Analýza klimatologických dat
Závěrečná zpráva projektu PA164

Pavel Drášil[1], Tomáš Gregar[1]

Fakulta informatiky Masarykovy univerzity v Brně
Botanická 68a, 602 00 Brno
{xdrasil,xgregar}@fi.muni.cz

Klíčová slova: strojové učení, SVM, naive Bayes, decision trees

1 Zadání

Zadáním projektu bylo vyzkoušet na dodaných datech (získaných z Geografického ústavu Přírodovědecké fakulty MU) postupy extrakce informace z textu pomocí strojového učení. Hlavním cílem práce bylo zjistit, zda jsou metody strojového učení použitelné pro klasifikaci takovýchto záznamů.

2 Data

Objektem našeho zkoumání jsou záznamy historické klimatologie z území České republiky od roku 1500. Většinou se jedná o několikavěté popisy z kronik a jiných pramenů, popisující klimatologicky zajímavé události.
Geografové taková data dělí na

- přímá, zabývající se přímo klimatem
- nepřímá („proxy“), popisující intenzitu nebo povahu jevů a další data, která s počasím souvisí (úrodu, ceny potravin apod.)

Obecně mohou být klimatologická data rozdělena na:

- **dokumentární data**
  - písemné a grafické prameny
    - letopisy, kroniky, paměti
    - deníky, kalendáře (denní záznamy, Jan z Kunovic 1533–1545)
    - ekonomické záznamy (kniha počtů města Loun)
    - obrazy (Brueghel), letáky
  - meteorologické záznamy (až od 17. století)
  - archeologické prameny (vykopávky v záplavových územích)
- **data vázaná na přírodní zdroje** (z ledovců, sedimentů...)

Data, která máme k dispozici, byla získána z písemných pramenů. Jedná se o databázi, která byla na daném ústavu nejprve organizována „analogově“ – na kartičkách (od 1990). V roce 2002 byly záznamy přepsány do počítačové podoby, umožňující lepší indexaci a klasifikaci podle času i tématu (*Dobrovolný,*

P. a Fukátko, J., 2003). Většina záznamů je z let 1500–1570, celkem databáze obsahuje 850 štítků.

Obsah záznamů je klasifikován do následujících kategorií (kategorie převzaty z: *Schülle, H. a Pfister, Ch., 1992, 235-262*):

- teplotní podmínky
- srážkové podmínky
- ukazatele sucha, vlhka, mrazu nebo tepla
- stav ovzduší (povětrnostní podmínky, optické jevy, atd.) a astronomické jevy
- vodní režim na stojatých a tekoucích vodách, ledové jevy
- charakteristiky týkající se moří (výška hladiny moře) nebo ledovců
- zemětřesení, sesuvy půdy, výbuchy sopky apod.
- charakteristiky týkající se úrody
- pozorování vegetace
- sezónní výskyt živočichů
- charakteristiky týkající se cen zboží, surovin, potravin
- zdraví obyvatelstva (epidemie, nemoci)
- dopady extrémních projevů počasí atd.

Tato kategorizace odpovídá praxi z mezinárodních projektů, do nichž je ČR zapojena (EuroClimHist, CLMDAT...).

2.1 Struktura dat

Data nám byla předána jako „databáze“ v jedné tabulce vytvořené v Excelu. První zevrubná analýza odhalila velmi vágní textové popisy časového umístění události (roční období, svátky...) a fakt, že ne všechny záznamy jsou v češtině, jak jsme původně předpokládali.
Atributy dat:

Kod_stit – ID záznamu
Alfanumericky kod – viz. následující odstavec „Klíč“
Klic_sym – označení pro použití v GIS
Klic – textová varianta klíče definovaného alfanumerickým kódem (seznam odpovídajících klíčových slov)
Datum – textový popis data události
Zacatek – popis počátečního data události ve formátu RRRRMMDD
Konec – popis koncového data události ve formátu RRRRMMDD
Presnost – určuje přesnost zadaného časového rozložení d (den), k (kolem určitého data), p (po určitém datu), m (měsíc), o (období), r (rok)
Umist – lokalita výskytu (oblast, pohoří, obec...)
Nazok – pojmenování okresu
Kodob – kód oblasti, které se záznam týká
Nazev_udal – označuje druh události
Popis_udal – obsahuje kompletní přepis celé zprávy (přináší případné další časové popisy když je tam popsáno víc informací)
Pramen – zdroj záznamu

