MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

DOBÝVÁNÍ ZNALOSTÍ Z ASTRONOMICKÝCH DAT

JAROSLAV VÁŽNÝ

BRNO, JARO 2009

Prohlášení

Prohlašuji, že tato bakalářská práce je mým původním autorským dílem, které jsem vypracoval samostatně. Všechny zdroje, prameny a literaturu, které jsem při vypracování používal nebo z nich čerpal, v práci řádně cituji s uvedením úplného odkazu na příslušný zdroj.

V Brně dne:

Poděkování

Děkuji Filipovi Hrochovi, vedoucímu této práce, za jeho pomoc a trpělivost. Za přečtení a konstruktivní kritiku děkuji Báře Mikulecké, Boženě Kováčové, Lence Matěchové a Josefovi Paculovi. Petrovi Šafaříkovy a Jiřímu Šperkovi děkuji za užitečné typografické připomínky.

Následujícím institucím pak za to, že poskytly data a nástroje, které umožnily vznik této práce.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is `http://www.sdss.org/`.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

This research has made use of data obtained from the High Energy Astrophysics Science Archive Research Center (HEASARC), provided by NASA's Goddard Space Flight Center.

This research has made use of the SIMBAD database, operated at CDS, Strasbourg, France.

ANOTACE

Bakalářská práce *Dobývání znalostí z astronomických dat* pojednává o metodách získávání a dobývání znalostí v kontextu astrofyziky. Diskutovány jsou možné zdroje dat, metody dobývání znalostí (klasifikační stromy a shluková analýza) a nezbytné související technologie. Praktická část je zaměřena na klasifikaci a hledání nových objektů typu BL Lac na základě charakteristik spektra v archívu Sloanovy digitální přehlídky oblohy. Využíván je algoritmus J48. Představeno je šest nalezených objektů, které byly následně vyhledány ve vědeckých článcích. Na základě výsledků lze předpokládat, že při zahrnutí dalších charakteristik (např. rádiové vyzařování) lze metodu zpřesnit a nalézt dosud neznámé objekty typu BL Lac.

ANNOTATION

My Bachelor's Thesis *Data Mining from Astronomical Data* deals with data mining methods in the contex of astrophysics. Possible data sources, data mining methods (classification trees and clustering) and corresponding technologies are discussed. Practical part of this work is focused on classification BL Lac objects based on spectral characteristics in the Sloan Digital Survey data warehouse. The algoritm J48 is used for classification purpose. There are introduced six classified objects. It might be possible to discover new BL Lac type objects when other parameters are considered (e.g. radio emission).

KLÍČOVÁ SLOVA

Dobývání znalostí, Sloanova digitální přehlídka oblohy, virtuální observatoř, BL Lac

KEYWORDS

Data mining, Sloan Digital Sky Survey, Virtual Observatory, BL Lac

Obsah

Úvod

S rozvojem pozorovací a výpočetní techniky se současná astrofyzika stává vědou velice bohatou na data a lze očekávat, že tento trend nadále poroste. Jednotlivé datové ostrůvky se nakonec slijí pod pláštěm virtuálních observatoří. Naše schopnosti na poli dobývání znalostí zde budou hrát klíčovou roli.

Tycho Brahe napozoroval za celý svůj život cca 500 kB[1] dat, kdežto dnešní přehlídky oblohy obsahují terabajty a není daleko doba, kdy budou naše přístroje produkovat petabajty dat[2]. Naše porozumění přírodě a jejím zákonitostem však neroste úměrně s našimi schopnostmi tato data získávat a ukládat. Je třeba najít metody, které nám umožní vyhledávat souvislosti a potenciálně užitečné informace. V práci jsem se zabýval právě takovými metodami v kontextu astrofyziky.

V první části jsem se zaměřil na možné zdroje dat, v druhé na dobývání znalostí. Třetí část je praktická a obsahuje vlastní výzkum a sice hledání objektů typu BL Lac v datovém skladu Sloanovy digitální přehlídky oblohy. Závěrečná kapitola obsahuje popis použitých technologií.

Pro lepší orientaci jsem na začátek každé kapitoly vložil obrázek zobrazující její strukturu. Celková struktura práce je pak na obrázku 1.

Úspěch je nemyslitelný bez použití informačních technologií, ty však s sebou přináší určitá úskalí, která se pokusím nastínit.

1. **Otevřenost a standardy**

   Dovolím si tvrdit, že úspěch vědy závisí do značné míry na její otevřenosti. Takové nároky je nutné klást i na použité softwarové nástroje. S uzavřeným softwarem a daty uloženými v nedostupných databázích zůstane tento potenciál nevyužit. V této práci jsou důsledně použity otevřené programy a formáty.

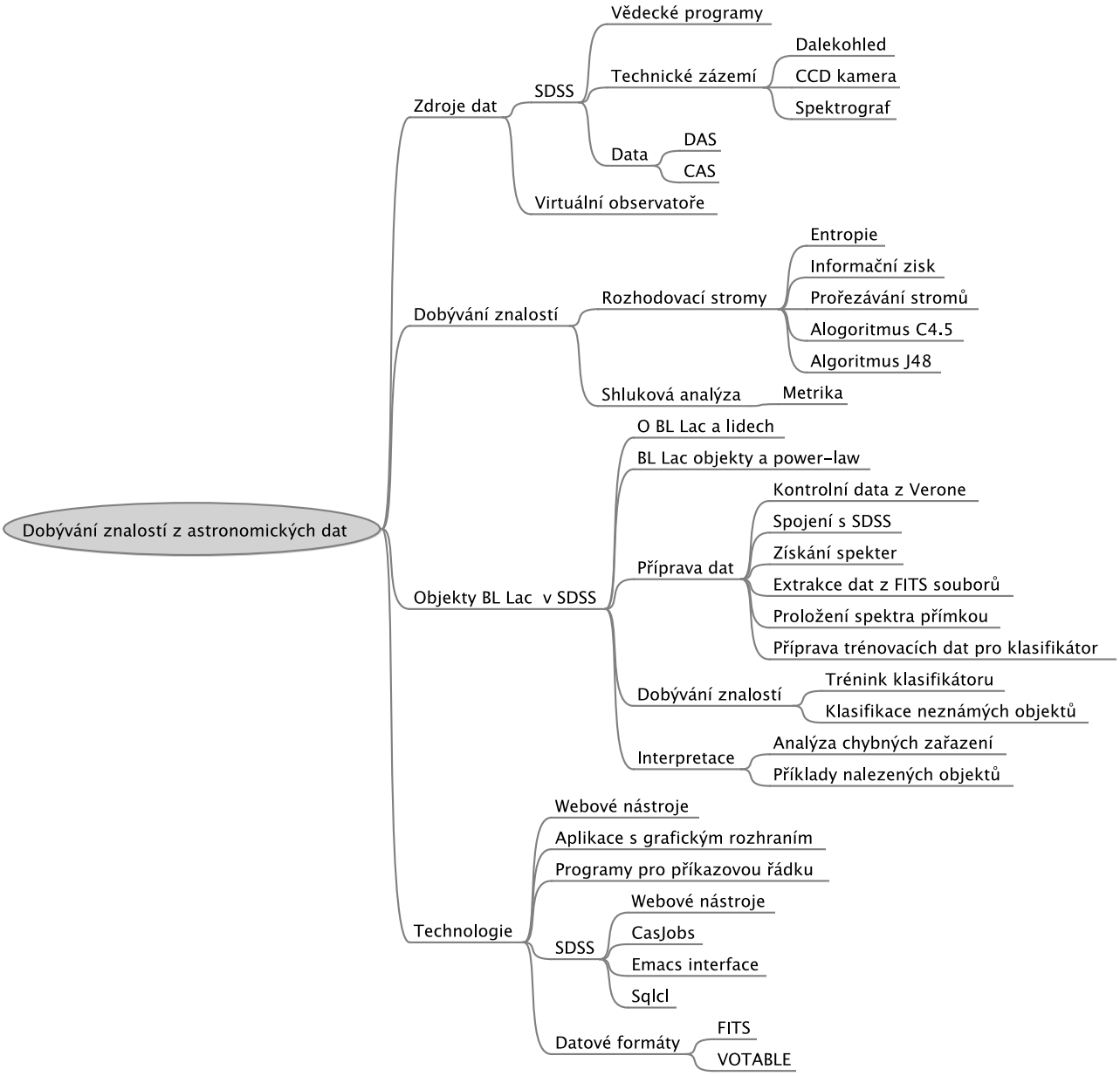
2. **Porozumění technologiím**

   Informační technologie mohou představovat účinný nástroj ve vědecké práci, naopak jejich nezvládnutí znamená obrovský handicap. Z tohoto důvodu je třeba vzdělání v této oblasti věnovat daleko větší pozornost, než se děje v současnosti. Počítačová gramotnost vědce musí být na zcela jiné úrovni, než je tomu u běžné populace.

---

1. RADDICK (2006)
2. DJORGOVSKI et al. (2001)

Poznámka k překladům: Většina zdrojů je v anglickém jazyce. Některé výrazy a slovní spojení nemají ustálené české ekvivalenty. V oblasti dobývání znalostí se držím terminologie užité v BERKA (2003). Pokud výraz překládám, uvádím i jeho anglický originál. Překládat některé výrazy by však bylo kontraproduktivní (jedná se většinou o softwarové nástroje a technologie např. SkyServer, CasJobs atp.).



Obrázek 1: *Struktura práce*

Slovník použitých pojmů a zkratek

ARFF Attribute-Relation File Format (vstupní formát pro program Weka)

ASCII American National Standard Code for Information

C4.5 algoritmus pro generování rozhodovacího stromu vytvořený Rossem Quinlanem

CAS Catalog Archive Server (databázová část dat SDSS přístupná pomocí SQL)

CasJobs program pro dávkové zpracování dotazů v SDSS

DAS Data Archive Server (souborová část dat SDSS)

FITS Flexible Image Transport System (formát pro výměnu astronomických dat)

GiB $(2^{10})^3 = 2^{30} = 1073741824$B (IEC60027-2 (2000))

GNU GPL General Public License (licence pro svobodný software)

J48 svobodná implementace algoritmu C4.5 v programu Weka

MiB $(2^{10})^2 = 2^{20} = 1048576$B (IEC60027-2 (2000))

RDBMS Relational database management system (relační databázový stroj)

SDSS Sloan Digital Sky Survey (Sloanova digitální přehlídka oblohy)

SkyServer přístupový bod k datům SDSS

SQL Structure Query Language (dotazovací jazyk používaný v RDBMS)

TDIDT metoda tvorby rozhodovacího stromu

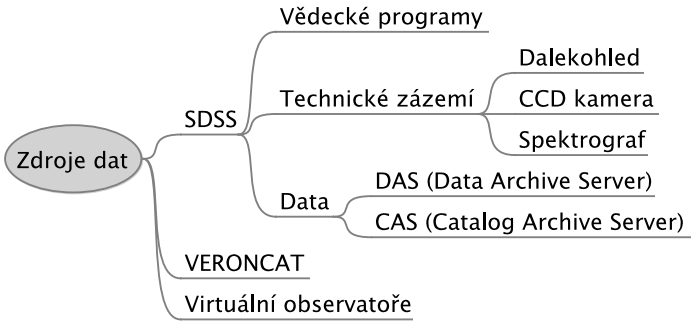
VERONCAT katalog kvasarů a AGN

Weka software pro dobývání znalostí a strojové učení

XML Extensible Markup Language (rozšiřitelný značkovací jazyk)

Kapitola 1

Zdroje dat



Obrázek 1.1: *Struktura kapitoly: Zdroje dat*

1.1 Sloanova digitální přehlídka oblohy

Sloan Digital Sky Survey (SDSS) je společný projekt 25 institucí včetně NASA, Princetonské univerzity a Fermilabu, financovaný primárně z nadace Alfreda P. Sloana.

