.1.2).

Našim cieľom je dať do súvisu distribúciu hodnoty klasifikovaného atribútu pre jednotlivé hodnoty, t.j. zoskupíme vždy tie vrcholy, ktoré majú v susedstve atribút distribuovaný rovnako a pre tieto vrcholy zistíme, aká je v priemere ich vlastná hodnota atribútu.

Pri vyhodnotení tvorí testovaciu množinu vždy len jeden sledovaný vrchol (čiže ide o *leave-one-out* vyhodnotenie). Ak by sme použili iné rozdelenie (napr. $X_{tr} : X_{tst} = 70 : 30$), znevýhodnili by sme pôvodnú metódu (priame susedstvo), ktorá by mala pre vrchol v priemere o toľko percent menej susedov, aké by bolo zastúpenie testovacej množiny.

6.3 Výsledky experimentu a diskusia

Výsledky sú znázornené na obr. 6.2. Os x znázorňuje podiel $\frac{W_{kc_+}}{W_k}$ a os y zodpovedá priemernej hodnote $homophily_s$ pre vrcholy združené podľa rovnakej hodnoty na osi x[2].

Na obr. 6.2 sú porovnané tri krivky, priebeh optimálnej homofílie v závislosti od $\frac{W_{kc_+}}{W_k}$ je porovnaný so skutočným pozorovaným priebehom pre jednoduché priame susedstvo a pre susedstvo generované šírením aktivácie. Krivka, ktorá lepšie kopíruje optimálny priebeh, vytvára homofílnejšie susedstvo. Ako vidieť z obr. 6.2, šírenie aktivácie tu prináša očividne lepšie výsledky. Ak odchýlky porovnáme pomocou RMSE (Root Mean Square Error), dostávame:

[2]Os x je vzorkovaná s krokom $step = 0.1$, teda napr. keď vrchol v_k má troch susedov prislúchajúcich k pozitívnej triede a jeden sused je z negatívnej triedy, potom $\frac{W_{kc_+}}{W_k} = \frac{3}{4}$.

- typ vrcholu `spoločnosť`: $RMSE_{zakl.\_susedstvo} = 0.360$ a $RMSE_{sirenie\_akt.} = 0.219$,
- typ vrcholu `osoba`: $RMSE_{zakl.\_susedstvo} = 0.374$ a $RMSE_{sirenie\_akt.} = 0.222$.

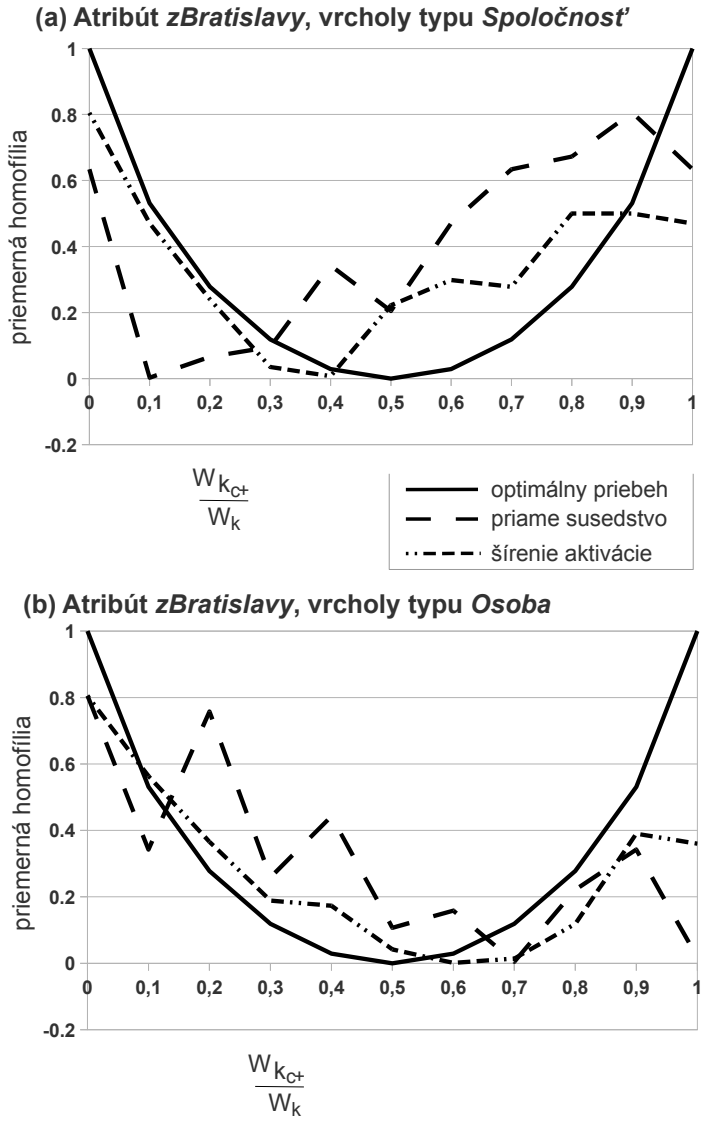
Pre porovnanie uvádzame v tab. 6.1 miery tabuľky podmienenosti. Vidieť, že šírenie aktivácie tu predčuje jednoduché priame susedstvo vo všetkých mierach okrem Recall pri type vrcholu `osoba`. Dôvodom je nerovnováha medzi $c_+ : c_-$ v dátovej vzorke, kde k triede c_- patrí 73% vrcholov, ale Recall sa počíta na podmnožine vrcholov triedy c_+.

Tabuľka 6.1: Výkonnosť klasifikátora podľa mier tabuľky podmienenosti.

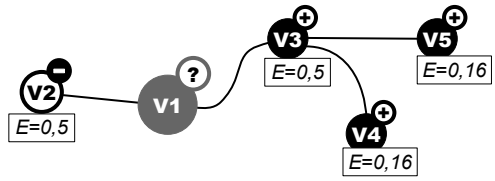
	`spoločnosť`		`osoba`	
	priame sused.	šír. akt.	priame sused.	šír. akt.
recall [%]	85.8	90.8	71.0	56.8
precision [%]	18.2	59.5	24.3	89.1
f1 [%]	74.7	86.1	77.5	79.4
accuracy [%]	30.0	71.9	36.2	69.4
RMSE	0.360	0.219	0.374	0.222

Je viacero dôvodov, pre ktoré šírenie aktivácie prináša kvalitnejšie výsledky a lepšie aproximuje optimálny priebeh krivky homofílie. Ak uvážime príklad na obr. 6.3, pri aplikovaní jednoduchého susedstva má vrchol v_1 takýchto susedov; $V_1 = \{v_2, v_3\}$. Táto konštelácia nie je z hľadiska klasifikácie príliš výhodná, pretože $trieda(v_2) = c_-$ a $trieda(v_3) = c_+$, $\frac{W_{1c_+}}{W_1} = 0.5$. Ak však uvážime susedstvo získané šírením aktivácie (s počiatočnou energiou $E = 1.0$ a prahom šírenia $T = 0.15$), dostaneme susedov $V_1 = \{v_2, v_3, v_4, v_5\}$. Potom $\frac{W_{1c_+}}{W_1} = 0.625$ a tomu zodpovedá vyššia miera homofílie než pri predošlom prípade.

Počiatočná energia šírenia aktivácie bola nastavená na $E = 300.0$ a prah na $T = 1.0$, takže sme zvyčajne dostali od 10 do 100 vrcholov v susedstve a proces ohodnotenia sa rýchlo ustálil. Zvýšenie aktivačnej energie alebo zníženie prahu by prinieslo širšie okolie vrcholu, resp. presnejšie ohodnotenie, ale čas výpočtu by sa predĺžil. Zníženie aktivačnej energie by naopak prinieslo menej susedov, prípadne iba priamych susedov, čím by sme získali rovnakú informáciu ako pri jednoduchom získavaní susedstva.



Obrázok 6.2: Porovnanie priebehu homofílie pre dva prístupy získavania susedných vrcholov grafu, atribút `zBratislavy`.

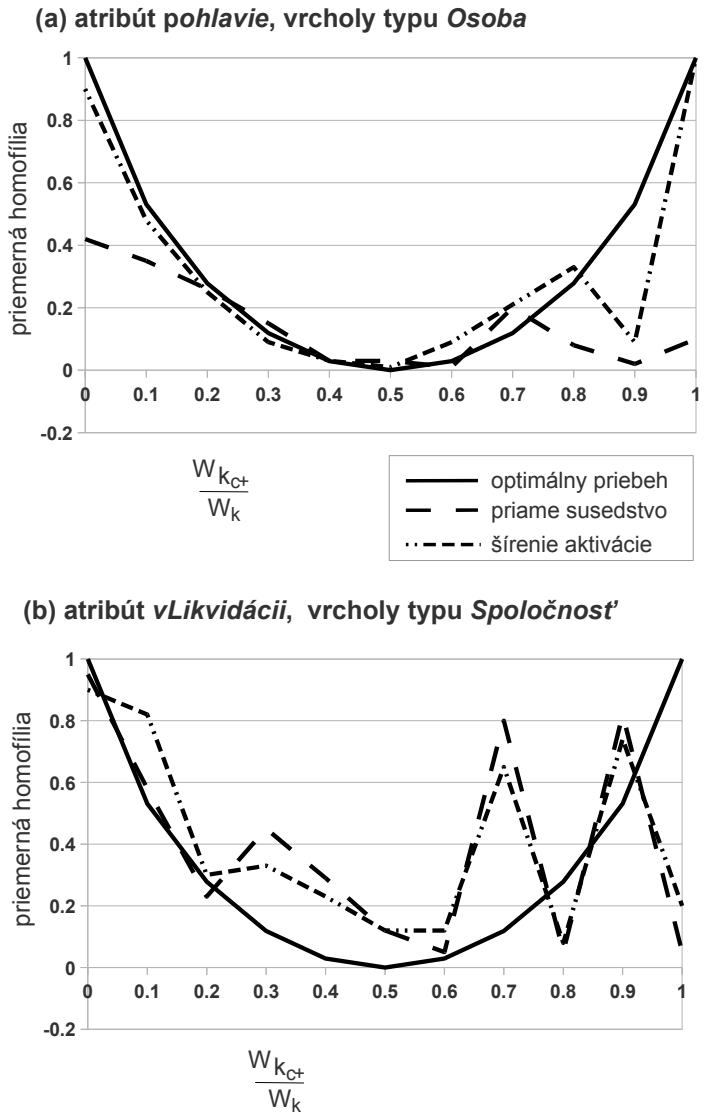


Obrázok 6.3: Príklad grafu s meniacou sa homofíliou v závislosti od druhu susedstva.