2.2 Klíč

Alfanumerický kód má hierarchickou strukturu o třech úrovních:

1. položka – hlavní kategorie události (26 – „povětrnostní podmínky“)
2. položka – dílčí skupina (26_5 – „vichřice“)
3. položka – popis události (26_3_22 – „vítr (ostrý)“)

Hlavní rozdělení klíčů:

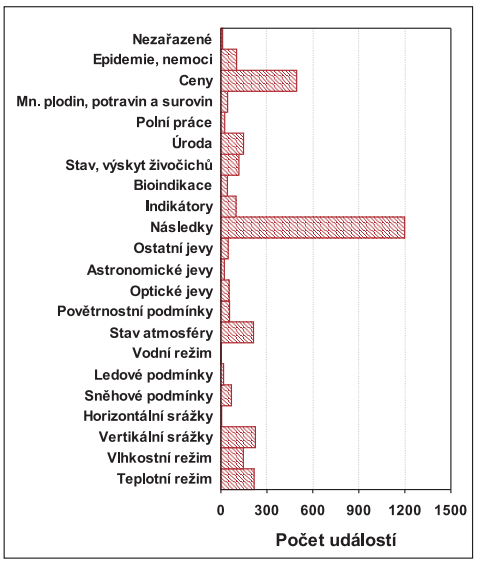
- **Teplotní režim** – 11 (chladno _1, mírně _2, teplo _3)
- **Vlhkostní režim** – 15 (sucho_1, vlhko_2)
- **Vertikální srážky** – 16 (déšť_1, kroupy_2, krupky_3, sněžení_4)
- **Horizontální srážky** – 17 (jinovatka_1, námraza_3, náledí_4)
- **Sněhové podmínky** – 18 (bez sněhu, obleva, lavina,...)
- **Ledové podmínky** – 19
- **Vodní režim** – 20
- **Oblačnost** – 25 (oblačnost, dohlednost, bouřka,...)
- **Povětrnostní podmínky** – 26 (problém s terminologií – povětří jako vzduch i vítr)
- **Optické jevy** – 27 (duha, polární záře, halo,...)
- **Astronomické jevy** – 28
- **Ostatní jevy** – 30 (zemětřesení, ...)
- **Následky** – 35 (nejčastější kategorie – smrt, škody,...)
- **Indikátory** – 40 (proxy informace – sucho, vlhko, sníh)
- **Bioindikátory** – 45 (kvetení, růst hub a podobně)
- **Stav, výskyt živočichů** – 50
- **Úroda** – 55
- **Polní práce** – 56
- **Množství surovin** – 60
- **Ceny** – 65
- **Epidemie** – 80
- **Nezařazené** – 99 (o mostech, zemědělských nezdarech,...)

2.3 Analýza dat

Analýza záznamů odhalila, že nezanedbatelná část jich není psána česky. Některé jsou německy, jiné latinsky. Latinské záznamy jsou navíc většinou doplněny o český překlad. Na několika místech je k původnímu záznamu přidána krátká vysvětlivka člověka, který daný záznam přidával do systému.

Původní téma projektu – extrakci informace z textu pro kontrolu přiřazení kódů k jednotlivým záznamům – jsme si proto rozšířili o klasifikaci záznamů dle jejich jazyka (pro otestování vlivu tohoto faktoru na výsledky přiřazování kódů).

Vzhledem k zadání můžeme v záznamech uvažovat jen atributy „identifikace záznamu“, „obsah“ a „alfanumerický kód klíče“. Ostatní atributy buď popisují pro analýzu nepodstatné údaje (místo, čas, autora záznamu, apod.), nebo jsou závislé na některém z výše zmíněných. Pro experimenty použijeme jen nejvyšší úroveň klíče. Získáme tím více příkladů nutných pro strojové učení a případné zahrnutí podklíčů už bude jen úpravou námi prezentovaného postupu.



Obrázek 1. Četnosti klíčů v datech

3 Přiřazení jazyka

Pro rozlišení jazyka záznamů jsme vyzkoušeli a porovnali výsledky několika technik.

3.1 Statistická metoda

Obsahy záznamů jsme porovnávali se stoplisty jednotlivých jazyků. U každého záznamu jsme vypočetli frekvenci, s jakou byla jednotlivá v něm obsažena slova nalezena ve stoplistu. Na základě této frekvence byl pak vybrán jazyk. V případě rovnosti frekvencí (například v případě, že v záznamu nebylo nalezeno žádné slovo ze stoplistů) měla přednost čeština.