Vědecké programy

SDSS-I: 2000–2005

V první fázi projektu bylo zmapováno více než 8000 čtverečných stupňů oblohy v pěti vlnových délkách. Bylo získáno spektrum kvasarů a galaxií z oblasti 5700 čtverečných stupňů.

SDSS-II: 2005–2008

- **The Sloan Legacy Survey**

  Dokončení snímkování a spektroskopie z SDSS-I. Výsledný soubor obsahuje 230 miliónů objektů v 8400 čtverečných stupních a spektra u 930 tisíc galaxií, 120 tisíc kvasarů a 225 tisíc hvězd.

- **SEGUE** (the **S**loan **E**xtension for **G**alactic **U**nderstanding and **E**xploration)

  Měření struktury a historie Mléčné dráhy. Nasnímáno 3500 čtverečných stupňů a spektrum 240 tisíc hvězd.

- **The Sloan Supernova Survey**

  Opakované snímání 300 čtverečných stupňů v jižní části rovníku pro objevení a měření supernov a jiných proměnných objektů. Bylo objeveno téměř 500 spektroskopicky potvrzených supernov typu Ia. Tyto znalosti byly využity k určení rychlosti rozpínání vesmíru v posledních čtyřech miliardách let.

SDSS-III: 2008–2014

- **BOSS B**aryon **O**scillation **S**pectroscopic **S**urvey

  Měření červeného posuvu galaxií v čáře *Lyman* $-\alpha$ a 160 tisíc kvasarů s velkým červeným posuvem. To umožní určit vzdálenosti s přesností 1,0 % pro $z = 0{,}35$, 1,1 % pro $z = 0{,}6$ a 1,5 % pro $z = 2{,}5$. Takto přesné údaje umožní například studování vlastností temné hmoty, porozumění původu struktur ve vesmíru, evoluci galaxií atp.

- **SEGUE-2**

  Mapování struktury, kinematiky a chemického vývoje vnější části galaktického disku a hala.

- **APOGEE**

  Použití infračervené spektroskopie s vysokým rozlišením umožní „vidět“ skrz prach do vnitřních částí galaxie.

- **MARVELS**

  Měření radiálních rychlostí 11 tisíc hvězd pro získání informací o vlastnostech obřích planet.

Technické zázemí

Dalekohled

Dvouapůlmetrový teleskop je umístěn na observatoři **A**pache **P**oint **O**bservatory v Novém Mexiku ve výšce 2788 metrů nad mořem. Vzhledem k faktu, že musí mapovat velkou část oblohy, byla zvolena velice důmyslná konstrukce: přístroj obsahuje dvě zrcadla o průměrech 2,5 a 1,08 metrů. Světelný signál putuje skrz primární zrcadlo přes korekční čočky do CCD kamery. Dalekohled umožňuje ostrý obraz oblasti o ploše tří stupňů (plocha 30 měsíců v úplňku). Uložení dalekohledu je také netypické (viz obr 1.2). Vnější kryt však způsobuje problémy s teplem, které způsobuje turbulence vzduchu a rozostřuje snímky.

Obrázek 1.2: *Dalekohled projektu SDSS*

CCD kamera

Kamera obsahuje 30 CCD čipů, řazených po pěti do sloupců. Každý čip má rozlišení 2048×2048 obrazových bodů (celkově tedy cca 120 megapixelů). Každý řádek má jiný optický filtr s vlnovými délkami 354, 476, 628, 769 a 925 nanometrů. Vše je uloženo ve vakuové komoře a chlazeno tekutým dusíkem na teplotu 190 kelvinů.

Spektrograf

Spektrograf se skládá z hliníkové desky, v níž je vyvrtáno 640 děr, do nichž jsou vloženy optické kabely. Polohy odpovídají pozorovaným objektům. Takto je možné pořídit simultánně 640 spekter. Výsledné spektrum je zaznamenáno na CCD v rozsahu od 380 do 920 nanometrů. Za jednu noc je vyměněno 6–9 desek (tj. cca 5000 objektů).

Data

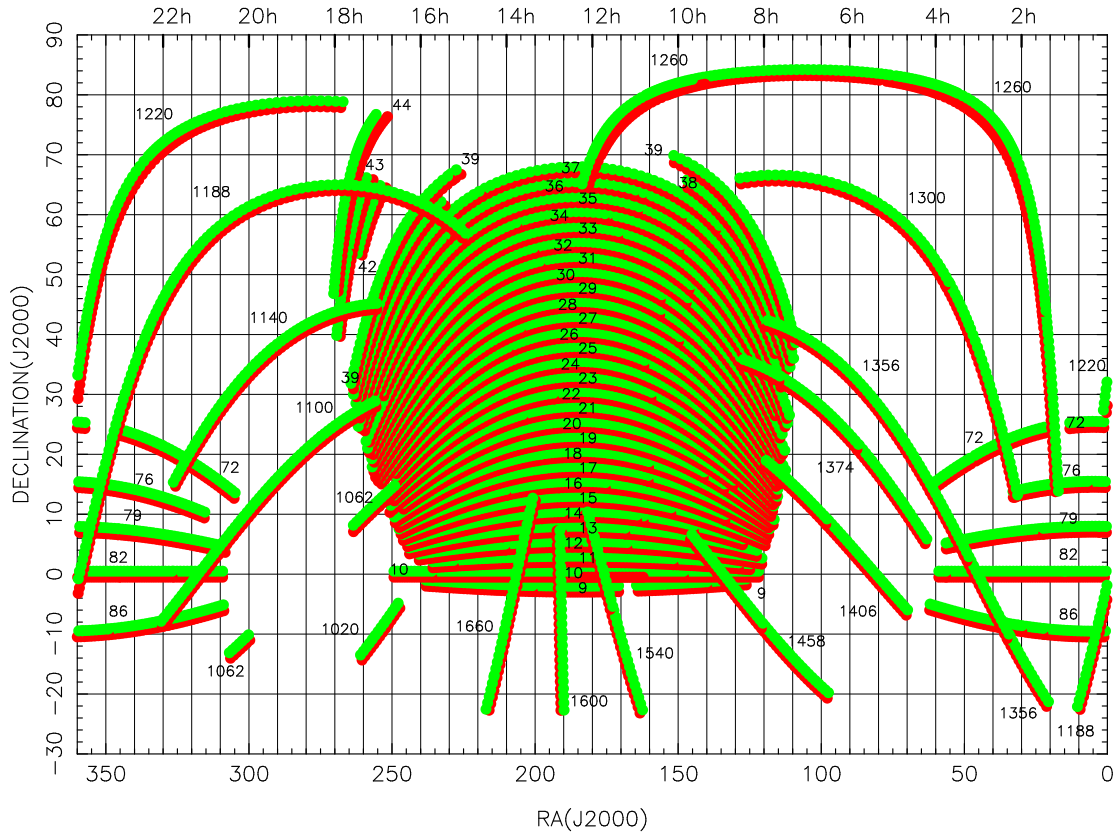
Datové výstupy jsou dostupné po blocích v tzv. „**SSD Data Release**“ DR+číslo. Posledním je DR7, uvolněný v roce 2008 (DR8 je plánovaný na rok 2010). Jednotlivé bloky jsou dostupné na `http://cas.sdss.org/drX`, kde X je číslo vydání. To obsahuje dva typy dat: datové (FITS soubory) a databázové (dostupné v SQL databázi).Vše dále uvedené platí pro DR7.

DAS (**D**ata **A**rchive **S**erver)

Obsahuje 15,7 TB obrazových dat a 26,8 TB ostatních dat (katalogy, masky, jpeg obrázky) Do tohoto archívu lze přistupovat pomocí webového rozhraní (`http://das.sdss.org/`) nebo lze použít program **wget** resp. **curl**. Je možné také použití programu **rsync** pro opakované stahování.

CAS (**C**atalog **A**rchive **S**erver)

Je 18 TB velká SQL databáze. Přístup je pomocí webového rozhraní (`http://cas.sdss.org/dr7/en/`) nebo programů pro příkazovou řádku.



Obrázek 1.3: *Pokrytí oblohy projektem SDSS. Červeně jsou označeny obrazová data, zeleně spektroskopická*

1.2 VERONCAT - Veron Catalog of Quasars & AGN

Tento katalog obsahuje ve své dvanácté edici[1] 85221 kvasarů, 1122 BL Lac objektů a 21737 aktivních galaxií (včetně 9628 Seyfertových galaxií). Katalog je dostupný jak k přímému stažení, tak přes rozhraní virtuální observatoře.

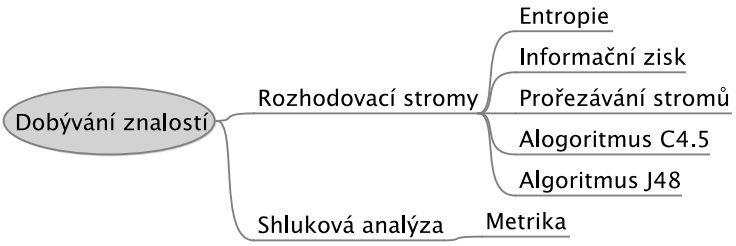
1.3 Virtuální observatoře

Virtuální observatoř (VO) je koncept přístupu k astronomickým datům. Skládá se ze standardů, datových archívů a nástrojů pro přístup a zpracování dat. Kromě jednotlivých národních a nadnárodních VO existuje mezinárodní aliance IVO (International Virtual Observatory Alliance), která udržuje a prosazuje standardy pro zajištění jednotného přístupu k datům.

1. VÉRON-CETTY and VÉRON (2006).

Kapitola 2

Dobývání znalostí



Obrázek 2.1: *Struktura kapitoly: Dobývání znalostí*

Dobývání znalostí (data mining) je dle definice, „netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat“ FAYYAD et al. (1996). Skládá se typicky z těchto fází: selekce, předzpracování, transformace a vlastního „dolování“ BERKA (2003).

Poznámka: Spojení „data mining“ je lépe překládat jako dobývání **z** dat, než dobývání dat. Ještě lépe pak dobývání znalostí (viz BERKA (2003)).

Existuje celá řada metod, které se využívají pro dobývání znalostí (např. rozhodovací stromy, shluková analýza, regresní analýza, genetické algoritmy, neutronové sítě, strojové učení a mnohé další). Vždy je třeba najít metodu, která je nejbližší řešenému problému. Není možné popisovat všechny tyto metody. Podíváme se podrobněji na ty, které byly zvažovány a použity v praktické části práce.

2.1 Rozhodovací stromy

K často používaným způsobem reprezentace znalostí patří rozhodovací stromy (taxonomie živočichů, b-stromy v relačních databázích atp.). Pro rozdělení do jednotlivých podskupin se používá metoda *rozděl a panuj* (divide and conquer), kdy se data rozdělují na stále menší skupiny tak, aby v takto vzniklých podmnožinách převládaly příklady jedné třídy. Při automatickém procesu je třeba vyřešit několik problémů:

- volbu atributů, podle kterých budeme data třídit,
- ukončení dělení stromu,
- zatížení testovacích dat šumem.

Pro automatické třídění se používá metoda TDIDT (top down induction of decision tree), která sestává z následujících kroků.