Pre získanie lepšieho prehľadu sme ešte vykonali dva dodatočné experimenty, v ktorých sme nad tým istým grafom vyhodnocovali homofíliu dvoch ďalších atribútov. Pre vrcholy typu `osoba` sme zvolili atribút `pohlavie`, ktorý sme určili pomocou heuristiky na základe koncovky priezviska osoby. Pre vrcholy typu `spoločnosť` sme určili binárny atribút `vLikvidácii`, ktorý určuje, či je spoločnosť v likvidácii (túto informáciu vkladá Obchodný register SR priamo do názvu spoločnosti). Keďže každý atribút je špecifický pre jeden typ vrcholu, nemohli sme použiť priame susedstvo, ale zvolili sme susedstvo druhého stupňa (*susedia susedov*). Množinu susedov získavaných cez šírenie aktivácie sme z rovnakého dôvodu obmedzovali iba na daný typ vrcholu (`osoba` pre atribút `pohlavie`, `spoločnosť` pre atribút `vLikvidácii`). Výsledky tohto experimentu sú na obr. 6.4. Hodnoty RMSE sú nasledovné:

- typ vrcholu `osoba`, atribút `pohlavie`:
  $RMSE_{zakl.\_susedstvo} = 0.367$, $RMSE_{sirenie\_akt.} = 0.142$
- typ vrcholu `spoločnosť`, atribút `vLikvidácii`:
  $RMSE_{zakl.\_susedstvo} = 0.393$, $RMSE_{sirenie\_akt.} = 0.331$

Z výsledkov vidieť, že šírenie aktivácie aj pri týchto atribútoch lepšie aproximuje optimálny priebeh homofílie. V porovnaní s atribútom `zBratislavy` sa ostatné dva atribúty líšia v pomere pozitívnych a negatívnych príkladov. Pre atribút `pohlavie` je podiel *ženy : muži* $= 22 : 78$ a pre atribút `vLikvidácii` je podiel $c_+ : c_- = 5 : 95$. Veľká (ale prirodzená) nerovnováha distribúcie atribútu `vLikvidácii` v dátovej vzorke spôsobuje veľké výkyvy v grafe na obr. 6.4(b) pre interval $\frac{W_{kc_+}}{W_k} \in (0.6, 1.0)$.



Obrázok 6.4: Porovnanie priebehu homofílie pre dva prístupy získavania susedných vrcholov grafu, atribúty `pohlavie` a `vLikvidácii`.

V predošlej a tejto kapitole sme navrhli a overili dve metódy, ktoré dokážu zohľadnit' homofíliu v grafe, pričom oba prístupy sa správajú v súlade s hypotézami stanovenými v časti 1.2. V kapitole 5 sme navrhli metódu na moderovanie výmeny informácií v grafe a ukázali sme, že vhodne zvolené obmedzovanie zdieľaných informácií medzi vrcholmi, založené na miere homofílie vrcholu, vedie k zlepšeniu kvality klasifikácie. V aktuálnej kapitole sme porovnali bežný spôsob získavania susedstva v priamom relačnom klasifikátore so susedstvom získaným lokálnym ohodnocovaním. Ukázali sme, že druhý typ susedstva prináša nad tým istým grafom homofílnejšie susedstvo a tým zvyšuje kvalitu klasifikácie. Podrobnejšia analýza nášho prínosu je v kap. 8.

Kapitola 7

Ďalšie smery výskumu

V tejto kapitole uvádzame viaceré oblasti relačnej klasifikácie, ktoré neboli v priamom smere nášho výskumu, avšak považujeme ich za zaujímavé a predpokladáme, že ide o smery v ktorých možno očakávať ďalší výskum.

7.1 Priradenie viacerých tried inštancii

Pri riešení klasifikačných problémov sme v kap. 6 pristupovali k označovaniu tried klasickým disjunktným prístupom podľa Aristotela (kap. 2.1), kedy inštancia po skončení klasifikačného procesu nadobúda nanajvýš jednu triedu (angl. single-label).

V praxi sa však s potrebou viactriedneho priradenia stretávame, napr. v experimente v kap. 5 alebo pri identifikácii jazyka dokumentu, kde klasifikujeme texty podľa ich obsahu. Ak je napr. 60% dokumentu v anglickom jazyku a 40% v slovenskom jazyku, pri tradičnom prístupe bude dokument označený ako napísaný v anglickom jazyku, keďže tento prevažuje. Je zrejmé, že keby bol dokument zároveň označený triedou *slovenský jazyk*, klasifikátor by sme hodnotili ako kvalitnejší. Podobne klasifikácia novinových článkov často naráža na potrebu zaradiť článok do viacerých tried, takisto ako úloha zatrieďovania vedeckých publikácií v kap. 5.

Viactriedna klasifikácia je odlišný pojem ako fazetová klasifikácia – v prvom prípade ide o priraďovanie viacerých tried z jednej klasifikácie C, teda $trieda(v_k) \in C$, kým v prípade fazetovej klasifikácie jednému vrcholu prideľujeme viacero tried, ale každú z inej klasifikácie.

Najjednoduchším prístupom k viactriednej klasifikácii je vykonať n binárnych klasifikácií, $n = |C|$, teda klasifikáciu s výsledkom *áno* alebo *nie* pre každú triedu $c_m \in C$ [Gao *et al.*, 2004]. Takýto prístup je aplikovateľný na akúkoľvek klasifikačnú metódu (aj relačnú), ako sme ukázali v kap. 5. Môže mať však tieto nevýhody:

- pomer pozitívnych a negatívnych príkladov triedy v trénovacej množine môže byť značne nevyrovnaný, čo neprospieva kvalite výsledku,
- jednotlivé binárne modely o sebe navzájom *nevedia* – každý model je tvorený zo značne obmedzeného sveta.

Druhý uvedený nedostatok sa snaží riešiť prístup uvedený v práci [Ghamrawi & McCallum, 2005], kde sa vytvárajú klasifikačné modely pre všetky dvojice (resp. trojice) tried. Alternatívny prístup spočíva v použití vektora príslušnosti k triede a stanovení prahu akceptovania triedy [Schapire & Singer, 2000 Voľba víťaznej triedy pre inštanciu v_k sa tak nestanoví pomocou $trieda(v_k) = argmax_{c_i}[p(c_i|v_k)]$, ale $trieda(v_k) = C_k$, pričom pre prah akceptovania t platí:

$$\forall c_l \in C : c_l \in C_k \; ak \; p(c_l|v_k) > t \tag{7.1}$$

Uvedené riešenia sú publikované pre atribútovo-viazané prístupy. Pri relačných metódach nám nie je známa žiadna práca, ktorá by sa tejto oblasti venovala – vidíme tu priestor pre ďalší výskum, keďže trendom vo viactriednej klasifikácii je posun k metódam, schopným využiť čo najviac informácií z dátovej vzorky, pričom relačné dáta v sebe obsahujú ešte vrstvu informácií navyše, ktorou sú explicitné vzťahy medzi inštanciami.

7.2 Bias v metódach vyhodnocovania úspešnosti klasifikácie

V kap. 2.5 sú uvedené prístupy k vytváraniu trénovacej a testovacej množiny s cieľom čo najlepšie sa priblížiť distribúcii inštancií, ktorá zodpovedá skutočnej chybe klasifikácie. Pre zvýšenie štatistickej relevancie experimentu sa pri atribútových metódach používa krížová validácia a *bootstrapping*. Tieto prístupy predpokladajú, že inštancie sú nezávisle a identicky distribuované. Daná podmienka však neplatí pre relačné dátové vzorky, ktoré naopak explicitne obsahujú závislosti medzi inštanciami. Výberu trénovacej a testovacej množiny preto treba venovať pozornosť. Existujú viaceré prístupy na rozdeľovanie inštancií do týchto množín tak, aby došlo k minimalizácii biasu vneseného reláciami [Korner & Wrobel, 2005, Jensen & Neville, 2002]:

- Rozdeliť graf inštancií na X_{tr} a X_{tst} tak, aby medzi nimi neboli žiadne hrany. To je niekedy možné z povahy grafu, inokedy je nutné hrany odstrániť, čím sa však vnášajú nepresnosti.
- Ak sú hrany orientované, rozdelit' X_{tr} a X_{tst} tak, aby hrany viedli len z X_{tr} do X_{tst}, nie naopak.
- Ak sú inštancie časovo označkované, rozdeliť ich podľa časovej informácie tak, že tie ktoré vznikli do času t patria do X_{tr} a tie čo vznikli po čase t patria do X_{tst}.

Uvedené jednoduché prístupy v sebe nesú riziko nerovnomerného zastúpenia tried v X_{tr} a X_{tst}. Z tohto dôvodu je v práci [Korner & Wrobel, 2005] uvedený zovšeobecnený postup vzorkovania trénovacej množiny, výskum v tejto oblasti však možno považovať za aktuálne prebiehajúci.

7.3 Celulárne automaty

Celulárne automaty predstavujú diskrétny model, ktorý umožňuje simulovať rôzne prírodné javy reprezentovateľné na pravidelnej mriežke a vykonávať týmto spôsobom výpočty v komplexných systémoch. V tejto časti sa venujeme viacerým črtám, v ktorých sa celulárne automaty a relačné klasifikačné metódy na seba podobajú.

7.3.1 Mriežka ako graf

Celulárny automat je definovaný na mriežke, zvyčajne štvorcového tvaru. Bunka v mriežke predstavuje základnú jednotku abstrakcie [Marshall, 2008].