Postup byl implementován v Javě (aplikace *LanguageGuesser*). Implementace umožňuje použít i větší množství stoplistů pro jednotlivé jazyky. Pro otestování jsme použili stoplisty češtiny, angličtiny, němčiny a holandštiny.
Testy proběhly následujícím způsobem:

- většina souborů (631) byla klasifikována jako české. Patří sem ty, které byly doopravdy klasifikovány, tak ty, jejichž obsah nebyl identifikován žádným stoplistem, z toho 625 správně
- méně textů (ale více než jsme předpokládali) bylo klasifikováno jako německé (210 souborů), z toho 209 správně
- 8 souborů bylo klasifikováno jako anglické, z toho 0 správně
- 2 soubory byly klasifikovány jako holandské, z toho 0 správně

Ruční klasifikací jsme pak zjistili, že se v záznamech dokonce vyskytuje i několik německých záznamů s českou vysvětlivkou a několik čistě latinských záznamů. Vzhledem k malému množství latinských záznamů, z nichž většina navíc byla doplněna o český překlad, jsme se rozhodli tyto záznamy klasifikovat jako české. Německé záznamy s českou vysvětlivkou jsme klasifikovali jako německé. Celkově tedy v sadě dat bylo:

- 10 latinských záznamů (k českým záznamům)
- 31 česko-latinských záznamů (k českým záznamům)
- 2 česko-německé záznamy (k německým záznamům)
- 211 čistě německých záznamů
- 597 čistě českých záznamů

Tedy:

- 638 záznamů klasifikovaných jako české
- 213 záznamů klasifikovaných jako německé

Tím je určena úspěšnost baseline klasifikátoru (přiřazena častější třída, tj. čeština) **74,97%**
P(Čeština) = 625/631 = **0,99**
R(Čeština) = 625/628 = **0,995** (nejsou brány v úvahu čistě latinské záznamy)
P(Němčina) = 209/210 = **0,995**
R(Němčina) = 209/213 = **0,981**

Výsledek použití této statistické metody může být jen těžko překonán – jeho přesnost i pokrytí je prakticky stoprocentní. Navíc kdybychom při testu porovnali jen nad českým a německým stoplistem, úspěšnost by byla ještě vyšší. I tak by pravděpodobně mohly být testy preciznější, pokud bychom:

- nalezli stoplisty bližší jazyku záznamů (stoplisty počítají s moderní verzí jazyků, kdežto záznamy jsou v češtině a němčině 16. století)
- využili lemmatizace
- nahradili stoplisty slov frekvenčním stoplistem n-gramů písmen

3.2 Zjištění jazyka pomocí strojového učení

Spíše pro porovnání jsme se rozhodli otestovat také metody strojového učení. Využili jsme učící algoritmy SMO, NaiveBayes a rozhodovací strom J48 (všechny se standardním nastavením). Protože se ale jedná o data nepoužívající dnešní jazyk (jak čeština tak němčina je několik století stará, navíc je vysoce pravděpodobné použití nářečí), vytvoříme učících množin několik:

- z datového souboru – množina několika delších záznamů přímo z databáze
- z moderních textů – texty získané z českých a německých zpravodajských serverů a beletrie (z guttenberg.org)
- ze starších textů – texty z odpovídající doby (16.–17. století) získané z guttenberg.org a dalších internetových zdrojů (těchto jsme nalezli velmi málo)

Pomocí existujících (*floods*) či nových skriptů a nástrojů (*LanguageTools*) jsme všechny záznamy vertikalizovali a převedli do formátu ARFF využívaného Wekou. Testovací množinu jsme získali ruční klasifikací všech záznamů (viz výše). Samotný experiment probíhal tak, že nejprve byly učící algoritmy naučeny na zadaných učících datech, a poté byly vygenerované klasifikátory spuštěny na zadaných testovacích datech. Výsledek experimentů byl prezentován pomocí html (*floods/rWeka*), zjištěny byly také nejisté klasifikace (*floods/uncertainHtml*).

Tabulka 1. Výsledky rozpoznávání jazyka pomocí metod strojového učení.