1. Zvol jeden atribut jako kořen dílčího stromu.
2. Rozděl data v tomto uzlu na podmnožiny podle hodnot zvoleného atributu a přidej uzel pro každou podmnožinu.
3. Existuje-li uzel, pro který nepatří všechna data do téže třídy, pro tento uzel opakuj postup od bodu 1, jinak skonči.

Kritická je zde volba atributu. Ten lze určit pomocí entropie, informačního zisku, poměrného informačního zisku a dalších veličin.

Entropie

Vyjadřuje míru neuspořádanosti, kromě svého původního fyzikálního významu má uplatnění i v teorii informací. Je definována jako:

$$H = -\sum_{t=1}^{T} p_t \log_2 p_t, \tag{2.1}$$

kde p_t je pravděpodobnost výskytu třídy t a T je počet tříd. Pro názornost uvádím příklad znázornění pro hod mincí na obrázku 2.2.

Informační zisk

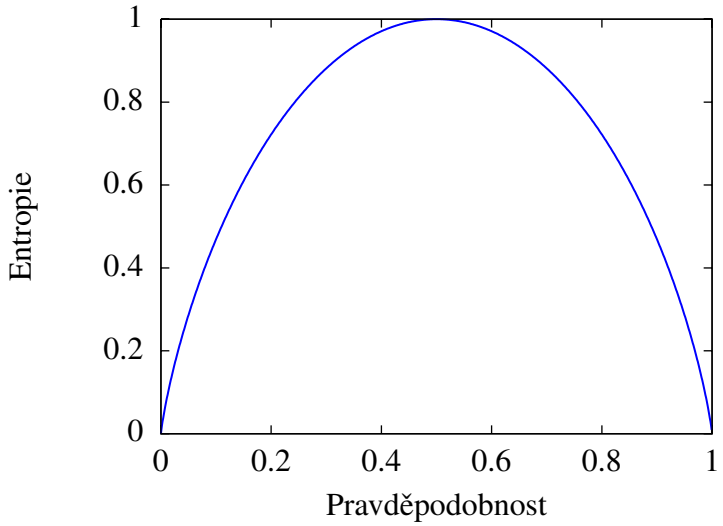
Je míra odvozená z entropie. Spočítá se jako rozdíl entropie $H(C)$ pro celá data a $H(A)$ pro uvažovaný atribut. Informační zisk udává redukci entropie způsobenou volbou atributu A.

$$\text{Zisk}(A) = H(C) - H(A), \tag{2.2}$$

$$H(C) = -\sum_{t=1}^{T} \frac{n_t}{n} \log_2 \frac{n_t}{n}. \tag{2.3}$$

Jako atribut vhodný pro větvení vybereme ten s maximální hodnotou informačního zisku. První člen rovnice (2.2) odpovídá entropii pro celá data a je konstantní. Rozdíl bude maximální v případě, kdy bude druhý člen minimální.

Další metody volby atributu pro větvení rozhodovacího stromu, jako je Gini index, poměrný informační zisk a další, jsou popsané v BERKA (2003).



Obrázek 2.2: *Závislost entropie na pravděpodobnosti při hodu mincí. Pokud je mince „férová“, pak je pravděpodobnost, že padne panna nebo orel,* 0,5 *a entropie maximální. V obou krajních případech (mince má obě strany stejné) je entropie rovna nule*

Prořezávání stromů

Pokud bychom měli šumem nezatížená data a postupovali důsledně podle algoritmu TDIDT (viz 2.1), dospěli bychom k bezchybné klasifikaci všech objektů. Takový stav však nemusí být ani žádoucí, ani možný. Jednak je výsledný strom příliš složitý a tudíž nepřehledný, dále jsou v praxi vstupní data zatížena šumem. Z těchto důvodů se v praktických implementacích požaduje, aby v koncovém uzlu převažovaly příklady jedné třídy. Jednou z metod, jak vytvořit takovýto redukovaný strom, je právě prořezávání stromů založené na pravidlech. Ta má následující fáze:

1. Převeď strom na pravidla.
2. Generalizuj pravidlo odstraněním podmínky za předpokladu, že dojde ke zlepšení odhadované správnosti.
3. Uspořádej prořezaná pravidla podle odhadované správnosti: v tomto pořadí budou pravidla použita pro klasifikaci.

BERKA (2003)

Systém C4.5

V praktické části této práce jsem použil algoritmus J48, jenž je implementací algoritmu C4.5[1]. Tato metoda tvorby rozhodovacího stromu využívá všechny výše zmíněné postupy. Navíc umožňuje práci s numerickými atributy, chybějícími hodnotami a bere do úvahy ceny za různá chybná rozhodnutí.

1. QUINLAN (1986)

2.2 Shluková analýza (clustering)

Tato metoda se snaží rozdělit objekty do přirozených skupin, na základě vzdálenosti mezi nimi. Vycházíme tedy z předpokladu, že umíme tuto vzdálenost měřit. Pro měření vzdálenosti se používají metriky. Přestože tato metoda nebyla nakonec v praktické části použita, považuji za rozumné ji zde zmínit.

Pojem metriky

Definice 2.2.1 *Metrickým prostorem nazýváme dvojici (P, ρ), kde P je libovolná neprázdná množina a zobrazení ρ: $P \times P \to \mathcal{R}^+$ splňuje pro každé x, y z P následující tři axiomy:*

- *(M1) $\rho(x, y) = 0$ právě když $x = y$ (axiom totožnosti);*
- *(M2) $\rho(x, y) = \rho(y, x)$ (axiom symetrie);*
- *(M3) $\rho(x, y) + \rho(y, z) \geq \rho(x, z)$ (trojúhelníková nerovnost).*

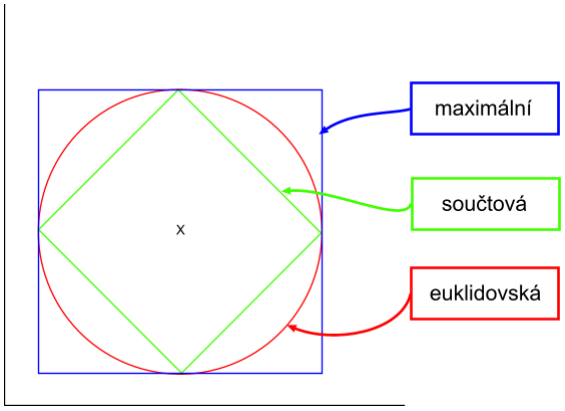
Zobrazení ρ nazýváme metrikou na P, prvky množiny P obvykle nazýváme body prostoru (P, ρ), číslo $\rho(x, y)$ nazýváme vzdáleností bodů x, y v prostoru (P, ρ). DOŠLÁ (2006).

Rovnice (2.4) představuje součtovou, (2.5) maximální a (2.6) eukleidovskou metriku.

$$\rho_1 = \sum_{k=1}^{n} |x_k - y_k|, \tag{2.4}$$

$$\rho_\infty = \max_{1 \leq k \leq n} |x_k - y_k|, \tag{2.5}$$

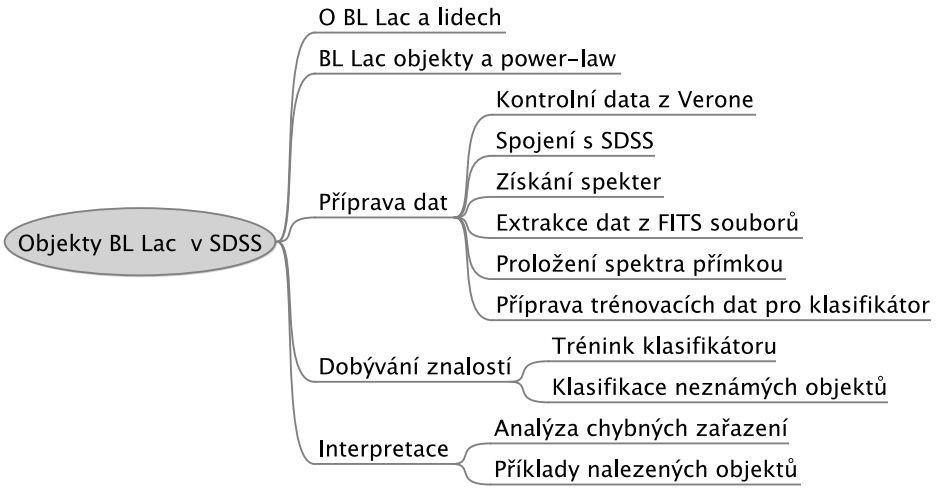
$$\rho_E = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}. \tag{2.6}$$



Obrázek 2.3: *Grafické znázornění metrik. Převzato z BERKA (2003)*

Kapitola 3

Klasifikace objektů typu BL Lac v SDSS



Obrázek 3.1: *Struktura kapitoly: Klasifikace objektů typu BL Lac v SDSS*

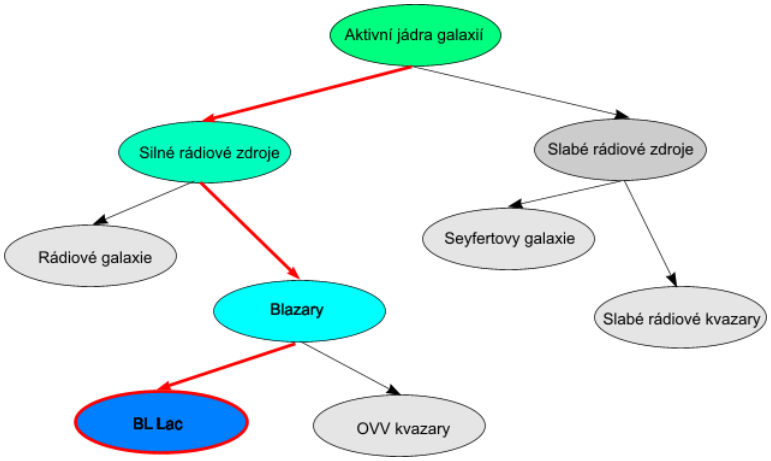
3.1 O BL Lac a lidech

V roce 1968 byl identifikován[1] velmi proměnný rádiový zdroj (VRO 42.22.01), nacházející se na stejném místě jako objekt BL Lacertae,[2] do té doby považovaný za proměnnou hvězdu. Podrobný výzkum ukázal, že tento objekt je opravdu zvláštní. Optické spektrum bylo bez čar, s kontinuem rostoucím do červené a infračervené oblasti, splňující mocninný zákon (power-law), emise vykazovala silnou lineární polarizaci. Podobné charakteristiky nebyly pozorované u hvězd. Ukázalo se ale, že se shodují s některými vlastnostmi nově objevených rádiových kvasarů. Další studium těchto objektů vedlo vědce k vytvoření zcela nové skupiny extra-galaktických objektů. Ta byla pojmenována podle prvního příslušníka BL Lacertae, zkráceně BL Lac. Klasický BL Lac objekt je silný rádiový zdroj s optickým kontinuem bez čar, vykazující silnou polarizaci a proměnnost (v řádu dní a více) (ROBSON (1996)).