Definícia 7.1: Funkcia susednosti. Nech C je množina buniek na mriežke N. Potom funkcia susednosti, určená ako $n : C \times C \rightarrow \{True, False\}$, vracia hodnotu *True* práve vtedy, keď vstupný pár buniek susedí. ▶

Existuje viacero typov susednosti, na štvorcovej mriežke sa najčastejšie stretávame s von Neumannovym okolím[1]:

$$N = \{\{0, -1\}, \{-1, 0\}, \{0, 0\}, \{+1, 0\}, \{0, +1\}\} \tag{7.2}$$

alebo Moorovym okolím bunky, ktoré je definované takto:

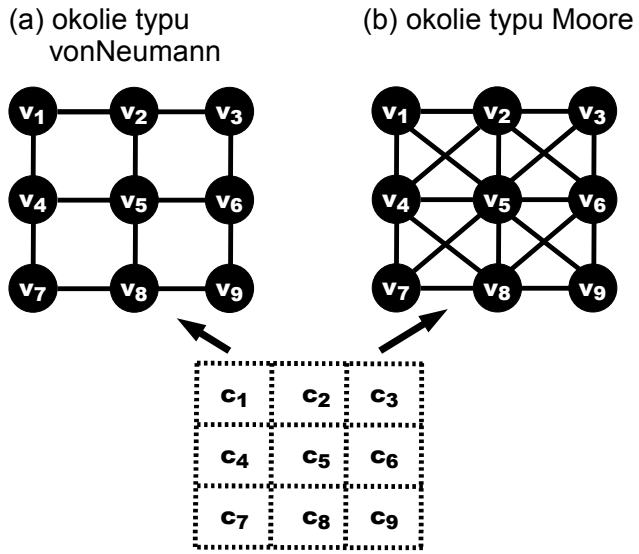
$$\begin{aligned} N = \{&\{-1, -1\}, \{0, -1\}, \{1, -1\}, \{-1, 0\}, \{0, 0\}, \\ &\{+1, 0\}, \{-1, +1\}, \{0, +1\}, \{+1, +1\}\} \end{aligned} \tag{7.3}$$

Okolie bunky c_0 potom získame takto:

$$N_{c0} = c \in C | n(c_0, c) = True \tag{7.4}$$

Mriežku vieme transformovať na graf, kde množina vrcholov grafu V zodpovedá množine buniek C, množina hrán medzi vrcholmi E je reprezentovaná v celulárnych automatoch funkciou susednosti n. Na obr. 7.1 je znázornená transformácia dvojrozmernej mriežky s okrajmi (nie je toroidálna) na graf v závislosti podľa zvoleného typu okolia. V grafe potom môžeme uvažovať o rovnakej funkcii susednosti ako v (7.4).

[1] Prevzaté z `http://mathworld.wolfram.com/vonNeumannNeighborhood.html` [cit. 2009-08-15].



Obrázok 7.1: Transformácia mriežky na graf v závislosti od typu okolia.

7.3.2 Prechodová funkcia ako iteratívny klasifikátor

Každá bunka celulárneho automatu interaguje s okolím nadobúda v diskrétnom čase jeden zo stavov S. Prechodová funkcia, ktorá sa aplikuje na každú bunku v C je v tvare

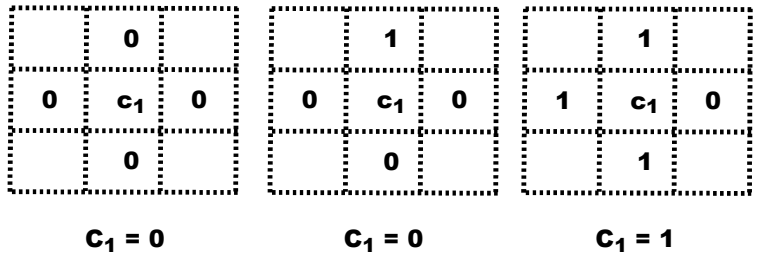
$$u : S^{|N|} \rightarrow S \tag{7.5}$$

Prechodová funkcia teda závisí od stavu buniek vo svojom susedstve v danom čase. V relačných klasifikačných metódach vidieť značnú podobnosť. Prechodovej funkcii u zodpovedá priamy relačný klasifikátor zhmotňujúci metódu $\hat{c}_r$. Vrchol grafu, zodpovedajúci bunke, má príslušnosť k triede (zodpovedá stavu bunky). V najjednoduchšom prípade je príslušnosť k triede binárna, $C = \{0, 1\}$, čo zodpovedá dvojstavovému celulárnemu automatu.

Ak napríklad uvážime prechodovú funkciu pre Simple Relational Classifier:

$$p(c|v_k) = \frac{\sum_{v_j \in V_k | trieda(v_j)=c} w(v_k, v_j)}{\sum_{v_j \in V_k} w(v_k, v_j)} \tag{7.6}$$

túto vieme pre štvorcovú mriežku prepísať prístupom podľa obr. 7.2.



Obrázok 7.2: Časť prechodovej funkcie pre Simple Relational Classifier pri okolí typu vonNeumann.

Uvedená prechodová funkcia naráža na problém, ako sa má bunka správať v prípade že je v susedstve zhodný počet buniek v oboch stavoch (t.j. práve polovica buniek má kladnú a polovica zápornú príslušnosť k triede). V praxi pri použití relačného klasifikátora táto *nerozhodnosť* zvyčajne nespôsobuje problém, keďže sa používa spojitý prechod medzi stavmi, takže klasifikovaný vrchol môže nadobúdať hodnoty na intervale $\langle 0, 1 \rangle$ a v takomto nerozhodnom prípade je hodnota stavu 0.5. Násobná aplikácia prechodovej funkcie má iteratívny charakter, čo zodpovedá kolektívnemu usudzovaniu $\hat{c}_{ci}$.

7.3.3 Využitie a obmedzenia

Na blízkosť celulárnych automatov a relačných klasifikačných prístupov v tejto práci upozorňujeme z dôvodu, že sme takéto premostenie napriek našej snahe v literatúre nenašli a v prípade vypracovania príslušnej teórie

by sme získali možnosť aplikovať poznatky z oblasti celulárnych automatov na klasifikačné metódy (a naopak). Ide najmä o otázku konvergencie: pre novonavrhnutú klasifikačnú metódu s kolektívnym usudzovaním je vhodné ukázať jeho schopnosť ustáliť sa pri klasifikácii, zvyčajne za pomoci Markovovskej vlastnosti, čo nie je vždy jednoduchá úloha. Ak by sa namiesto toho podarilo vytvoriť transformáciu klasifikačnú metódu na celulárny automat so známymi vlastnosťami konvergentnosti[2], vedeli by sme, či je schopný dosiahnuť ustálený stav.

Jednou z aplikácií môžu byť klasifikačné problémy v grafoch pravidelného tvaru, napr. rozpoznávanie vzorov na obraze, ktorý je tvorený mriežkou pixelov.

[2]Napr. štyri základné triedy celulárnych automatov podľa [Wolfram, 1984].

Kapitola 8

Zhodnotenie a prínosy práce

V súčasnosti je relačná klasifikácia ešte stále novou vetvou klasifikácie. V oblasti prebieha výskum, avšak ešte nedošlo k masovému nasadeniu do komerčných aplikácií. Napríklad známy softvér na dolovanie v dátach PASW Statistics 18[1] (bývalý SPSS) obsahuje viaceré atribútové klasifikačné metódy (rozhodovacie stromy, neurónové siete, logistickú regresiu), ale žiadne relačné metódy. Jedným z dôvodov je počiatočná nedôvera i skutočnosť, že väčšina z nás nie je zvyknutá nenazerať na svet a problémy, ktoré v ňom riešime ako na úlohy reprezentovateľné grafom.

V práci sme rozvinuli relačný grafový prístup na organizáciu a analýzu dát. Identifikovali sme vlastnosti grafu, ktoré majú vplyv na výsledok klasifikácie a priniesli sme poznatok, že nie je vhodné, aby sme tieto vlastnosti grafov brali ako samozrejmé, ako to doteraz bolo zvykom.

Konkrétnejšie, cieľom práce bolo vytvoriť novú klasifikačnú metódu, ktorá bude zohľadňovať a využívať vplyv homofílie dátovej vzorky. V práci sme vychádzali z existujúcich prehľadov relačných klasifikačných metód [Jensen *et al.*, 2004, Macskassy & Provost, 2007], ktoré zaviedli *klasifikáciu* klasifikačných metód. Na základe analýzy a po vykonaní experimentov sme identifikovali pre túto vetvu metód väzbu na homofíliu v dátach, nevyhnutnú pre ich správnu funkčnosť. K postrehu previazanosti s homofíliou nás inšpirovala práca [Gallagher *et al.*, 2008]. Pri skúmaní všadeprítomnosti javu homofílie bola pre nás významná práca [Mcpherson *et al.*, 2001].

[1] `http://www.spss.com/software/statistics/` [cit. 2009-10-13].

Na základe získaného prehľadu sme vytvorili dve nové metódy na zvýšenie robustnosti klasifikácie vzhľadom na meniacu sa homofíliu v dátach, konkrétne:

- metódu založenú na moderovaní výmeny informácií medzi inštanciami (kap. 5),
- metódu s rozšíreným získavaním susedstva vrcholu pomocou lokálneho ohodnocovacieho algoritmu (kap. 6),

V rámci pôvodného prínosu sme v spolupráci vytvorili dve grafové dátové vzorky:

- MAPEKUS (príloha A),
- foaf.sk (príloha B),

na ktorých sme overili funkčnosť a prínos oboch prístupov.

Obe uvedené metódy, moderovanie výmeny informácií i rozšírené okolie vrcholu, obohacujú relačnú klasifikáciu o vnímanie homofílnych tendencií v grafe a prinášajú kvalitnejšie výsledky v zmysle zníženia chyby klasifikačného modelu. V práci sme naše metódy aplikovali na tieto relačné klasifikačné metódy:

- moderovanie príslušnosti k triede sme zapojili do metódy IRC, patriacej medzi metódy využívajúce kolektívne usudzovanie, ktorým zodpovedá model $\hat{c}_{ci}$,
- ohodnotenie vrcholu šírením aktivácie sme prepojili s metódou SRC, ktorá spadá medzi priame relačné metódy s modelom $\hat{c}_r$.

Na našom prínose je podstatné, že uvedené zapojenie je voľne zameniteľné, teda moderovanie príslušnosti k triede je integrovateľné do ktorejkoľvek metódy uvedenej v kap. 3.2.4 a ohodnotenie vrcholu šírením aktivácie je zapojiteľné do ktorejkoľvek metódy v kap. 3.2.3. Dokonca, obe naše metódy sú kombinovateľné, teda môžeme použiť ľubovoľnú kombináciu modelov $\hat{c}_{ci} \leftarrow \hat{c}_r \leftarrow \hat{c}_a$, kde $\hat{c}_{ci}$ využíva moderovanie príslušnosti a zároveň model $\hat{c}_r$ používa šírenie aktivácie.