Učící množina	**SMO**	**DT J48**	**naiveBayes**
Záznamy	93,2 %	85,8 % (dle „der“)	79 %
Staré texty	93,1 %	73,7 % (dle „a“)	51,5 %
Nové texty	95,1 %	85,8 % (dle „der“)	27 %

Nejlepší výsledky mělo SVM, naivní Bayes na těchto datech pohořel. Rozhodovací strom, vzhledem k tomu, že rozhoduje jen na základě jednoho slova (německé „der“ nebo české „a“), má úspěšnost celkem dobrou, v případě použití „der“ lepší, než baseline 74,97%.

4 Získání informace z textu – kontrola přidělení kódu

Naším hlavním úkolem bylo otestovat, zda klasifikace do taxonomie kategorií (prováděná ručně) byla provedena správně – to jest, zkontrolovat přidělení klíče záznamům. Pro potřeby těchto experimentů jsme uvažovali jen nejvyšší úroveň klíče (hlavní kategorie).
I v tomto případě jsme využili dva různé přístupy:

- Porovnání **1 třídy versus všech ostatních** – záznamy rozdělíme do dvou tříd, jedna bude definována příslušností ke kategorii, do druhé budou patřit záznamy příslušné ke všem ostatním třídám.
- Porovnání **třídy a nějaké jiné třídy** – zde budeme zkoušet úspěšnost rozlišení dvou různých kategorií.

V obou případech jsme řešili problém s přidělením více klíčů, protože záznamy mohou patřit i do více kategorií (například záznam *„ukrutný déšť dopadl a přes noc následující byla zima a mráz, vody zmrzly a po ránu se na zamrzlém Labi utopilo pět lidí“* je přidělen do kategorie *„teplota“*, *„srážky vertikální-déšť“*, *„srážky horizontální-námraza“* i *„následky-úmrtí“*). V těchto případech jsme záznam přidělili pouze první uvažované třídě. V druhé (ať už se jedná o *„seznam zbývajících“* či *„jinou“*) je vymazán.

Tento přístup není optimální, hlavně při porovnávání dvojic tříd. Snižuje se jím totiž již tak dost malý počet záznamů, které jsou pro učení i testování k dispozici. Navíc – i pokud záznam patřící do obou tříd z jedné z nich smažeme,

stejně se tímto způsobem nezbavíme souvislosti obou tříd. Záznam patřící do obou vždy „zašpiní“ třídu A termy třídy B. Řešením tohoto problému by bylo vymazání „společného záznamu“ z obou tříd. Rozhodli jsme se ale upřednostnit vyšší počet záznamů k učení i testování a tento přístup nepoužili.

4.1 Metodika testů

Samotnou kontrolu jsme provedli pomocí desetisložkové křížové validace (tj. datová množina je rozdělena na deset částí, na devíti se učící algoritmy učí a na poslední se vytvořené klasifikátory otestují, postup se opakuje pro všech 10 složek). Při křížové validaci byly množiny stratifikované (tj. poměr tříd v jednotlivých podmnožinách byl stejný jako v celé množině).

Kromě porovnání úspěšnosti jednotlivých algoritmů (opět J48, SMO a NaiveBayes se standardním nastavením), jsme sledovali i závislost jejich úspěšnosti na počtu významných termů vybíraných pomocí metody filtrování na základě metriky information gain. Nejprve jsme testovali na 350 a 1000 termech, později jsme experimenty rozšířili i na 5, 10, 20, 50, 100 a 200 termů. Zjistili jsme totiž, že při nižším počtu rysů pravděpodobně bude dosaženo lepších výsledků.

Paralelně jsme zjišťovali výsledky dosažené na čistě českých a všech záznamech.

4.2 Průběh testů

Základem byly skripty vytvořené Janem Blaťákem, doplněné a upravené námi do podoby Javových nástrojů. Pro všechny třídy, učící algoritmy i množiny „nejzajímavějších“ termů byl spuštěn nástroj `kdd.jar`, kterým jsou data připravena pro křížovou validaci. Následně byly spuštěny jednotlivé průchody a z jejich výsledků získán průměrný výsledek jednoho experimentu (daného třídou, počtem termů a použitým učícím algoritmem). Výsledky experimentů, které se prováděly několik dní, byly uloženy do adresářů `data_vysledky` a `data_vysledky_cz` (podle toho, jestli byly jako zdroj dat použity záznamy všechny, nebo jen české).