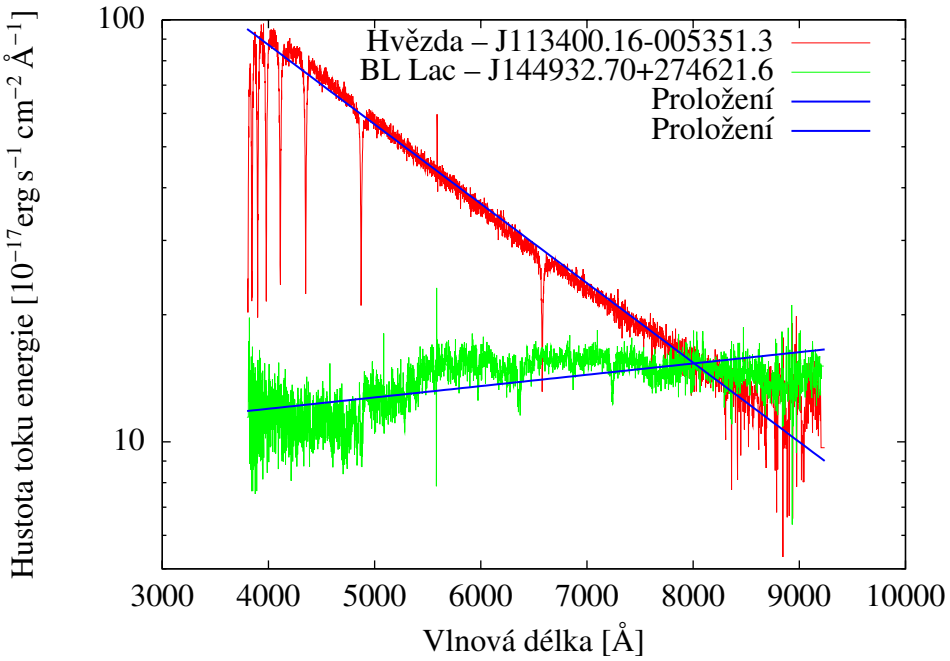
1. Johnem Schmittem na observatoři Davida Dunlapa.
2. Objevený Cuno Hoffmeisterem v roce 1929.

BL Lac jsou jednou ze skupin aktivních galaktických jader. Jejich zařazení v rámci této kategorie je na obrázku 3.2.

Základní myšlenka tedy zněla: Lze na základě jednoduchých charakteristik spektra rozlišit objekty typu BL Lac od ostatních objektů? Jak je vidět na obrázku 3.3, jsou spektra hvězdy a objektu poměrně dobře rozlišitelná.



Obrázek 3.2: *Rozdělení aktivních galaktických jader. Vytvořeno podle ROBSON (1996)*



Obrázek 3.3: *Příklad spekter hvězdy a BL Lac objektu*

Problémy však nastávají v případech, kdy spektrum nepochází pouze od BL Lac objektu, ale i okolních objektů. V tomto případě by pomohlo pozorování v rádiové oblasti. Další problém nastal v klasifikaci spekter v SDSS. Jako kontrolní objekty byly použity objekty typu hvězda a galaxie, avšak některé z objektů, které jsou klasifikovány v katalogu VERONCAT jako BL Lac jsou v SDSS označeny jako galaxie.

3.2 BL Lac objekty a power-law

Záření BL Lac objektů je synchrotronové povahy (magnetické brzdné záření), produkované relativistickými elektrony kroužícími v magnetickém poli. Výsledné spektrum lze v úzké (optické) oblasti popsat mocninným zákonem[3] (power-law)

$$I(\lambda) = I_0 \, \lambda^{-\alpha}. \tag{3.1}$$

Logaritmováním lze rovnici (3.1) převést na tvar:

$$\log I = \log I_0 - \alpha \log \lambda. \tag{3.2}$$

V malém okolí kolem referenční vlnové délky λ_0 lze pomocí Taylorova rozvoje $\log \lambda$ se zanedbatelnou chybou aproximovat logaritmus rovnicí (3.3)

$$\log(\lambda) = \log \lambda_0 + \frac{\ln 10}{\lambda_0}(\lambda - \lambda_0) + \ldots, \tag{3.3}$$

$$\log I \approx \log I_0 - \alpha \log \lambda_0 + \alpha \ln 10 - \alpha \frac{\ln 10}{\lambda_0} \lambda. \tag{3.4}$$

První tři členy rovnice (3.4) udávají absolutní člen přímky, koeficient před čtvrtým její směrnici. Je vidět, že logaritmus spektra lze aproximovat lineární funkcí a zjistit hodnotu α pro objekty typu BL Lac. Ostatní objekty s jiným průběhem spektra by měly vykazovat větší odlišnosti oproti kontrolní skupině. Tímto způsobem je možné klasifikovat a následně vyhledat objekty podezřelé z příslušnosti ke skupině objektů typu BL Lac.

3.3 Příprava dat

Pro dobývání znalostí bylo třeba dvou tříd dat: kontrolní data pro trénink klasifikátoru, se spolehlivými údaji o objektech typu BL Lac a data objektů, jenž byly následně klasifikovány. Výsledné soubory musely obsahovat následující data ve formátu *arff* používaný programem *Weka*, který byl k účelu dobývání znalostí použit.

```
1 @relation testovaciData #
2 @attribute objekt_id string # id objektu
3 @attribute a numeric  # parametr
4 @attribute b numeric  # parametr
5 @attribute rez numeric # parametr
6 @attribute typ {1,2}  # typ objektu 1=bllac, 2=ostatni
```

3. ROBSON (1996)

```
7
8  @data
9  587722983895335009 1.729 -7.112E-005 4.050 1
10 587724197205573657 2.089 -7.108E-005 4.660 1
11 587724233174024233 2.384 -8.893E-005 1.900 1
```

V deklarační části jsou definovány proměnné a jejich datové typy. Kontrolní soubor obsahuje proměnnou typ, na jejímž základě lze určit chybu klasifikátoru.

Bylo tedy třeba získat souřadnice testovacích objektů, ty ztotožnit s objekty v SDSS, získat jejich spektra, tato spektra proložit přímkou a výsledné údaje transformovat do souboru typu arff. Tato data byla použita k trénování klasifikátoru a tento model následně použit na klasifikaci nových objektů.

Cest, jak tato data získat, je více, ta mnou zvolená jistě není nejefektivnější. Spíše jsem volil postupy maximalizující pochopení problematiky, než ty minimalizující čas potřebný na práci. Záměrně jsem se vyhýbal webovým aplikacím a grafickým nadstavbám.

Kontrolní data z VERONCAT

Jako zdroj kontrolních dat byl zvolen katalog VERONCAT, obsahující 1122 objektů typu BL Lac, ovšem pouze 662 je jich skutečně potvrzených. Toto byla první restrikce kontrolních dat. Z dostupných formátů jsem zvolil formát VOTable, který jsem pomocí nástroje CasJobs importoval do databáze MyDB[4] v CAS SDSS.

Spojení s daty v CAS SDSS

Souřadnice rektascenze a deklinace z VERONCAT katalogu bylo třeba zkorelovat se souřadnicemi v SDDS CAS archívu. Použil funkci *dbo.fGetNearestObjIdEq(ra, dec, 1)*, která pro dané souřadnice vrátí identifikátor objektu. Celý SQL dotaz vypadal následovně:

```
1 SELECT ra,dec,dbo.fGetNearestObjIdEq(ra, dec, 1) AS id
2 FROM mydb.bllac
3 INTO mydb.bllac2
```

Kde *mydb.bllac2* byla tabulka s BL Lac objekty z VERONCAT katalogu, vytvořená importem katalogu ve formátu VOTable do uživatelského prostoru CAS archívu. Tímto způsobem se podařilo ztotožnit 325 objektů. Toto byla druhá restrikce na kontrolní data a zároveň jejich konečný počet.

Získání spekter

Spektra jsou v SDSS uložena v DAS. Funkce *dbo.fGetUrlFitsSpectrum(specObjID)* vrátí odkaz ke spektru specifikovaném identifikátorem *specObjID*. Následujícím dotazem jsem získal seznam URL adres s odkazy na soubory FITS pro BL Lac objekty z VERONCAT katalogu.

4. Tento koncept je popsán v kapitole Technologie

```
SELECT objID,dbo.fGetUrlFitsSpectrum(specObjID)
FROM SpecPhoto
WHERE ObjId in (SELECT id
            FROM mydb.bllac2
            WHERE id IS NOT NULL)
```

Výstup měl následující formát:

```
587722982823756025,http://das.sdss.org/spectro/1d_26/0298/1d/
   spSpec-51955-0298-279.fit
587722983895335009,http://das.sdss.org/spectro/1d_26/0294/1d/
   spSpec-51986-0294-629.fit
587722983912243248,http://das.sdss.org/spectro/1d_26/0342/1d/
   spSpec-51691-0342-423.fit
```

Stažení souborů je možné provést pomocí programu **wget**, nebo **rsync**. Z výstupu SQL dotazu jsem získal skript. Za jména souborů jsem zvolil identifikátor z SDSS, to umožnilo jednoznačnou identifikaci a jednoduchou kontrolu výsledků. Transformaci jsem provedl pomocí programu **awk** následujícím příkazem.

```
awk 'BEGIN { FS = "," } ;
{ print "wget " $2 " -O " $1".fits" }' SeznamSpektra >
   getSpektra.sh
```

Výstup měl následující formát.

```
wget "http://das.sdss.org/spectro/1d_26/0570/1d/spSpec
   -52266-0570-008.fit" -O 587726033844568450.fits
wget "http://das.sdss.org/spectro/1d_26/0481/1d/spSpec
   -51908-0481-361.fit" -O 587726033845026976.fits
wget "http://das.sdss.org/spectro/1d_26/0500/1d/spSpec
   -51994-0500-459.fit" -O 587726033845813746.fits
```

Extrakce dat z FITS souborů

Jednoduchým programem ve **Fortranu** jsem získal hodnoty vlnové délky a toku záření. Hodnoty vlnových délek nejsou ve FITS souborech uloženy přímo, ale bylo nutné je transformovat pomocí rovnice (3.5).[5] Data bylo třeba logaritmovat, abychom mohli použít lineární funkci. Z důvodu úspory místa a zvýšení srozumitelnosti uvádím jen klíčovou část programu pro výpis dat z FITS souboru. Podrobnější pojednání je v oddíle o technologiích. Výsledný program přikládám k práci.

5. Viz `http://www.sdss.org/DR5/products/spectra/read_spSpec.html`

$$\lambda(i) = 10^{(a+bi)}, \tag{3.5}$$

kde i je číslo řádku, $a = 3{,}5796$ a $b = 10^{-4}$ jsou konstanty získané z hlavičky souborů.

```
do i = 1,naxes(1)

   lambda = 10.0**(3.5796 + i*10.0**(-4))
   write(*,*) lambda,log10(d(i,j))

end do
```

Proložení spektra přímkou

K tomuto účelu jsem modifikoval **fortranovský** program Filipa Hrocha s názvem rline, který za pomoci robustní metody největší věrohodnosti[6] proloží daný graf přímkou a vypíše reziduální součet. Program používá knihovny z knihy Numerical Recipes (TEUKOLSKY et al. (1990)).

Spektra jsem proložil přímkou a získal její parametry (sklon, absolutní člen) a reziduální součet. Reziduální součet je úměrný odchylkám hodnot od této přímky. Méně spektrálních čar tedy znamená menší hodnotu reziduálního součtu. Naopak spektra odlišná od power-law mají větší hodnotu reziduálního součtu.