V širšom ponímaní je výsledkom tejto práce preskúmanie závislosti predpokladu homofílie v konštrukcii grafových relačných klasifikačných metód a upozornenie na negatívne dôsledky, ktoré tento predpoklad prináša.

V práci sme naznačili niekoľko smerov, v ktorých predpokladáme ďalší výskum v tejto oblasti – ide najmä o podobnosť grafových relačných klasi-

fikačných metód s celulárnymi automatmi a určenie miery závislosti vyhodnocovacích klasifikačných metrík so spôsobom, akým sa tvorí trénovacia a testovacia množina.

Viaceré výsledky prezentované v práci vznikli v rámci riešenia výskumných projektov:

- projekt MAPEKUS, podporovaný Agentúrou na podporu výskumu a vývoja, APVT-20-007104, so zameraním na modelovanie a získavanie, spracovanie a využitie znalostí o správaní používateľov webových portálov (`http://mapekus.fiit.stuba.sk`),
- projekt NAZOU v rámci Štátneho programu výskumu a vývoja s cieľom ustanovenia informačnej spoločnosti, číslo 1025/04, ktorý sa zameriaval na získavanie, organizáciu a údržbu znalostí v prostredí heterogénnych informačných zdrojov (`http://nazou.fiit.stuba.sk`),
- projekt s názvom "Adaptívny sociálny web a jeho služby pre sprístupňovanie informácií" podporovaný vedeckou grantovou agentúrou VEGA, grant VG1/0508/09,
- projekt s názvom "Modely softvérových systémov v prostredí webu so sémantikou" podporovaný vedecká grantovou agentúrou VEGA, grant VG1/3102/06,
- portál `http://foaf.sk` (sociálna sieť Obchodného registra SR) a dátová vzorka s ním spojená,
- spolupráca s denníkom SME pri vytváraní adaptívnej verzie portálu `http://www.sme.sk`,
- spolupráca s Alianciou Fair Play pri vytváraní portálu pre poskytovanie strojovo konzumovateľných informácií o podnikateľskom prostredí na Slovensku.

Dosiahnuté výsledky výskumu sme publikovali na viacerých vedeckých fórach, z ktorých najvýznamnejšie sú:

- klasifikačná metóda založená na moderovaní výmeny informácií medzi inštanciami: [Vojtek & Bieliková, 2009] (konferencia AWIC, zborník vydaný vydavateľstvom Springer).
- dátová vzorka MAPEKUS a proces jej vytvorenia: [Frivolt *et al.*, 2008] (konferencia SOFSEM, zborník vydaný vydavateľstvom Springer v sérii LNCS).

- klasifikačná metóda s rozšíreným získavaním susedstva vrcholu pomocou lokálneho ohodnocovacieho algoritmu: [Vojtek & Bieliková, 2010] (konferencia SOFSEM, zborník vydaný vydavateľstvom Springer v sérii LNCS).
- dátová vzorka foaf.sk: [Suchal & Vojtek, 2009].

Úplný prehľad publikačnej činnosti je v prílohe D.

Literatúra

[Anagnostopoulos *et al.*, 2008] Anagnostopoulos, Aris, Kumar, Ravi, & Mahdian, Mohammad. 2008. Influence and correlation in social networks. *Pages 7–15 of:* Li, Y., Liu, B., & Sarawagi, S. (eds), *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM.

[Borko & Bernick, 1963] Borko, Harold, & Bernick, Myrna. 1963. Automatic Document Classification. *J. ACM*, **10**(2), 151–162.

[Cai & Hofmann, 2003] Cai, Lijuan, & Hofmann, Thomas. 2003. Text Categorization by Boosting Automatically Extracted Concepts. *Pages 182–189 of: SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, USA: ACM Press.

[Cavnar & Trenkle, 1994] Cavnar, William B., & Trenkle, John M. 1994. N-Gram-Based Text Categorization. *Pages 161–175 of: Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*.

[Ceglowski *et al.*, 2003] Ceglowski, M., Coburn, A., & Cuadrado, J. 2003. *Semantic Search Of Unstructured Data Using Contextual Network Graphs*. Tech. rept. National Institute for Technology and Liberal Education, Middlebury College, Middlebury, Vermont, 05753 USA.

[Chakrabarti *et al.*, 1998] Chakrabarti, S., Dom, B. E., & Indyk, P. 1998. Enhanced Hypertext Categorization Using Hyperlinks. *Pages 307–318 of:* Haas, L., & Tiwary, A. (eds), *Proceedings of SIGMOD-98, ACM International Conference on Management of Data*. Seattle, US: ACM Press, New York, USA.

[Chin *et al.*, 2006] Chin, O. S., Kulathuramaiyer, N., & Yeo, A. W. 2006. Automatic Discovery of Concepts from Text. *Pages 1046–1049 of: Web Intelligence*. IEEE Computer Society.

[Chung *et al.*, 2002] Chung, Fan, Chung, Fan, Chung, Fan, Lu, Linyuan, & Lu, Linyuan. 2002. The Average Distances In Random Graphs With Given Expected Degrees. *Internet Mathematics*, **1**, 15879–15882.

[Clarke & Cooke, 1998] Clarke, G. M., & Cooke, D. 1998. *A Basic Course in Statistics*. 4th edn. Arnold Publishers.

[Cohen & Lefebvre, 2005] Cohen, Henri, & Lefebvre, Claire. 2005. *Handbook of Categorization in Cognitive Science*. Elsevier.

[Davis & Goadrich, 2006] Davis, Jesse, & Goadrich, Mark. 2006. The Relationship Between Precision-Recall and Roc Curves. *Pages 233–240 of:* Cohen, William, & Moore, Andrew (eds), *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA: ACM.

[Denton, 2003] Denton, William. 2003. *How to Make a Faceted Classification and Put It On the Web*. Tech. rept. Miskatonic University, California, USA.

[Diestel, 2005] Diestel, Reinhard. 2005. *Graph Theory (Graduate Texts in Mathematics)*. Springer.

[Doyle, 1965] Doyle, Lauren B. 1965. Is Automatic Classification a Reasonable Application of Statistical Analysis of Text? *J. ACM*, **12**(4), 473–489.

[Drucker *et al.*, 1999] Drucker, Harris, Wu, Donghui, & Vapnik, Vladimir. 1999. Support Vector Machines for Spam Categorization. *IEEE Transactions on Neural Networks*, **10**(5), 1048–1054.

[Dzeroski & Todorovski, 1995] Dzeroski, Saso, & Todorovski, Ljupco. 1995. Discovering Dynamics: From Inductive Logic Programming to Machine Discovery. *Journal of Intelligent Information Systems*, **4**(1), 89–108.

[Efron & Tibshirani, 1995] Efron, B., & Tibshirani, R. 1995. *Cross-Validation and the Bootstrap: Estimating the Error Rate of a Prediction Rule*. Manuscript.

[Fayyad *et al.*, 1996] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. 1996. The Kdd Process for Extracting Useful Knowledge From Volumes of Data. *Commun. ACM*, **39**(11), 27–34.

[Frank *et al.*, 2007] Frank, R., Moser, F., & Ester, M. 2007. A Method for Multi-relational Classification Using Single and Multi-feature Aggregation Functions. *Pages 430–437 of:* Kok, J. N., Koronacki, J., de Mántaras, R. López, Matwin, S., Mladenic, D., & Skowron, A. (eds), *Knowledge Discovery in Databases: PKDD*

2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007, Proceedings. Lecture Notes in Computer Science, vol. 4702. Springer.

[Friedman *et al.*, 1997] Friedman, N., Geiger, D., & Goldszmidt, M. 1997. Bayesian network classifiers. *Machine Learning*, **29**, 131–163.

[Friedman *et al.*, 1999] Friedman, N., Getoor, L., Koller, D., & Pfeffer, A. 1999. Learning Probabilistic Relational Models. *Pages 1300–1309 of: IJCAI '99: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[Friedman, 1999] Friedman, Thomas L. 1999. *The Lexus and the Olive Tree*. Farrar, Straus and Giroux.

[Frivolt *et al.*, 2008] Frivolt, György, Suchal, Ján, Vesely, Richard, Vojtek, Peter, Vozár, Oto, & Bieliková, Mária. 2008. Creation, Population and Preprocessing of Experimental Data Sets for Evaluation of Applications for the Semantic Web. *Pages 684–695 of:* Geffert, V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., & Bieliková, M. (eds), *SOFSEM 2008: Theory and Practice of Computer Science, 34th Conference on Current Trends in Theory and Practice of Computer Science, Nový Smokovec, Slovakia, January 19-25, 2008, Proceedings*. Lecture Notes in Computer Science, vol. 4910. Springer.

[Gallagher *et al.*, 2008] Gallagher, Brian, Tong, Hanghang, Eliassi-Rad, Tina, & Faloutsos, Christos. 2008. Using Ghost Edges for Classification in Sparsely Labeled Networks. *Pages 256–264 of:* Li, Y., Liu, B., & Sarawagi, S. (eds), *KDD '08: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA: ACM Press.

[Galstyan & Cohen, 2006] Galstyan, Aram, & Cohen, Paul R. 2006. Iterative Relational Classification Through Three-State Epidemic Dynamics. *Pages 83–92 of:* Mehrotra, S., Zeng, D. D., Chen, H., Thuraisingham, B. M., & Wang, Fei-Yue (eds), *Intelligence and Security Informatics, IEEE International Conference on Intelligence and Security Informatics, ISI 2006, San Diego, CA, USA, May 23-24, 2006, Proceedings*. Lecture Notes in Computer Science, vol. 3975. Springer.

[Ganti *et al.*, 2008] Ganti, Venkatesh, König, Arnd C., & Vernica, Rares. 2008. Entity Categorization over Large Document Collections. *Pages 274–282 of:* Li, Ying, Liu, Bing, & Sarawagi, Sunita (eds), *KDD '08: Proceeding of the 14th Acm*

SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press.

[Gao *et al.*, 2004] Gao, Sheng, Wu, Wen, Lee, Chin-Hui, & Chua, Tat-Seng. 2004. A MFoM Learning Approach to Robust Multiclass Multi-Label Text Categorization. *Page 42 of:* Gao, S., Wu, W., Lee, C., & Chua, T. (eds), *ICML '04: Proceedings of the 21st International Conference on Machine Learning*. New York, USA: ACM Press.