4.3 Výsledky

Vzhledem k velkému počtu výsledků jsme se rozhodli prezentovat je pomocí grafů – znázorňujících závislost úspěšnosti jednotlivých učících algoritmů na množství vybraných rysů. V každém grafu jsou vyneseny průměrné výsledky všech učících algoritmů (zvlášť pro všechna a jen česká data) a přímka znázorňující úspěšnost baseline klasifikátoru. Přesto, že jsou tímto způsobem v jednom grafu znázorněny výsledky 48 základních experimentů (8 testů na počet rysů $\times$ 6 použití učičů), získáme 255 grafů. Kompletní sada grafů je přílohou této zprávy, zde uvedeme jen několik vybraných.
Podařilo se nám odhalit zákonitostí:

– třídy **1vsVše** je baseline mnohem vyšší – prakticky vždy má převahu třída „vše“; zatímco u tříd **1vs1** často dochází k porovnávání tříd velmi podobných, kdy se baseline blíží padesáti procentům.

- Analýzou výsledků (a je to vidět i z uvedených příkladů) jsme zjistili, že ve většině případů je na čistě českých datech dosaženo stejných či častěji lepších výsledků. Zlepšení ale není příliš velké, většionou se jedná o 1–3%. Tyto rozdíly jsou větší u tříd **1vs1**.
- Výsledky testů **1vsVše** příliš na počtu rysů nezávisí, často se s rostoucím počtem rysů mírně zlepšují (minimálně cca do počtu 200). Výsledky testů 1vs1 oproti tomu variují mnohem více.
- V případě tříd **1vsVše** se nejlépe osvědčil algoritmus **SMO**. Dosahuje velmi vyrovnaných výsledků, vysoké úspěšnosti. Dost často se tato úspěšnost při zvyšování počtu rysů dále zvyšuje. Naproti tomu naivní Bayesovský klasifikátor dosahuje lepších výsledků pro malý počet rysů, při jejich zvyšování se až na výjimky zhoršuje (až pod baseline).
- V případě tříd **1vs1** se chová mnohem vyrovnaněji (vzhledem k počtu termů) **naivní Bayesovský klasifikátor** – a má zde většinou také nejlepší výsledky. SMO má pro nižší počet termů výsledky přibližně stejné, při zvyšování cca nad 100 se ale jeho úspěšnost velmi rychle snižuje.
- V jakýchkoli testech naivní Bayesovský klasifikátor dával nejlepší výsledky při malém počtu rysů (cca do 20).
- Obecně se zdá, že nejlepší výsledky klasifikátory dosahují v oblasti 50–200 rysů (přičemž nB na spodní hranici a SVM na horní).

5 Závěry

První část úkolu – rozpoznání jazyka záznamu – byla nejlépe zpracována pomocí statistické metody nástrojem LanguageGuesser. Kontrola pomocí metod strojového učení odhalila, že podobný problém nejlépe řeší SVM, nicméně i tak zaostává za prakticky stoprocentní úspěšností statistické metody. Naivní Bayesovská metoda se na tento problém nehodí vůbec, dávala úspěšnost i hluboce pod baseline klasifikátorem.

Druhá část úkolu pak přinesla několik zajímavých informací. Nejdůležitější je, že výsledky testů více závisí na počtu rysů než na tom, zda jsou studovány jen české, nebo všechny záznamy. Zatímco u tříd 1vsVše se nejlépe osvědčilo SMO a počet rysů cca 200, u tříd 1vs1 to byl naivní Bayesovský klasifikátor a počet rysů v okolí 10. Obecně u tříd 1vsVše je možné nalézt více společných rysů, než u tříd 1vs1.

Došli jsme tedy k závěru, že metod strojového učení lze ke klasifikaci záznamů použít – hlavně při rozhodování, zda záznam patří do určité konkrétní třídy má klasifikace úspěšnost většinou více než 95 %. Klasifikace mezi dvěma třídami je méně úspěšná.