Příprava trénovacích dat pro klasifikátor

Všechny předchozí kroky byly automatizovány pomocí bash skriptu tak, aby při opakovaném použití nebylo nutné provádět jednotlivé kroky ručně.

```
#!/bin/bash
log=rez
sep=" "
rm $log

for soubor in *.fits
do
vystup= `fitslist $soubor | rline | awk '{OFS=" "}{print $1,$2,
   $4 }'`
   vystup=$soubor$sep$vystup
   if [ -n "$1" ]
     then
         vystup=$vystup$sep$1
   fi

   echo $vystup >> $log
echo $$soubor " done"
```

6. Maximum Likehood (MLE)

Výsledný soubor obsahoval parametry přímky pro BL Lac objekty z VERONCAT katalogu a další objekty z SDSS se známým typem spektra (objekty klasifikované jako galaxie a hvězdy). Na základě těchto dat vytvořil klasifikátor rozhodovací strom.

Příprava dat pro klasifikaci nových objektů

Protože cílem této části práce bylo nalezení BL Lac objektů, bylo třeba vybrat vhodné objekty pro klasifikaci. Tyto objekty jsem vybíral podle následujících kritérií: neznámý typ spektra, tj. hodnota UNKNOWN pro specClass,[7] zároveň nesměl být součástí VERONCAT katalogu. Data splňující tyto požadavky dostaneme následujícím SQL dotazem:

```
SELECT TOP 1000 objID,dbo.fGetUrlFitsSpectrum(specObjID)
FROM SpecPhoto
WHERE ObjId NOT IN (SELECT id
            FROM mydb.BL~Lac objekt2
            WHERE id IS NOT NULL)
AND specClass = dbo.fSpecClass('UNKNOWN')
```

Analyzovaná data byla zpracována stejným způsobem jako trénovací data; tj. byla stažena spektra ve formátu FITS, z nich extrahována data a následně proložena přímkou. Výsledky byly zapsány do souboru a převedeny do formátu *arff*.

3.4 Dobývání znalostí

Vlastní dobývání bylo provedeno programem Weka, obsahující velké množství algoritmů pro datovou analýzu, strojové učení a přípravu dat. V praktické části byl použit klasifikátor J48, který je implementací volně šířeného algoritmu C4.5, jehož autorem je Ross Quinlan. Základní principy algoritmu jsou popsány v kapitole o dobývání znalostí.

Trénink klasifikátoru

Výsledkem tréninku klasifikátoru byl následující výstup.

```
=== Run information ===
Scheme:    weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:  rezidual6-weka.filters.unsupervised.attribute.Remove-
   R1
Instances: 598
Attributes: 4
           a
           b
           rez
           typ
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===
```

7. Sloupec tabulky SpecPhoto v CAS SDSS

```
13 J48 pruned tree
14 ------------------
15 rez <= 4.985968: 1 (56.0/5.0)
16 rez > 4.985968
17 |  a <= 1.725255
18 |  |  rez <= 15.266346
19 |  |  |  a <= 1.542392: 1 (35.0)
20 |  |  |  a > 1.542392
21 |  |  |  |  b <= -0.000021: 2 (8.0/2.0)
22 |  |  |  |  b > -0.000021: 1 (3.0)
23 |  |  rez > 15.266346
24 |  |  |  rez <= 34.054842
25 |  |  |  |  a <= 1.045598: 1 (10.0/1.0)
26 |  |  |  |  a > 1.045598: 2 (45.0/8.0)
27 |  |  |  rez > 34.054842: 2 (123.0/5.0)
28 |  a > 1.725255: 2 (318.0/9.0)
29
30 Number of Leaves : 8
31 Size of the tree : 15
32 Time taken to build model: 0.06 seconds
33
34 === Stratified cross-validation ===
35 === Summary ===
36 Correctly Classified Instances 550         91.9732 %
37 Incorrectly Classified Instances 48        8.0268 %
38 Kappa statistic                  0.7467
39 Mean absolute error              0.1174
40 Root mean squared error          0.2719
41 Relative absolute error         36.0758 %
42 Root relative squared error     67.4776 %
43 Total Number of Instances    598
44
45 === Detailed Accuracy By Class ===
46               TP Rate FP Rate Precision Recall F-Measure ROC Area
                    Class
47                0.77    0.042    0.825   0.77    0.797    0.869  1
48                0.958   0.23     0.942   0.958   0.95     0.869  2
49 Weighted Avg. 0.92   0.191    0.918   0.92    0.919    0.869
50
51 === Confusion Matrix ===
52    a   b   <-- classified as
53   94  28 | a = 1
54   20 456 | b = 2
```

První řádek ukazuje úspěšnost klasifikátoru (Correctly Classified Instances), která v tomto případě dosahuje 92 %. Z toho vyplývá, že pouze 48 z 550 objektů určil klasifikátor chybně. Tento výsledek dává solidní základ na použití pro neznámé objekty.

Na řádcích 18 až 31 je vytvořený rozhodovací strom. Jednotlivé parametry (sklon přímky, absolutní člen a reziduální součet) jsou rozděleny na intervaly a použity k zařazení

do jedné ze dvou skupin. První by měla reprezentovat objekty typu BL Lac a druhá ty ostatní (galaxie a hvězdy).

Klasifikace neznámých objektů

Program Weka lze spouštět z příkazové řádky a automatizovat pomocí skriptů. Následujícím příkazem jsem nastavil klasifikátor na J48 (řádek 3), specifikoval trénovací a testovací soubor (řádek 4) a nastavením parametru *-p 1* jsem uložil identifikátor objektu z SDSS do výstupního souboru.

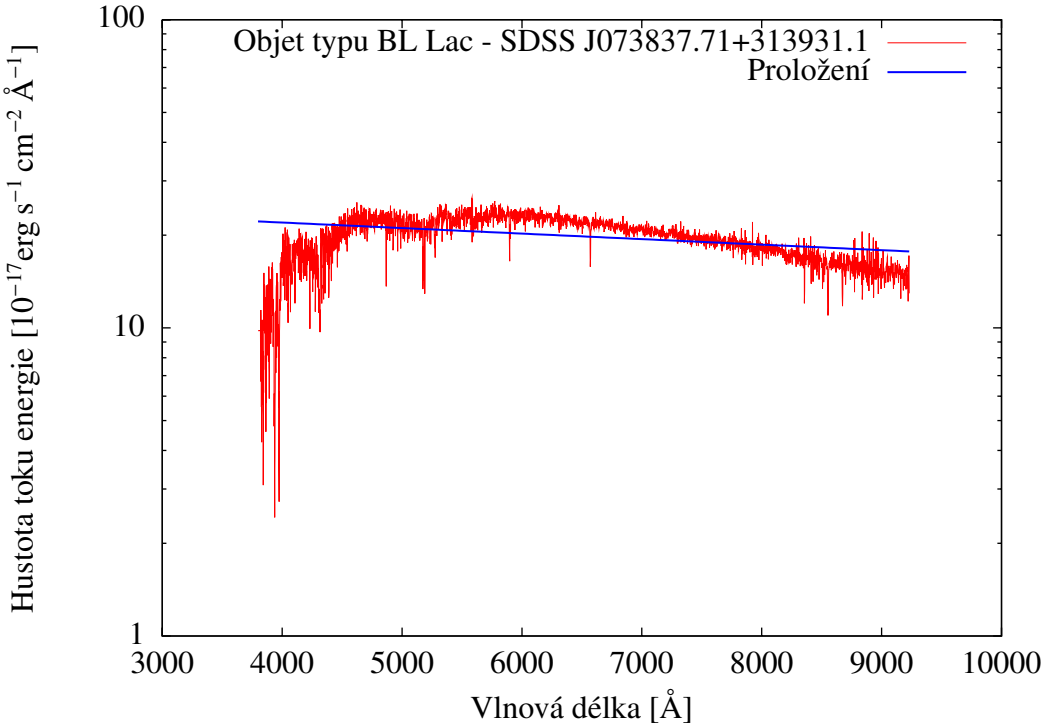
```
1 java weka.classifiers.meta.FilteredClassifier
2 -F weka.filters.unsupervised.attribute.RemoveType
3 -W weka.classifiers.trees.J48
4 -t train.arff -T test.arff -p 1
```

3.5 Interpretace

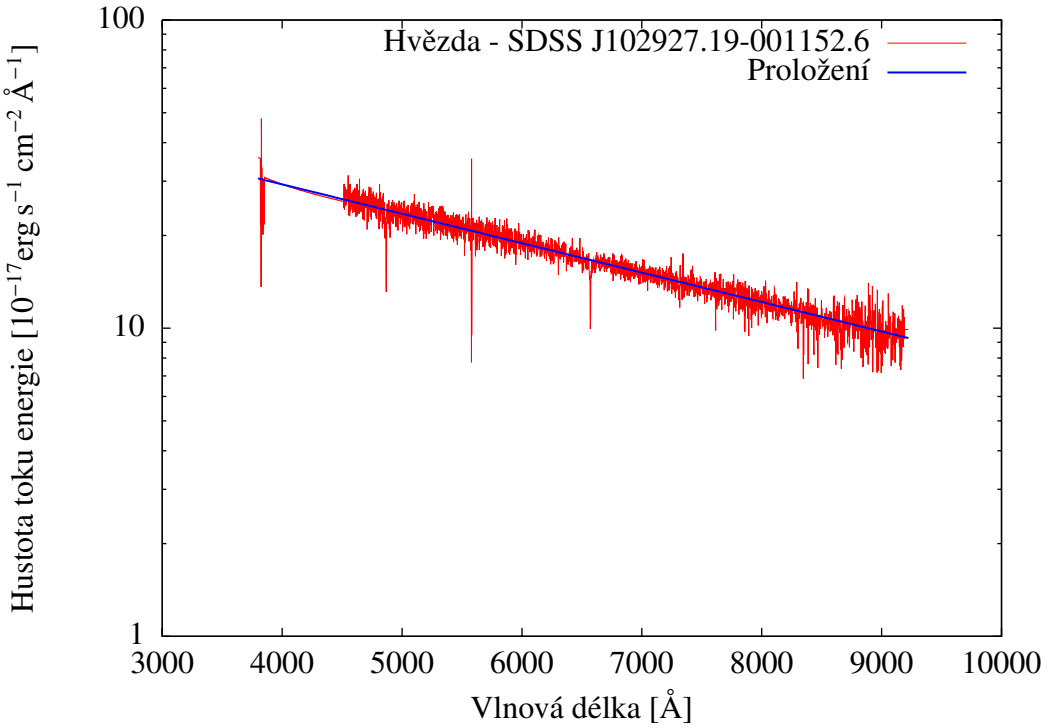
Protože se doposud jednalo o kontrolní data se známým typem objektu, bylo možné provést analýzu chyb klasifikátoru.

Analýza chybných zařazení

SDSS J073837.71+313931.1 je veden v katalogu VERONCAT jako objekt typu BL Lac, ale klasifikátor ho zařadil mezi hvězdy. Hodnoty proložené přímky jsou 1,41 (absolutní člen), $-1,80$ (sklon přímky) a 52,24 (reziduální součet). Z výpisu rozhodovacího stromu plyne, že pokud je reziduální součet vyšší něž 4,92, pak se rozhoduje podle hodnoty absolutního členu. Ta je v tomto případě 1,41, a tedy menší než hodnota 1,72. Opět se tedy rozhoduje podle hodnoty reziduálního součtu. Protože je ostře vyšší než 34,05, je objekt zařazen do skupiny dva (tj. ostatní objekty).
SDSS J102927.19-001152.6 je v SDSS zařazen mezi hvězdy, avšak klasifikátor jej zařadil do skupiny 1, tedy objekty typu BL Lac. Hodnoty proložené přímky jsou 1,84 (absolutní člen), $-9,55$ (sklon přímky) a 4,19 (reziduální součet). Zde je důvod přiřazení ještě přímější než v předchozím případě. Protože je reziduální součet menší než 4,98, byl objekt klasifikován jako objekt typu BL Lac.