[Geman *et al.*, 1993] Geman, Stuart, Geman, Donald, Abend, K., Harley, T. J., & Kanal, L. N. 1993. Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *Journal of Applied Statistics*, **20**(5), 25–62.

[Ghamrawi & McCallum, 2005] Ghamrawi, Nadia, & McCallum, Andrew. 2005. Collective Multi-label Classification. *Pages 195–200 of: CIKM '05: Proceedings of the 14th ACM International Conference on Information And Knowledge Management*. New York, USA: ACM Press.

[Giudici, 2003] Giudici, P. 2003. *Applied Data Mining: Statistical Methods for Business and Industry*. Wiley.

[Gnoli *et al.*, 2006] Gnoli, Claudio, Mei, & Hong. 2006. Freely Faceted Classification for Web-Based Information Retrieval. *New Review in Hypermedia and Multimedia*, **12**(1), 63–81.

[Gürel & Kersting, 2005] Gürel, Tayfun, & Kersting, Kristian. 2005 (October). On the Trade-Off Between Iterative Classification and Collective Classification: First Experimental Results. *Pages 25–36 of:* Nijssen, S., Meinl, T., & Karypis, G. (eds), *International Workshop on Mining Graphs, Trees and Sequences (MGTS 2005), in conjunction with ECML/PKDD 2005, Porto, Portugal*.

[Gutierrez-Osuna, 2001] Gutierrez-Osuna, Ricardo. 2001. *Validation (Intelligent Sensor Systems)*. Tech. rept. Wright State University.

[Han & Kamber, 2006] Han, Jiawei, & Kamber, Micheline. 2006. *Data Mining: Concepts and Techniques*. 2 edn. Morgan Kaufmann.

[Isaksson *et al.*, 2008] Isaksson, A., Wallman, M., Göransson, H., & Gustafsson, M. G. 2008. Cross-Validation and Bootstrapping are Unreliable in Small Sample Classification. *Pattern Recognition Letters*, **29**(14), 1960–1965.

[Jackson, 2008] Jackson, Matthew O. 2008. Average Distance, Diameter, and Clustering in Social Networks with Homophily. *Pages 4–11 of:* Papadimitriou, C., & Zhang, S. (eds), *WINE '08: Proceedings of the 4th International Workshop on Internet and Network Economics*. Berlin, Heidelberg: Springer-Verlag.

[Jacob, 2004] Jacob, E. K. 2004. Classification and Categorization: A Difference that Makes a Difference. *Pages 515–540 of: Library Trends 52 (3) Winter 2004: The Philosophy of Information*. Graduate School of Library and Information Science. University of Illinois at Urbana-Champaign. USA.

[Jensen *et al.*, 2004] Jensen, D., Neville, J., & Gallagher, B. 2004. Why Collective Inference Improves Relational Classification. *Pages 593–598 of:* Kim, W., Kohavi, R., Gehrke, J., & DuMouchel, W. (eds), *KDD '04: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA: ACM Press.

[Jensen & Neville, 2002] Jensen, David, & Neville, Jennifer. 2002. Linkage and Autocorrelation Cause Feature Selection Bias in Relational Learning. *Pages 259–266 of: ICML '02: Proceedings of the 19th International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[Keller *et al.*, 2000] Keller, Andrew D., Schummer, Michel, Hood, Lee, & Ruzzo, Walter L. 2000. *Bayesian Classification of DNA Array Expression Data*. Tech. rept. University of Washington, College of Computer Science and Engineering.

[Keller, 2001] Keller, Frank. 2001. *Evaluation (Connectionist and Statistical Language Processing course)*. Tech. rept. Computerlinguistik, Universität des Saarlandes, Germany.

[Ketkar *et al.*, 2005] Ketkar, N. S., Holder, L. B., & Cook, D. J. 2005. Qualitative Comparison of Graph-Based and Logic-Based Multi-Relational Data Mining: a Case Study. *Pages 25–32 of:* Dzeroski, S. (ed), *MRDM '05: Proceedings of the 4th International Workshop on Multi-Relational Mining*. New York, USA: ACM Press.

[Kohavi, 1995] Kohavi, Ron. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Pages 1137–1145 of: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95*. Morgan Kaufmann.

[Korner & Wrobel, 2005] Korner, C., & Wrobel, S. 2005. Bias-free Hypothesis Evaluation in Multirelational Domains. *Pages 33–38 of:* Dzeroski, Saso (ed), *MRDM '05: Proceedings of the 4th international workshop on Multi-relational mining.* New York, NY, USA: ACM Press.

[Kotsiantis *et al.*, 2006] Kotsiantis, S., Zaharakis, I., & Pintelas, P. 2006. Supervised Machine Learning: A Review of Classification Techniques. *Artificial Intelligence Review*, **26**(3), 159–190.

[Kwon & Lee, 2000] Kwon, O., & Lee, J. 2000. Web Page Classification Based on K-Nearest Neighbor Approach. *Pages 9–15 of:* Wong, K., Lee, D., & Lee, J. (eds), *IRAL '00: Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages.* New York, USA: ACM Press.

[Lewis, 1991] Lewis, D. D. 1991. Evaluating Text Categorization. *Pages 312–318 of: Proceedings of Speech and Natural Language Workshop.* Morgan Kaufmann.

[Li *et al.*, 2007] Li, X., Chen, H., Zhang, Z., & Li, J. 2007. Automatic Patent Classification using Citation Network Information: an Experimental Study in Nanotechnology. *Pages 419–427 of:* Rasmussen, E., Larson, R., Toms, E., & Sugimoto, S. (eds), *JCDL '07: Proceedings of the 2007 Conference on Digital libraries.* New York, USA: ACM Press.

[Ling *et al.*, 2004] Ling, C. X., Yang, Q., Wang, J., & Zhang, S. 2004. Decision Trees with Minimal Costs. *Page 69 of:* Gao, S., Wu, W., Lee, C., & Chua, T. (eds), *ICML '04: Proceedings of the 21st International Conference on Machine Learning.* New York, USA: ACM Press.

[Liu, 2006] Liu, B. 2006. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.* Springer.

[Lu & Getoor, 2003] Lu, Qing, & Getoor, Lise. 2003. Link-based Classification. *Pages 496–503 of:* Fawcett, Tom, & Mishra, Nina (eds), *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA.* AAAI Press.

[Macskassy & Provost, 2003] Macskassy, S., & Provost, F. 2003. A Simple Relational Classifier. *Pages 64–76 of:* Dzeroski, S., Raedt, L., & Wrobel, S. (eds), *Proceedings of the 2nd Workshop on Multi-Relational Data Mining, 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York, USA: ACM Press.

[Macskassy & Provost, 2007] Macskassy, Sofus A., & Provost, Foster. 2007. Classification in Networked Data: A Toolkit and a Univariate Case Study. *Journal of Machine Learning Research*, **8**(May), 935–983.

[Maron, 1961] Maron, M. E. 1961. Automatic Indexing: An Experimental Inquiry. *J. ACM*, **8**(3), 404–417.

[Marshall, 2008] Marshall, James A. R. 2008. *Cellular Automata, Computational Methods for Complex Systems*. Tech. rept. Department of Computer Science, University of Bristol.

[McCallum & Nigam, 1998] McCallum, A., & Nigam, K. 1998. A Comparison of Event Models for Naive Bayes Text Classification. *In:* Sahami, Mehran, Craven, Mark, Joachims, Thorsten, & McCallum, Andrew (eds), *Workshop on Learning for Text Categorization, AAAI-98: 15th National Conference on Artificial Intelligence*.

[Mcpherson *et al.*, 2001] Mcpherson, Miller, Lovin, Lynn S., & Cook, James M. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, **27**(1), 415–444.

[Mitchell, 1997] Mitchell, Tom M. 1997. *Machine Learning*. McGraw-Hill Science/Engineering/Math.

[Neville, 2006] Neville, Jennifer. 2006. *Statistical Models and Analysis Techniques for Learning in Relational Data*. Ph.D. thesis. Adviser-Jensen, David.

[Page *et al.*, 1999] Page, Lawrence, Brin, Sergey, Motwani, Rajeev, & Winograd, Terry. 1999. *The PageRank Citation Ranking: Bringing Order to the Web*. Tech. rept. Stanford InfoLab.

[Paralič, 2003] Paralič, J. 2003. *Knowledge discovery in Databases (in Slovak)*. FEI TU Košice, Slovakia.

[Pearl, 1998] Pearl, Judea. 1998. *The handbook of brain theory and neural networks*. Cambridge, MA, USA: MIT Press. Pages 149–153.

[Porter, 1980] Porter, M. F. 1980. An Algorithm for Suffix Stripping. *Program*, **14**(3), 130–137.

[Preisach & Schmidt-Thieme, 2006] Preisach, C., & Schmidt-Thieme, L. 2006. Relational Ensemble Classification. *Pages 499–509 of:* Clifton, C.W., Zhong, N.,

Liu, J., Wah, B.W., & Wu, X. (eds), *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society.

[Quinlan, 1993] Quinlan, Ross J. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

[Rogers *et al.*, 2003] Rogers, Everett M., E., Medina Una, & Rivera, Mario A. Wiley, Cody J. 2003. Complex Adaptive Systems and the Diffusion of Innovations. *The Innovation Journal: The Public Sector Innovation Journal*, **10**(3).

[Sacha, 1999] Sacha, Jaroslaw P. 1999. *New Synthesis of Bayesian Network Classifiers and Cardic SPECT Image Representation*. Ph.D. thesis.

[Salton & Buckley, 1987] Salton, G., & Buckley, C. 1987. *Term Weighting Approaches in Automatic Text Retrieval*. Tech. rept. Ithaca, NY, USA.

[Schapire & Singer, 2000] Schapire, Robert E., & Singer, Yoram. 2000. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning*, **39**(2/3), 135–168.

[Sebastiani, 2002] Sebastiani, Fabrizio. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, **34**(1), 1–47.

[Shannon *et al.*, 1998] Shannon, Claude E., Weaver, Warren, & Shannon. 1998. *The Mathematical Theory of Communication*. University of Illinois Press.

[Shearer, 2000] Shearer, C. 2000. The CRISP-DM model: the new Blueprint for Data Mining. *Journal of Data Warehousing*, **5**(4), 13–22.

[Slattery & Mitchell, 2000] Slattery, Seán, & Mitchell, Tom M. 2000. Discovering Test Set Regularities in Relational Domains. *Pages 895–902 of:* Langley, Pat (ed), *ICML '00: Proceedings of the 17th International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[Suchal, 2008] Suchal, Ján. 2008. On Finding Power Method in Spreading Activation Search. *Pages 124–130 of:* Geffert, V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., & Bieliková, M. (eds), *SOFSEM 2008: Theory and Practice of Computer Science, 34th Conference on Current Trends in Theory and Practice of Computer Science, Nový Smokovec, Slovakia, January 19-25, 2008, Student Research Forum Proceedings*. Safarik University, Kosice, Slovakia.

[Suchal & Vojtek, 2009] Suchal, Ján, & Vojtek, Peter. 2009. Navigation is Social Network of Slovak Companies Register (in Slovak). *Pages 145–151 of: DATAKON 2009, Proceedings of the Annual Database Conference. Srní, Czech Republic, October 10-13.*

[Teahan, 2000] Teahan, William J. 2000. Text Classification and Segmentation using Minimum Cross Entropy. *In:* Mariani, Joseph-Jean, & Harman, Donna (eds), *Proceeding of RIAO-00, 6th International Conference "Recherche d'Information Assistee par Ordinateur".*

[Thelwall, 2009] Thelwall, Mike. 2009. Homophily in MySpace. *J. Am. Soc. Inf. Sci. Technol.*, **60**(2), 219–231.

[Tong *et al.*, 2006] Tong, Hanghang, Faloutsos, Christos, & Pan, Jia-Yu. 2006. Fast Random Walk with Restart and Its Applications. *Pages 613–622 of:* Clifton, C.W., Zhong, N., Liu, J., Wah, B.W., & Wu, X. (eds), *ICDM '06: Proceedings of the 6th International Conference on Data Mining.* Washington, DC, USA: IEEE Computer Society.

[Vapnik, 1982] Vapnik, Vladimir. 1982. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics).* Secaucus, NJ, USA: Springer-Verlag New York, Inc.

[Vickery, 2008] Vickery, Brian. 2008. Faceted Classification for the Web. *Axiomathes*, **18**(2), 145–160.

[Vojtek, 2008] Vojtek, P. 2008. *Multi-relational Classification for Information Processing.* Tech. rept. Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.

[Vojtek, 2006] Vojtek, Peter. 2006. Natural Language Identification in the World Wide Web. *Pages 153–159 of:* Bieliková, M. (ed), *IIT.SRC 2006: Student Research Conference.* Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.

[Vojtek & Bieliková, 2009] Vojtek, Peter, & Bieliková, Mária. 2009. Moderated Class-membership Interchange in Iterative Multi-relational Graph Classifier. *Pages 229–238 of:* Snášel, V., Szczepaniak, P.S., Abraham, A., & Kacprzyk, J. (eds), *AWIC 2009: Proceedings of the 6th Atlantic Web Intelligence Conference.* Advances in Intelligent and Soft Computing, vol. 67. Springer.

[Vojtek & Bieliková, 2010] Vojtek, Peter, & Bieliková, Mária. 2010. Homophily of Neighborhood in Graph Relational Classifier. *Pages 721–730 of:* Geffert,

V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., & Bieliková, M. (eds), *SOFSEM 2010: Theory and Practice of Computer Science, 36th Conference on Current Trends in Theory and Practice of Computer Science, Spindleruv Mlyn, Czech Republic, January 23-29, 2010, Proceedings*. Lecture Notes in Computer Science, vol. 5901. Springer.

[Witten & Frank, 1999] Witten, Ian H., & Frank, Eibe. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (The Morgan Kaufmann Series in Data Management Systems)*. 1st edn. Morgan Kaufmann.

[Wolfram, 1984] Wolfram, Stephen. 1984. Universality and Complexity in Cellular Automata. *Physica D: Nonlinear Phenomena*, **10**(1-2), 1–35.

[Wolpert & Macready, 1997] Wolpert, D. H., & Macready, W. G. 1997. No Free Lunch Theorems for Optimization. *Evolutionary Computation, IEEE Transactions on*, **1**(1), 67–82.

[Xue *et al.*, 2006] Xue, G., Yu, Y., Shen, D., Yang, Q., Zeng, H., & Chen, Z. 2006. Reinforcing Web-object Categorization Through Interrelationships. *Data Min. Knowl. Discov.*, **12**(2-3), 229–248.

[Yang & Pedersen, 1997] Yang, Y., & Pedersen, J. O. 1997. A Comparative Study on Feature Selection in Text Categorization. *Pages 412–420 of:* Fisher, Douglas H. (ed), *ICML '97: Proceedings of the 14th International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[Yang & Liu, 1999] Yang, Yiming, & Liu, Xin. 1999. A Re-examination of Text Categorization Methods. *Pages 42–49 of:* Hearst, Marti A., Gey, Fredric, & Tong, Richard (eds), *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*. Berkeley, USA: ACM Press, New York, USA.

[Yu-Chia *et al.*, 2006] Yu-Chia, Hsieh, Tsung-Zu, Wu, Ding-Ping, Liu, Pei-Lan, Shao, Luan-Yin, Chang, Chun-Yi, Lu, Chin-Yun, Lee, Fu-Yuan, Huang, & Li-Min, Huang. 2006. Influenza Pandemics: Past Present and Future. *Journal of the Formosan Medical Association*, **105**(1), 1–6.

[Zhang, 2004] Zhang, Harry. 2004. The Optimality of Naive Bayes. *In:* Barr, V., & Markov, Z. (eds), *FLAIRS: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA*. AAAI Press.

[Zhang *et al.*, 2007] Zhang, Richong, Shepherd, Michael, Duffy, Jack, & Watters, Carolyn. 2007. Automatic Web Page Categorization using Principal Component Analysis. *Page 73 of: HICSS '07: Proceedings of the 40th Annual Hawaii International Conference on System Sciences.* Washington, DC, USA: IEEE Computer Society.

Príloha A

Dátová vzorka MAPEKUS

V rámci výskumného projektu MAPEKUS[1] (Modeling and Acquisition, Processing anf Employing Knowledge About User Activities in the Internet Hyperspace) sme sa spolupodieľali na vytvorení dátovej vzorky z oblasti vedeckých publikácií, nad ktorou sme vykonali experimenty týkajúce sa prínosu tejto dizertačnej práce (kap. 5). V tejto časti preto podrobnejšie predstavujeme proces vzniku dátovej vzorky a jej vlastnosti.

A.1 Úvod

Dátová vzorka MAPEKUS zahŕňa a agreguje metainformácie o vedeckých publikáciách zverejnených na týchto portáloch:

- ACM (`http://www.acm.org/`),
- DBLP (`http://www.informatik.uni-trier.de/~ley/db/`),
- Springer (`http://www.springer.com/`).

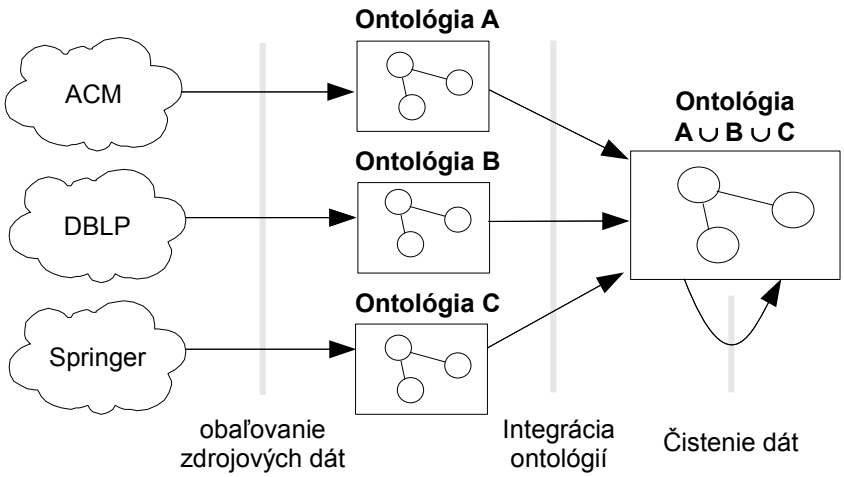
Cieľom bolo nielen zjednotiť údaje o článkoch, ich autoroch a súvisiacich informáciach, ale zabezpečiť ich jednotnú reprezentáciu a identifikovať zhodných autorov v priestore troch zdrojových databáz. Keďže cieľom je ukladať iba metainformácie o publikáciách, pre ukladanie dát sme zvolili ontologické úložisko.

[1] `http://mapekus.fiit.stuba.sk`

Z hľadiska dát sú v centre pozornosti publikácie. Na tieto nadväzujú pridružené typy entít, ako autori publikácií, kľúčové slová, zaradenie do klasifikačnej schémy a text abstraktu. Rozsah zdrojových údajov je uvedený v tabuľke A.1. Obr. A.1 znázorňuje priebeh spracovania údajov.

Tabuľka A.1: Počty inštancií z databáz ACM, DBLP a Springer.

	Zdroj dát		
Inštancia	ACM	DBLP	Springer
Autor	126 589	69 996	57 504
Organizácia	17 161	—	6 232
Publikácia	48 854	47 854	35 442
Kľúčové slovo	49 182	—	—
Referencia	454 997	—	—



Obrázok A.1: Hlavný proces spracovania a integrácia dát.

A.2 Štruktúra dátovej vzorky

V tejto časti je opísaná ontológia pre oblasť vedeckých publikácii, ktorá bola navrhnutá a vytvorená na základe charakteristiky domény vedeckých publikácii.[2].

A.2.1 Importované ontológie

V rámci ontológie publikácii sú použité dve externé ontológie a to `Party` a `Region`.

Party

Slúži pre opis rôznych druhov organizácii a jednotlivcov. Z hľadiska nami spracovávanej domény je dôležitá pre opis autorov, editorov alebo vydavateľov.

Region

Slúži na reprezentáciu geografických oblastí a krajín a používa sa napríklad pri opise pôvodu publikácie alebo miesta konania konferencie.

A.2.2 Triedy a ich dátové väzby

V tejto časti sú opísané všetky triedy ontológie spoločne s ich dátovými atribútmi. Tento postup sme zvolili preto, lebo dátové atribúty sa týkajú vždy len jednej triedy a sú podstatnou časťou jej charakteristiky. Všetky dátové väzby sú funkcionálne, to znamená, že pre každú inštanciu existuje vždy len jeden atribút.