Obrázek 3.4: *Příklad chybně zařazeného objektu. Klasifikátor určil, že se o BL Lac objekt nejedná, avšak ve VERONCAT katalogu je identifikován jako BL Lac objekt*



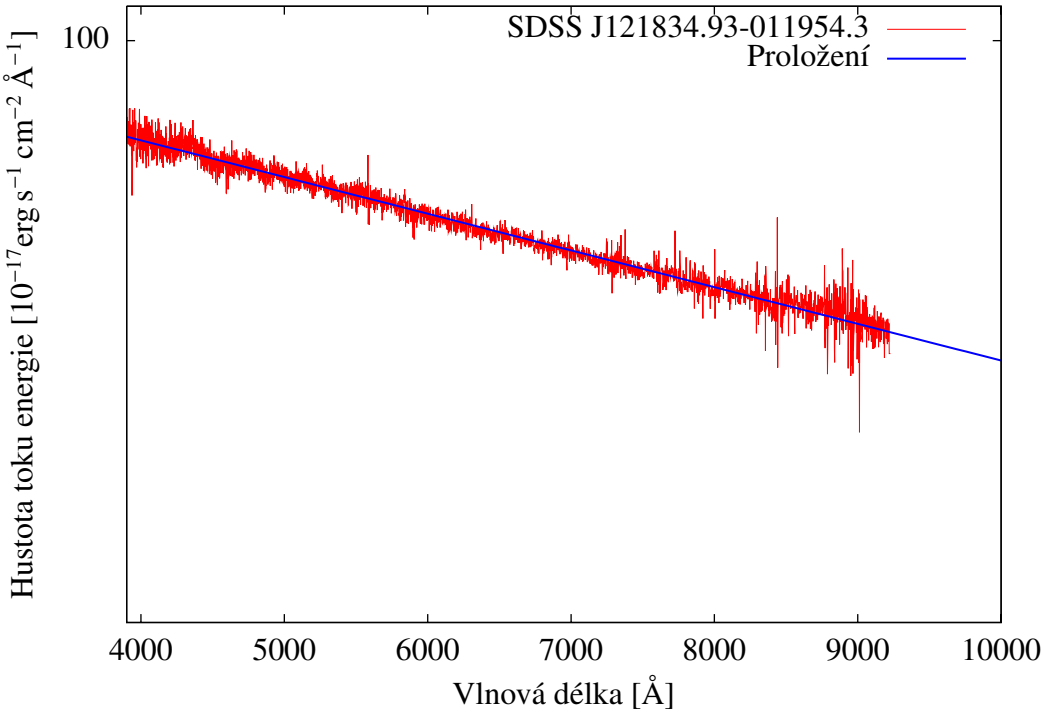
Obrázek 3.5: *Příklad chybně zařazeného objektu. Klasifikátor určil, že jde o BL Lac objekt, avšak v SDSS je identifikován jako hvězda*

Příklady nalezených objektů

Poté, co byly náhodně vybrané objekty klasifikovány, byl proveden pomocí databáze vědeckých ADS[8] článků a astronomické databáze SIMBAD[9] průzkum jejich skutečné povahy.

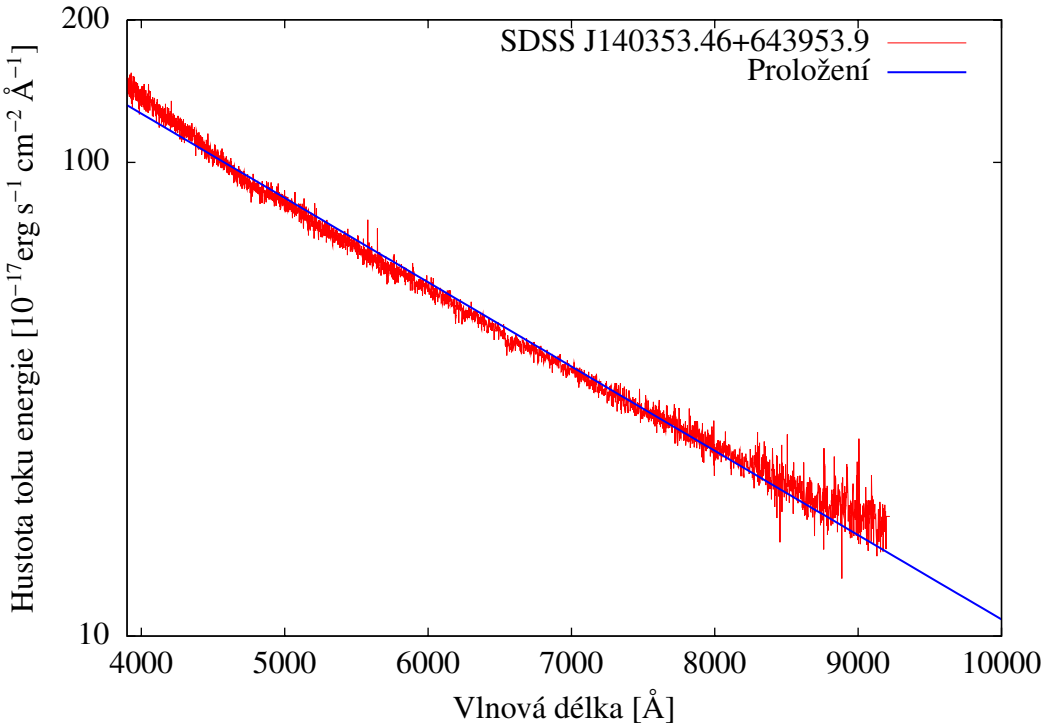
Jméno objektu	α[°]	δ[°]	Typ objektu	Odkaz na zdroj
SDSS J121834.93-011954.3	184,65	-1,33	BL Lac	LONDISH et al. (2007)
SDSS J140353.46+643953.9	210,97	64,66	bílý trpaslík	EISENSTEIN et al. (2006)
SDSS J102653.03+644459.0	156,72	64,75	kvasar	SCHNEIDER et al. (2007)
SDSS J105829.60+013358.7	164,62	1,57	kvasar	LISTER et al. (2009)
SDSS J015441.74+140308.0	28,67	14,05	bílý trpaslík	EISENSTEIN et al. (2006)
SDSS J142923.91+024023.1	217,35	2,67	bílý trpaslík	EISENSTEIN et al. (2006)

Tabulka 3.1: *Klasifikované objekty*

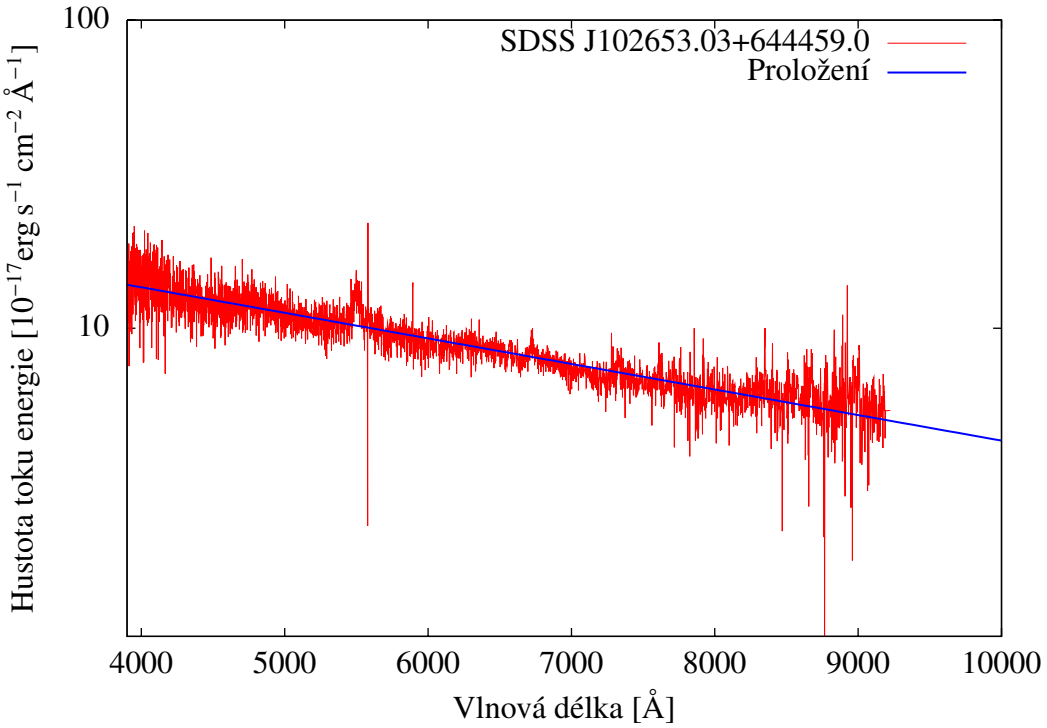


Obrázek 3.6: *Příklad nalezeného objektu: SDSS J121834.93-011954.3*

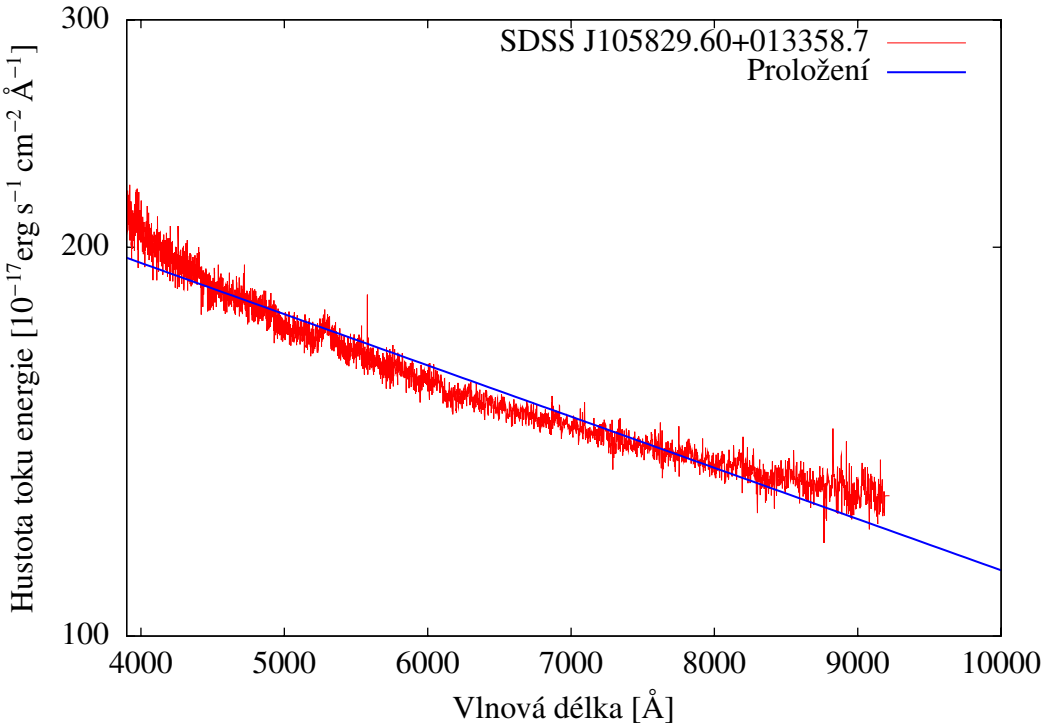
8. Digital Library for Physics and Astronomy
9. SIMBAD4 1.120 - 27-Apr-2009