Trieda Publication a jej podtriedy

Je to bázová trieda plne definovaná svojimi podtriedami, z ktorej sú odvodené všetky druhy publikácii opísaných v ontológii. Obsahuje atribúty uvedené v tab. A.2.

[2]Obsah kapitoly je prevzatý z výskumnej správy, časť z nej je dostupná na adrese `http://mapekus.fiit.stuba.sk/other/mapekus_domainOnto.pdf` [cit. 2009-09-10]

Tabuľka A.2: Atribúty triedy `Publication`.

Názov atribútu	*Dátový typ*	*Opis*
Title	Ret'azec	Názov publikácie
Year	Číslo	Rok vydania
Month	Číslo (hodnoty od 1 do 12)	Mesiac vydania
Day	Číslo (hodnoty od 1 do 31)	Deň vydania
FirstPage	Číslo	Strana, na ktorej publikácia začína
LastPage	Číslo	Strana, na ktorej publikácia končí
Abstract	Ret'azec	Abstrakt
Source	Ret'azec	Zdroj publikácie
Web	Ret'azec	Adresa zdroja

Hierarchia podtried triedy `Publication`, čiže hierarchia všetkých publikácii v ontológii, sa nachádza na obr. A.2. Jednotlivé triedy predstavujú uzly grafu reprezentované obdĺžnikom a sú spojené orientovanými hranami zobrazenými formou šípky znázorňujúcimi generalizačný vzt'ah. Túto notáciu budeme používat' aj pri d'alších obrázkoch hierarchií tried.

Nasleduje opis jednotlivých podtried a ich d'alších dátových väzieb.

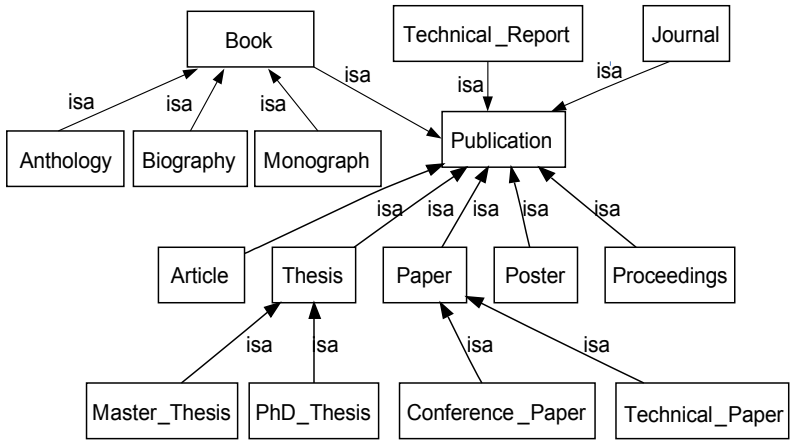
Article < Publication

Trieda reprezentujúca článok napríklad v časopise.

Book < Publication

Predstavuje knihu a obsahuje d'alšie podtriedy:

- `Anthology`: Zbierka prác od rôznych autorov,
- `Biography`: Životopisné dielo,
- `Monography`: Odborné dielo opisujúce jednu problematiku.



Obrázok A.2: Hierarchia publikácií.

Špecifické atribúty:

Názov atribútu	*Dátový typ*	*Opis*
ISBN	Ret'azec	Medzinárodné identifikačné číslo knihy

Journal < Publication

Reprezentuje periodikum určené pre odbornú verejnost'. Obsahuje tieto špecifické atribúty:

Názov atribútu	*Dátový typ*	*Opis*
Number, volume	Ret'azec	Identifikačné čísla periodika

Paper < Publication

Predstavuje odborný príspevok. Delí sa na d'alšie podtriedy:

- `Conference paper`: konferenčný odborný príspevok,
- `Technical paper`: odborný príspevok opisujúci napríklad technické aspekty určitého systému.

Proceedings < Publication

Súbor odborných príspevkov vydaných v kontexte nejakej konferencie alebo iného stretnutia odbornej verejnosti. Obyčajne sa vydáva v knižnej podobe.

Poster < Publication

Reprezentuje plagáty s odbornou tematikou.

Technical report < Publication

Formálna správa, ktorá opisuje prínos v oblasti aplikovaného výskumu, poukazuje na detaily a výsledky riešenia nejakého vedeckého problému.

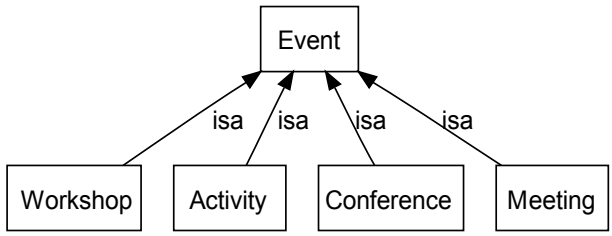
Thesis < Publication

Práca obsahujúca výsledky výskumu vypracovaná kandidátom na titul v rámci štúdia. Člení sa na podtriedy:

- `MasterThesis`: práca vypracovaná na druhom stupni vysokoškolského štúdia,
- `PhDThesis`: práca vypracovaná na tret'om (doktorandskom) stupni vysokoškolského štúdia.

Trieda Event a jej podtriedy

Trieda `Event` a jej podtriedy slúžia na reprezentáciu nejakej udalosti v rámci odbornej komunity. Hierarchia podtried je znázornená na obr. A.3.



Obrázok A.3: Hierarchia triedy `Event`.

Trieda sa člení na tieto podtriedy:

- `Activity`,
- `Conference`: konferencia,
- `Meeting`: odborné stretnutie,
- `Workshop`: udalosť zahrňujúca získavanie poznatkov, diskusiu a reakcie viažúca sa k nejakej odbornej téme.

Trieda `Event` obsahuje tieto atribúty:

Názov atribútu	*Dátový typ*	*Opis*
StartDate	Dátum	Dátum, kedy udalosť začína
EndDate	Dátum	Dátum, kedy udalosť končí
Web	Reťazec	Adresa zdroja o danej udalosti

Hierarchia triedy IndexTerms

Trieda `IndexTerm` a jej podtriedy predstavujú rozdelenie disciplín informačnej vedy do hierarchickej štruktúry. Táto klasifikácia bola prebratá z digitálnej knižnice ACM[3]. Obr. A.4 znázorňuje stromovú štruktúru jednotlivých podtried, pričom pre zachovanie prehľadnosti je do ďalšej úrovne rozvedená vždy len jedna vetva. Listy stromu už nereprezentujú triedy, ale inštancie.

Project

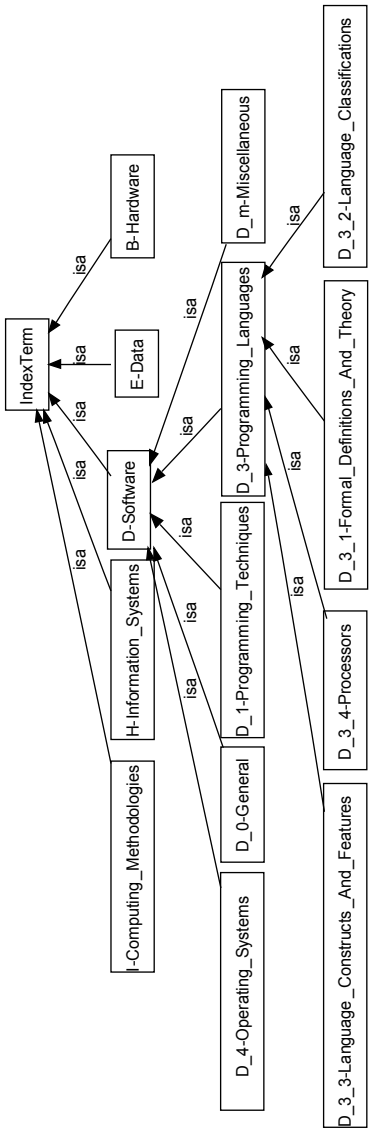
Trieda reprezentujúca nejaký projekt. Názov projektu je uložený v názve inštancie.

Author

Trieda predstavujúca autora minimálne jednej publikácie. Je odvodená od triedy `Person` reprezentujúcej všeobecne osobu, ktorá sa nachádza v importovanej ontológii `Party`. Obsahuje tieto zdedené atribúty:

Názov atribútu	*Dátový typ*	*Opis*
GivenName	Reťazec	Krstné meno
FamilyName	Reťazec	Priezvisko

[3] `http://www.acm.org/about/class/ccs98-html`



Obrázok A.4: Hierarchia triedy `IndexTerm`.

Editor

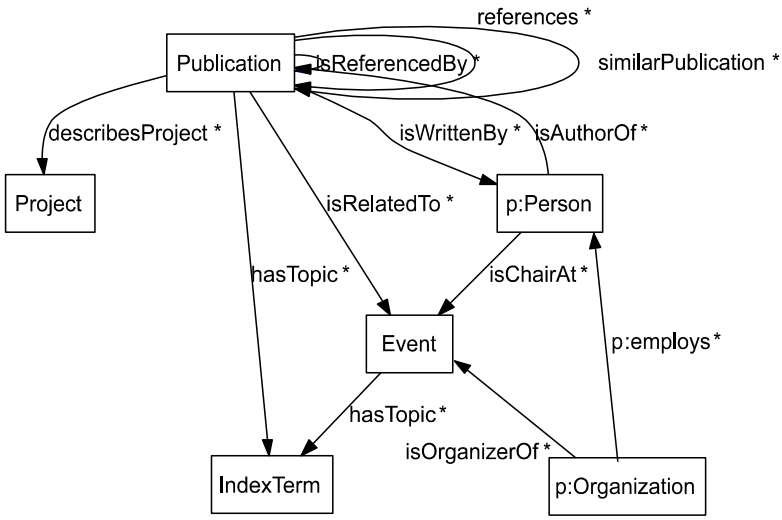
Trieda reprezentujúca editora minimálne jednej publikácie. Podobne ako trieda `Author` je odvodená od importovanej triedy `Person` a obsahuje rovnaké zdedené atribúty.

Publisher

Predstavuje vydavateľa minimálne jednej publikácie. Je odvodená od triedy `Organization` importovanej v rámci ontológie `Party`.

A.2.3 Väzby medzi triedami

Najvýznamnejšie väzby sú spoločne s triedami zobrazené na obr. A.5.



Obrázok A.5: Hierarchia triedy `Event`.

Príloha B

Dátová vzorka foaf.sk

Portál `http://foaf.sk/` vznikol z vlastnej iniciatívy v spolupráci s Jánom Suchalom[1]. Portál bol spustený v roku 2008 s cieľom vytvoriť alternatívu k oficiálnemu portálu Obchodného registra (`http://orsr.sk/`) spravovanému Ministerstvom spravodlivosti SR. Prínos spočíva v zjednodušení prehľadávania väzieb medzi osobami s spoločnosťami.

Dátová vzorka foaf.sk pozostáva z dvoch zložiek: graf podnikateľov a firiem a logy používateľov portálu. V nasledujúcich častiach sa venujeme obom dátovým vzorkám.

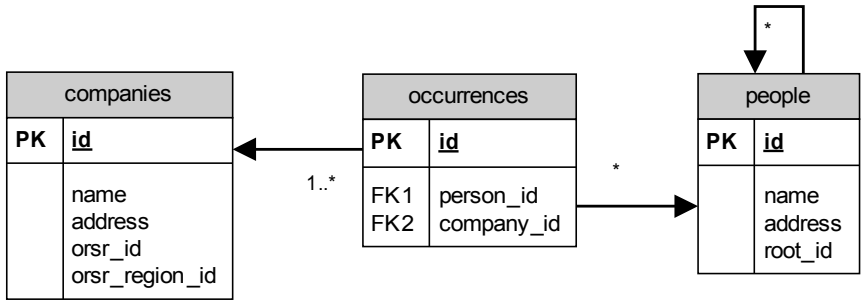
B.1 Graf obchodného registra

Záznamy z obchodného registra Slovenskej republiky (`http://orsr.sk/`) sme stiahli pomocou inkrementácie URL. Jednotlivé záznamy zodpovedajú spoločnostiam. Pomocou ručne vytvoreného obaľovača (ang. wrapper) sme z týchto záznamov extrahovali graf, jeho databázová schéma je na obr. B.1.

V januári 2009 mal tento graf 168 000 vrcholov typu `spoločnosť` (tabuľka *companies*), 352 000 vrcholov typu `osoba` (tabuľka *people*) a 460 000 hrán medzi nimi (tabuľka *occurrences*).

Tento graf má typické vlastnosti sietí malého sveta, zdôrazňuje ich najmä exponenciálne rozdelenie, ktorým sa riadia prinajmenšom nasledovné vzťahy:

[1] `http://www2.fiit.stuba.sk/~suchal/`

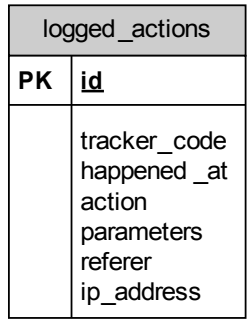


Obrázok B.1: Entitno-relačná schéma grafu obchodného registra.

- spoločnosti združené podľa počtu osôb činných v spoločnosti,
- komponenty grafu (neprepojené subgrafy) združené podľa počtu vrcholov na komponent.

B.2 Návštevníci portálu

Po nasadení portál foaf.sk pritiahol návštevnosť, napr. za január 2009 mal portál približne 30 000 unikátnych návštevníkov, ktorí za toto obdobie vykonali 280 000 akcií (pageviews).Tieto akcie sme zaznamenávali do tabuľky, ktorej schéma je na obr. B.2.



Obrázok B.2: Entitno-relačná schéma záznamov návštevnosti.

Existujú tri typy akcií používateľa portálu (atribút *action*, obr. B.2):

- `show_company`: zobrazenie spoločnosti (obr. B.3),
- `show_person`: zobrazenie osoby,
- `search`: fulltextové vyhľadávanie osoby alebo spoločnosti.

Obrázok B.3: Náhľad na vrchol typu `spoločnosť` na portáli `http://foaf.sk/`.

Okrem prehľadnejšej orientácie a vyhľadávania je pridanou hodnotou portálu zobrazovanie širšieho okolia vrcholu (zoznamy blízkych ľudí a firiem na obr. B.3, ktoré sa pre zadaný vrchol počíta pomocou lokálneho ohodnocovacieho algoritmu s názvom šírenie aktivácie. Jeho pseudokód je uvedený v kap. 6.

Príloha C

O autorovi

Peter Vojtek sa narodil 19. septembra 1982 v Bratislave. Titul inžinier získal v roku 2007 na Fakulte informatiky a informačných technológií Slovenskej technickej univerzity v Bratislave s diplomovou prácou *Identifikácia prirodzených jazykov v textových dokumentoch.*

Od roku 2007 po súčasnosť je študentom doktorandského štúdia na Fakulte informatiky a informačných technológií Slovenskej technickej univerzity v Bratislave, kde sa venuje oblasti dolovania v dátach, klasifikačným metódam a spracovaniu grafovo orientovaných dát a spolupracuje na viacerých výskumných projektoch. V priebehu štúdia publikoval viacero vedeckých príspevkov, ich zoznam je v prílohe D. Od roku 2006 je členom výskumnej skupiny Personalised Web (PeWe, `http://pewe.fiit.stuba.sk/`).

Medzi jeho ďalšie záujmy patrí vytrvalostný beh, nezávislé cestovanie, digitálna fotografia a spracovanie geografických dát.

Príloha D

Publikácie autora

D.1 Medzinárodné vedecké konferencie

1. Vojtek, P., & Bieliková, M. 2010. Homophily of Neighborhood in Graph Relational Classifier. *Pages 721–730 of:* Geffert, V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., & Bieliková, M. (eds), *SOFSEM 2010: 36th Conf. on Current Trends in Theory and Practice of Computer Science, Spindleruv Mlyn, Czech Republic, 2010, Proc.* LNCS, vol. 5901. Springer.

2. Vojtek, P., & Bieliková, M. 2009. Moderated Class-membership Interchange in Iterative Multi-relational Graph Classifier. *Pages 229–238 of:* Snášel, V., Szczepaniak, P.S., Abraham, A., & Kacprzyk, J. (eds), *AWIC 2009: Proc. of the 6th Atlantic Web Intelligence Conf.* AISC, vol. 67. Springer.

3. Frivolt, G., Suchal, J., Vesely, R., Vojtek, P., Vozár, O., & Bieliková, M. 2008. Creation, Population and Preprocessing of Experimental Data Sets for Evaluation of Applications for the Semantic Web. *Pages 684–695 of:* Geffert, V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., & Bieliková, M. (eds), *SOFSEM 2008: 34th Conf. on Current Trends in Theory and Practice of Computer Science, Nový Smokovec, Slovakia, 2008, Proc.* LNCS, vol. 4910. Springer.

D.2 Lokálne a národné vedecké konferencie

1. Barla, M., Kompan, M., Suchal, J., Vojtek, P., Zeleník , D., & Bieliková, M. 2010. Recommendation of News (in Slovak). *In: Znalosti 2010.* Fakulta managementu Vysoké školy ekonomické, Jindřichuv Hradec.
2. Vojtek P., & Bieliková M.. 2009. Local Graph Ranking in Social Network of Slovak Companies (in Slovak). *In: Proc. of 4th Workshop on Intelligent and Knowledge Oriented Technologies (WIKT 2009).*
3. Suchal, J., & Vojtek, P. 2009. Navigation is Social Network of Slovak Companies Register (in Slovak). *Pages 145–151 of: DATAKON 2009, Proc. of the Annual Database Conf. Srní, Czech Republic.*
4. Vojtek P., & Bieliková M. 2008. Increasing the Robustness of Relational Classifier in Datasets with Low Homophily (in Slovak). *In:* Návrat, P., & Vranić, V. (eds), *Proc. of 3rd Workshop on Intelligent and Knowledge Oriented Technologies (WIKT 2008).*
5. Gatial, E., Balogh, Z., Hluchý, L., & Vojtek, P.. 2007. Identification and Acquisition of Domain dependent Internet Resources (in Slovak). In: Návrat, P., Bartoš, P., Bieliková, M., Hluchý, L., and Vojtáš, P., (eds),*Tools for Acguisition, Organisation and Presenting of Information and Knowledge: Research Project Workshop*, Poľana, 2007.
6. Suchal, J., Vojtek, P., & Frivolt, G. 2007. Interactive Navigation in Large Graphs based on Clustering (in Slovak). In: Návrat, P., Bartoš, P., Bieliková, M., Hluchý, L., and Vojtáš, P., (eds), *Tools for Acguisition, Organisation and Presenting of Information and Knowledge: Research Project Workshop*, Poľana, 2007.
7. Laclavík, M., Ciglan, M., Šeleng, M., Krajčí, S., Vojtek, P., & Hluchý, L. 2007. Semi-automatic Semantic Annotation of Slovak Texts. *Pages 126–137 of: Proc. of 4th International Seminar on NLP, Computational Lexicography and Terminology (SLOVKO 2007).*
8. Vojtek, P., & Bieliková, M. 2007. Comparing Natural Language Identification Methods based on Markov Processes. *Pages 271–281 of: Proceedings of 4th International Seminar on NLP, Computational Lexicography and Terminology (SLOVKO 2007).*

D.3 Kapitoly v knihách

1. Bieliková, M., & Návrat, P. 2007. *Advanced methods of designing program systems. Edition of research texts in in-formatics and information technologies*, 226 strán. Faculty of informatics and information technologies STU Bratislava.

D.4 Študentské vedecké konferencie

1. Vojtek, P. 2009. How Graph Generated from User Logs Extends Collective Classifier. *Pages 224–231 of:* Bieliková, M. (ed.), *IIT.SRC 2009: Student Research Conf.* Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.
2. Vojtek, P. 2008. Moderate Iterative Multi-Relational Classification. *Pages 269–276 of:* Bieliková, M. (ed), *IIT.SRC 2008: Student Research Conf.* Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.
3. Vojtek, P. 2007. Improving Text Categorization Based on Markov Models. *In:* Bieliková, M. (ed), *IIT.SRC 2007: Student Research Conf.* Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.
4. Vojtek, P. 2006. Natural Language Identification in the World Wide Web. *Pages 153–159 of:* Bieliková, M. (ed), *IIT.SRC 2006: Student Research Conf.* Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.