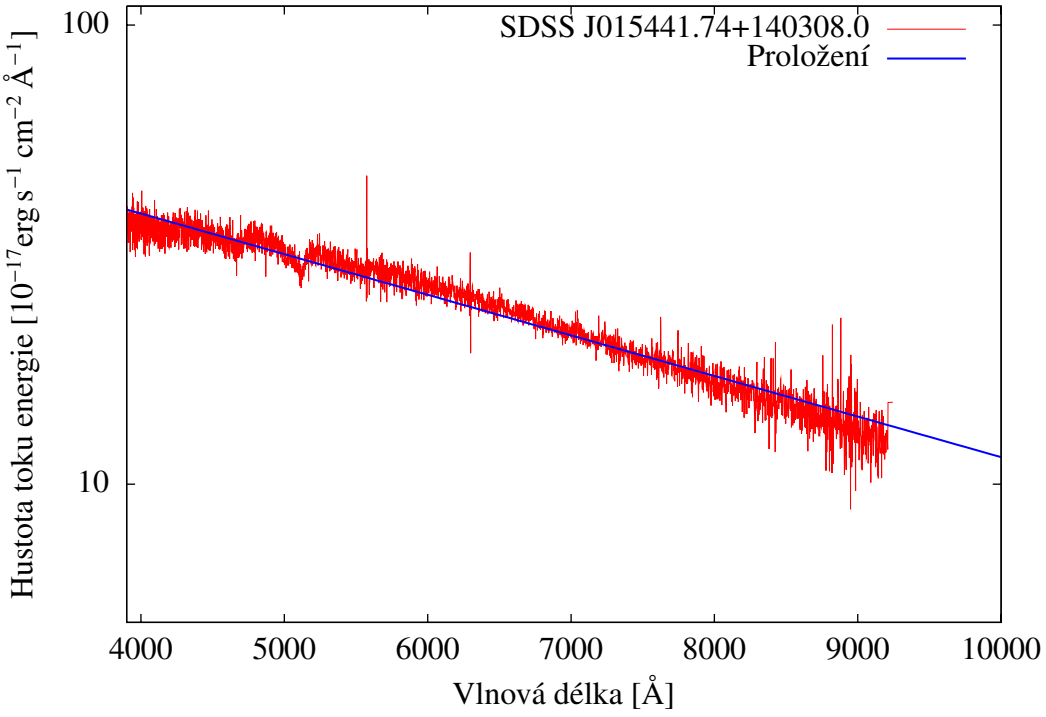
Obrázek 3.7: *Příklad nalezeného objektu: SDSS J140353.46+643953.9*



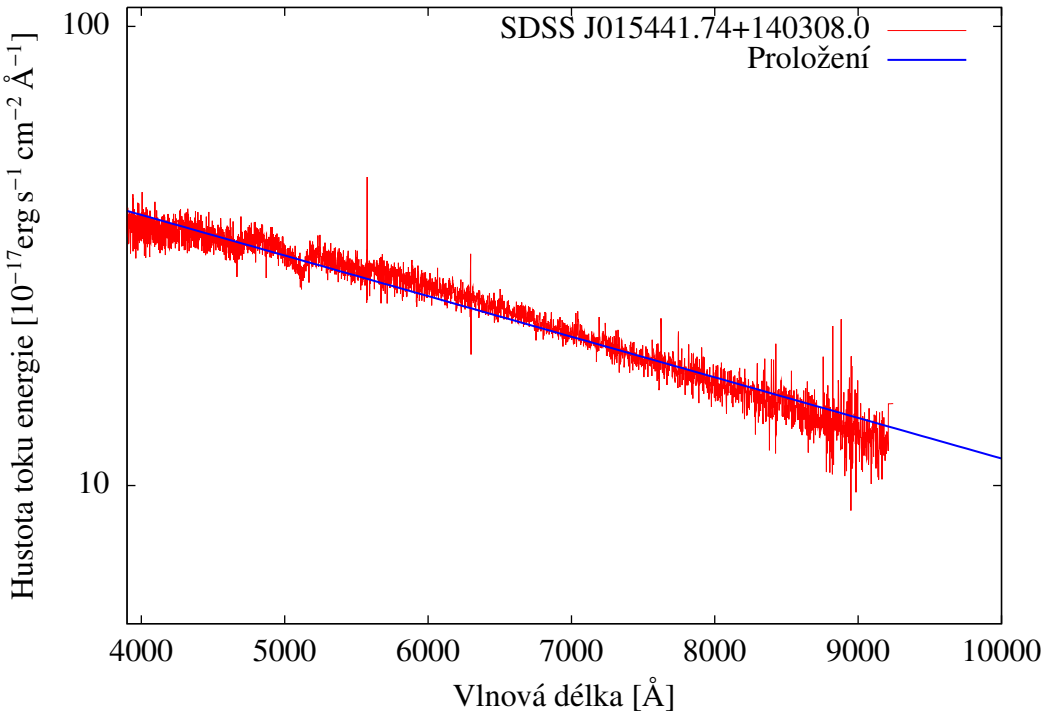
Obrázek 3.8: *Příklad nalezeného objektu: SDSS J102653.03+644459.0*



Obrázek 3.9: *Příklad nalezeného objektu: SDSS J105829.60+013358.7*



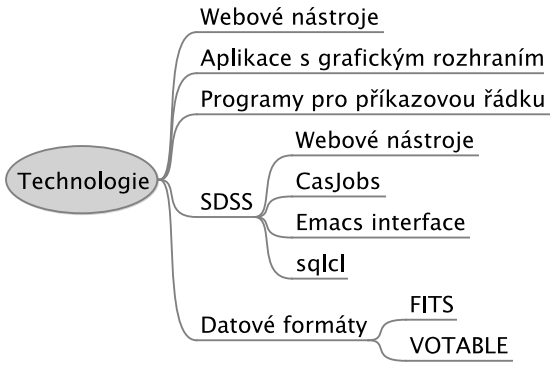
Obrázek 3.10: *Příklad nalezeného objektu: SDSS J015441.74+140308.0*



Obrázek 3.11: *Příklad nalezeného objektu: SDSS J142923.91+024023.1 - Kačenka*

Kapitola 4

Technologie



Obrázek 4.1: *Struktura kapitoly: Technologie*

Z hlediska uživatelského rozhraní lze softwarové nástroje rozdělit na webové aplikace, lokální klienty s grafickým rozhraním a programy pro příkazovou řádku. Každá z těchto skupin je vhodná pro určitý druh práce, což musí mít na paměti nejen jejich tvůrci, ale zejména uživatelé. Použití nevhodného typu nástroje může mít mnoho negativních dopadů.

4.1 Webové aplikace

Silnou stránkou webových aplikací je jejich téměř absolutní přenositelnost, jsou nezávislé na platformě a téměř všude dostupné. Problémem však zůstává rychlost odezvy.

4.2 Aplikace s grafickým rozhraním

Jsou obecně nutně závislé na platformě a musí být tedy psány s ohledem na operační systém. Tento problém se snaží vyřešit programovací jazyky využívající virtuální počítač.[1] Idea je taková, že pro každou platformu se jednou vyvine abstraktní vrstva (Virtual Machine), která odstíní specifika daného operačního systému a aplikace lze vyvíjet pro tento

1. Tato technologie řeší i další problémy, jako je správa paměti (garbage collector) atp.

virtuální počítač. Nespornou výhodou oproti webovým aplikacím je rychlost odezvy. Možnost spolupráce mezi programy je silně omezená a musí být explicitně implementována. Pěkný příklad takové spolupráce ukazuje technologie PLASTIC[2] používaná v programech Virtuální observatoře.

4.3 Programy pro příkazovou řádku

Přinášejí množství výhod oproti dříve jmenovaným kategoriím. Důležitá je schopnost spolupráce s dalšími programy, která v důsledku vede k možnostem, které autor programu nemusel explicitně implementovat. Další výhodou (sic!) je nutnost znalosti zkoumané problematiky před vlastní prací s programem. Snad všechny vědecké programy mají možnost používat program tímto efektivním způsobem.

4.4 Nástroje SDSS

SDSS je velice pečlivě vedený projekt a to se týká i možnosti přístupu k datům. Webový server obsahuje velké množství tutoriálů, článků a prezentací. Po zaregistrování má uživatel k dispozici vlastní datový prostor, kam může importovat svá data, vytvářet tabulky s pojovat tato data s daty z CAS databáze.

MyDB

Velice zajímavý a užitečný je koncept vlastní části[3] databáze uvnitř datového prostoru CAS nazvaný MyDB. Po zaregistrování je uživateli přiděleno 0,5GiB prostoru, kde může vytvářet své vlastní datové struktury, importovat do nich data a následně je spojovat se zbytkem CAS. Tato možnost byla použita v praktické části práce. K MyDB lze přistupovat pomocí webového rozhraní nebo prostřednictvím programu CasJobs.

Webové nástroje

SDSS obsahuje velmi širokou paletu aplikací všech výše zmíněných typů. Všechny fungují na podobném principu: uživatelův dotaz na objekt, oblast nebo skupinu objektů se převede na SQL dotaz, ten zpracuje databázový stroj relační databáze (CAS), pokud je to žádoucí, zkombinuje se výstup s daty ze souborové části (DAS) a výsledek je prezentován uživateli. Velmi praktické je i zobrazení prováděného SQL dotazu, který lze zkopírovat, modifikovat a použít v jiném nástroji.

CasJobs

Je na Javě postavený klient pro příkazovou řádku, sloužící k zadávání úloh, práci s datovými strukturami MyDB databáze a extrakci dat do různých formátů. Následuje ukázka použití

2. Platform for Astronomy Tool InterConnection – protokol pro předávání dat mezi klientskými programy.
3. V databázovém světě označovaném jako schéma.

```
casjobs execute 'select ra,dec,dbo.fGetNearestObjIdEq(ra,dec,
   1) from MyDB.bllac'
```

Pokud je dotaz rozsáhlejší lze ho zařadit do fronty.

```
casjobs submit -n "jmeno" dotaz.sql

Submitting new query...

Target/Queue:DR7/1
Taskname:test
Query is file?:false
Query:dotaz.sql

Query succesfully submitted!
JobID is 3448925
```

Po dokončení zpracování je vygenerován odkaz a data je možné stáhnout.

```
casjobs output
  Table: MyTable
  Type: CSV
  URL: http://casjobs.sdss.org/CasJobsOutput2/CSV/MyTable_astar.
     csv
  Completed: 2/7/2009 7:49
  Table: test0
```

Emacs interface

Emacs je rozšiřitelný textový editor. Jedno z těchto rozšíření umožňuje velice pohodlnou[4] práci s CAS databází SDSS. Toto rozšíření je k dispozici na stránkách projektu SDSS. Stejně jako samotný textový editor je napsáno v jazyce Lisp. Instaluje se obdobně jako ostatní rozšíření. Stačí v konfiguračním souboru *.emacs* v domovské adresáři uvést *(load " /cesta/skyserver.el")*. Funkce jsou pak dostupné standardně přes *M-x skyserver-příkaz*. Například *skyserver-submit-region* pošle k zpracovaní databází označený blok SQL příkazů. Práce s tímto rozšířením je velice efektivní, neboť máme k dispozici veškerý komfort editoru Emacs.

4.5 Weka

Weka (Waikato Environment for Knowledge Analysis) je soubor algoritmů pro dobývání znalostí a strojové učení. Je napsán v jazyce Java a pochází z Waikatské univerzity na

4. Za předpokladu, že Vám práce v Emacsu připadá pohodlná.

Novém Zélandu. Šířen je pod licencí GNU GPL. Umožňuje předzpracování dat, shlukovou analýzu a klasifikační algoritmy. S programem lze pracovat jak pomocí grafické rozhraní tak přes příkazovou řádku. Program je dostupný na adrese `http://www.cs.waikato.ac.nz/ml/weka/`. Existuje i rozšíření AstroWeka umožnující přímou práci s daty VO a formátem VOTable.

4.6 Datové formáty

FITS

Flexible Image Transport System je datový formát používaný pro ukládání, výměnu a práci s vědeckými daty převáženě v astronomii. Může obsahovat jak obrazová data, tak ASCII a binární tabulky. V jednom souboru může být uloženo několik skupin dat. Začátek souboru je tvořen ASCII hlavičkou, kde jsou uloženy informace o datové části souboru. Pro každou skupinu dat slouží jedna hlavička. Z důvodu jeho důležitosti uvádím přehled klíčových momentů ve vývoji tohoto formátu.

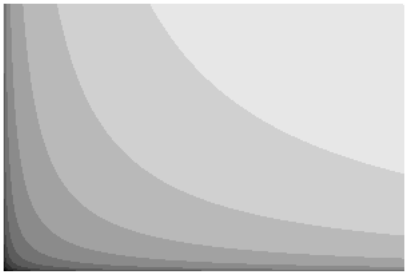
Datum	Milník
1979	Předběžný návrh FITS a první výměna souborů
1981	Publikována původní (jednoduché HDU) definice hlavičky
1982	Formálně schváleno IAU
1988	Definovány pravidla pro soubory s více skupinami dat
1988	Ustanovena pracovní skupina uvnitř IAU zabývající se FITS formátem
1988	Rozšíření o ASCII tabulky
1988	Formální schválení IAU ASCII tabulek
1990	Rozšíření zahrnující data s plovoucí čárkou
1994	Rozšíření o více obrazových polí
1995	Rozšíření o binární tabulky
1997	Adoptován čtyřmístný formát roku
2002	Adoptovány konvence pro světový souřadnicový systém
2004	Adoptovány MIME typy
2005	Rozšíření o pole proměnné délky v binárních tabulkách
2005	Adoptovány konvence pro spektrální souřadnicový systém
2005	Rozšíření o 64bitové datové typy

Tabulka 4.1: *Historické milníky ve vývoji FITS fromátu. Převzato z HANISCH et al. (2001)*

Tento formát je podporován v programech[5] na zpracování astronomických dat. Pro porozumění a flexibilní použití je však více než vhodné umět s tímto formátem pracovat prostřednictvím knihovny CFITSIO a některého z podporovaných[6] jazyků. Následuje ukázka vytvoření obrázku ve FITS formátu. Velmi podobně lze vytvořit a číst ASCII a binární tabulky. Vše je důkladně popsané na domovských stránkách knihovny FITSIO `http://heasarc.nasa.gov/docs/software/fitsio/fitsio.html`.

5. Např. fv, SAOImage ds9, FITS Utils, FTOOLS
6. Fortran, C++, Python, Ruby, Tcl

```
program image
  integer status,unit,x,y,obrazek(300,200),naxes(2)

  status=0

  call ftgiou(unit, status )
  call ftinit(unit,'obrazek.fit',1,status)

  naxes(1)=300
  naxes(2)=200

  call ftphpr(unit,1,16,2,naxes,0,1,1,status)

  obrazek=log(real(x*y))

  call ftpprj(unit,1,1,naxes(1)*naxes(2),obrazek,status)

  call ftclos(unit, status)
  call ftfiou(unit, status)
end
```



Obrázek 4.2: *Obrázek ve formátu FITS vytvořený programem image*

VOTable

VOTable je datový formát vyvíjený pro potřeby Virtuálních observatoří. Odvozen je ze standardu Astrores s použitím XML formátu. Datová část může být uložena jako TABLEDATA (XML), FITS nebo BINARY. Flexibilní návrh umožnuje mnoho způsobů použití,

jako například paralelní a distribuované zpracování (grid computing), kontinuální zpracování, ukládání rozsáhlých struktur a multidimezionálních polí (OCHSENBEIN et al. (2004)). Následuje ukázka VOTable, v které jsou uložena data katalogu VERONCAT, který jsem využíval v praktické části této práce.

```
<?xml version="1.0"?>
<!DOCTYPE VOTABLE SYSTEM "http://us-vo.org/xml/VOTable.dtd">
<VOTABLE version="1.0">
  <DEFINITIONS>
    <COOSYS system="eq_FK5" equinox="2000" />
  </DEFINITIONS>
  <RESOURCE ID="T9001">
     <DESCRIPTION>
       HEASARC Browse data service
       Please send inquiries to mailto:request@athena.gsfc.nasa.
          gov
    </DESCRIPTION>
    <PARAM ID="default_search_radius" ucd="OBS_ANG-SIZE" value="
       0.0333333333333333" ></PARAM>
    <TABLE>
      <DESCRIPTION> Veron Catalog of Quasars &amp; AGN, 12th
         Edition </DESCRIPTION>
      <FIELD name="unique_id" datatype="int" >
        <DESCRIPTION> Integer key </DESCRIPTION>
      </FIELD>
.
.
.
      <DATA>
        <TABLEDATA>
<TR><TD>85251</TD><TD>Q J00278-2935</TD><TD>6.9650</TD><TD>
   -29.5853</TD><TD>BL</TD></TR>
```

Poznámka k použitým nástrojům: Celou práci jsem vytvořil pomocí otevřeného softwaru. Práci jsem napsal v editoru GNU Emacs s rozšířením AucTEXna operačním systému GNU/Linux a vysázel sázecím systémem LATEX. Myšlenkové mapy na začátku kapitol jsou vytvořeny programem Freemind. Zdrojové kódy a vlastní text práce jsem verzoval pomocí systému Subversion, což umožnovalo, mimo jiné, efektivní spolupráci s vedoucím práce.

Závěr

Metody dobývání znalostí se snaží vypořádat s problémem množství dat, která moderní společnost vytváří. Věda a speciálně astrofyzika je v této produkci na předním místě. Astrofyzikální data jsou navíc velmi dobře přístupná. Tato kombinace představuje ideální prostředí pro nové objevy a všeobečně prospěšný rozvoj této a přilehlých oblastí informatiky, astrofyziky a dalších věd.

V této práci jsem se snažil na praktickém příkladu hledání objektů typu BL Lac ve Sloanově digitální přehlídce oblohy ukázat celý postup od získání dat, přes jejich zpracování, vlastní dobývání znalostí až k interpretaci nalezených objektů. Použitá metoda se ukázala jako úspěšná. Jeden ze šesti představených objektů skutečně takovým objektem je a zbylé objekty jsou z hlediska vzhledu spektra těmto objektům blízké (i když jejich fyzikální vlastnosti jsou velmi odlišné). Pokud se metoda rozšíří o další parametry specifické pro objekty tohoto typu (např. rádiové vyzařování) mohla by metoda sloužit k nalezení zcela nových objektů.

Obecně řečeno se velice nadějná jeví možnost spojení jednotlivých přehlídek oblohy a dalších datových zdrojů v různých vlnových délkách, za použití metod dobývání znalostí, což pravděpodobně povede k nalezení zcela nových objektů a skupin. Také můžeme rozpoznat dříve netušené souvislosti mezi již známými objekty vesmíru. V současné době se velké úsilí věnuje projektům Virtuálních observatoří, které takovéto výzkumy umožní (viz např. DJORGOVSKI et al. (2001)).

Klíčové bylo v této práci použití informačních technologiií. V každém kroku jsem se snažil nespoléhat na hotový software, ale pokud možno výtvářet vlastní programy, případně upravovat již existující. Takový přístup je sice velice časově náročný, avšak vede k hlubšímu porozumění zkoumané problematiky.

Na závěr bych rád připojil varování. Spoléhání na informační technologie se může stát velice zrádné. Je třeba si uvědomit, že se jedné pouze o velice užitečný nástroj, ale nikoliv o řešení problémů. Mnohokrát řešíme spíše komplikace spojené s těmito technologiemi, než skutečný vědecký problém. V důsledku jejich používání také ztrácíme některé schopnosti, které byly pro generace vědců bez přístupu k těmto prostředkům klíčové.

Literatura

P. BERKA. *Dobývání znalostí z databází*. Academia, 2003.

SG DJORGOVSKI, RJ BRUNNER, AA MAHABAL, SC ODEWAHN, et al. Exploration of large digital sky surveys. *Mining the Sky*, 2001.

Z. DOŠLÁ, O. DOŠLÝ. Metrické prostory: teorie a příklady. 3. vyd. Brno: Masarykova univerzita, 2006. Technical report, ISBN 80-210-4160-9, 2006.

D. J. EISENSTEIN, J. LIEBERT, H. C. HARRIS, S. J. KLEINMAN, A. NITTA, N. SILVESTRI, S. A. ANDERSON, et al. A Catalog of Spectroscopically Confirmed White Dwarfs from the Sloan Digital Sky Survey Data Release 4. *apjs*, 167:40–58, November 2006. doi: 10.1086/507110.

U. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMYTH, et al. From data mining to knowledge discovery in databases. *Communications of the ACM*, 39(11):24–26, 1996.

RJ HANISCH, A. FARRIS, EW GREISEN, WD PENCE, et al. Definition of the flexible image transport system (FITS). *Astronomy and Astrophysics*, 376(1):359–380, 2001.

IEC60027-2. Letter symbols to be used in electrical technology-part 2: Telecommunications and electronics. *IEC standard*, pages 60027–2, 2000.

M. L. LISTER, H. D. ALLER, M. F. ALLER, M. H. COHEN, et al. MOJAVE: Monitoring of Jets in Active Galactic Nuclei with VLBA Experiments. V. Multi-Epoch VLBA Images. *APJ*, 137:3718–3729, March 2009. doi: 10.1088/0004-6256/137/3/3718.

D. LONDISH, S. M. CROOM, J. HEIDT, B. J. BOYLE, E. M. Sadler, M. Whiting, T. A. Rector, T. Pursimo, and K. Chynoweth. The 2QZ BL Lac survey - II. *mnras*, 374: 556–578, January 2007. doi: 10.1111/j.1365-2966.2006.11165.x.

F. OCHSENBEIN, R. WILLIAMS, C. DAVENHALL, D. DURAND, et al. VOTable Format Definition Version 1.1. *IVOA VOTable WG Recommendation*, 2004.

JR QUINLAN. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Jordan RADDICK. Color-Magnitude Diagram of Galaxies. *Cooking with Sloan: SDSS Data Demos*, 2006.

I. ROBSON. Active galactic nuclei(Book). *Chichester, United Kingdom: John Wiley & Sons, 1996.*, 1996.

D. P. SCHNEIDER, P. B. HALL, G. T. RICHARDS, M. A. STRAUSS, D. E. VANDEN BERK, S. F. ANDERSON, et al. The Sloan Digital Sky Survey Quasar Catalog. IV. Fifth Data Release. *APJ*, 134:102–117, July 2007. doi: 10.1086/518474.

S.L.A. TEUKOLSKY, B.N.P. FLANNERY, and W.M.T. VETTERLING. *Numerical Recipes: FORTRAN*. Cambridge University Press New York, NY, USA, 1990.

M.-P. VÉRON-CETTY and P. VÉRON. A catalogue of quasars and active nuclei: 12th edition. *AAP*, 455:773–777, August 2006. doi: 10.1051/0004-6361:20065177.