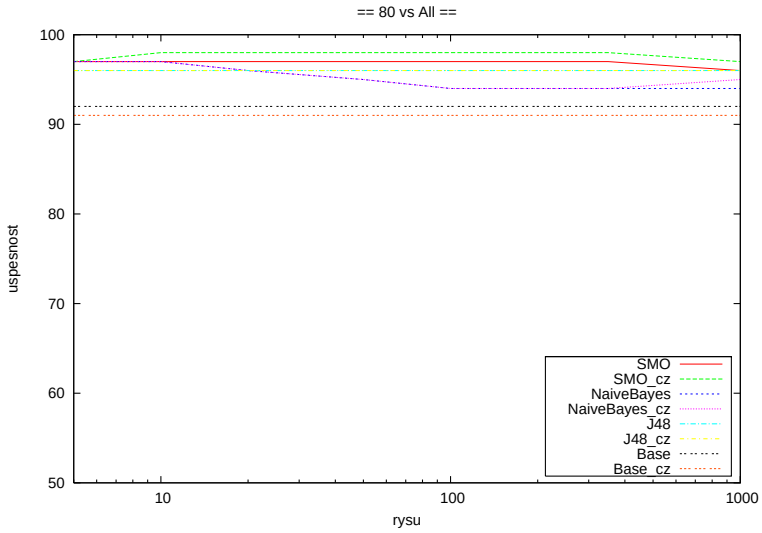
Reference

1. SCHÜLLE, H. – PFISTER, CH.: *Coding climate proxy information for the EURO-CLIMHIST Data Base, revised version (Summer 1993).* In: Frenzel, B. (ed.): In: Climatic trends and anomalies in Europe 1675-1715. Paleoclimate Research, Vol. 13, Verlag, Stuttgart, Jena, New York, 1994, s. 461-475)

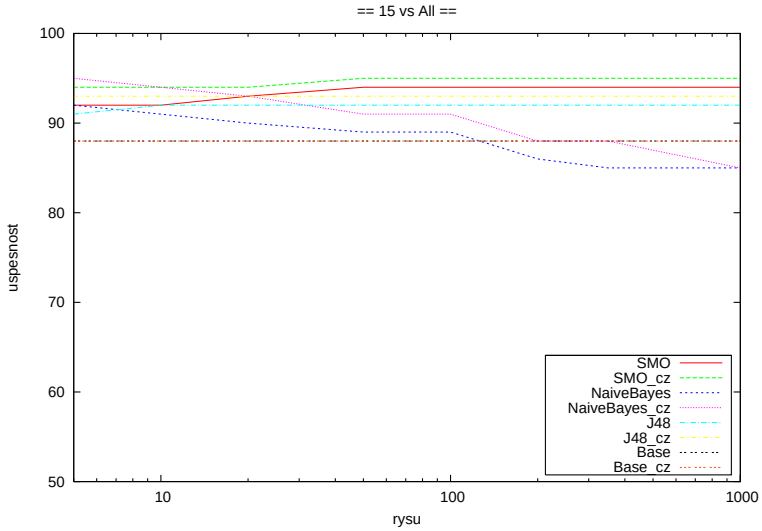
2. DOBROVOLNÝ, P. – FUKÁTKO, J.: *Digital mapping of climate history of the Czech Republic from documentary sources.*. In: Konečný, M., (ed.). In: International symposium on Digital Earth. Proceedings. CD ROM, 2003, s. 146-153. ISBN 80-210-3223-5.

6 Příloha

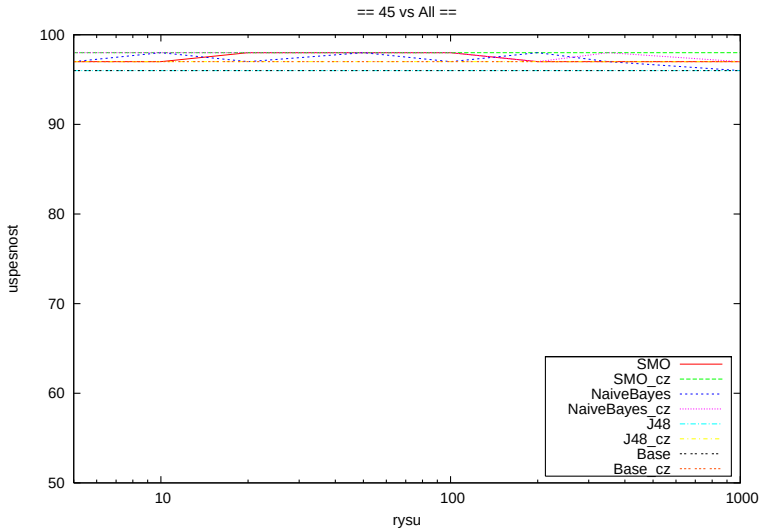
Následuje několik grafů s výsledky testů tříd 1 ku všem i jedna ku jedné. Ostatní grafy jsou přiloženy. Datové soubory, skripty taktéž.



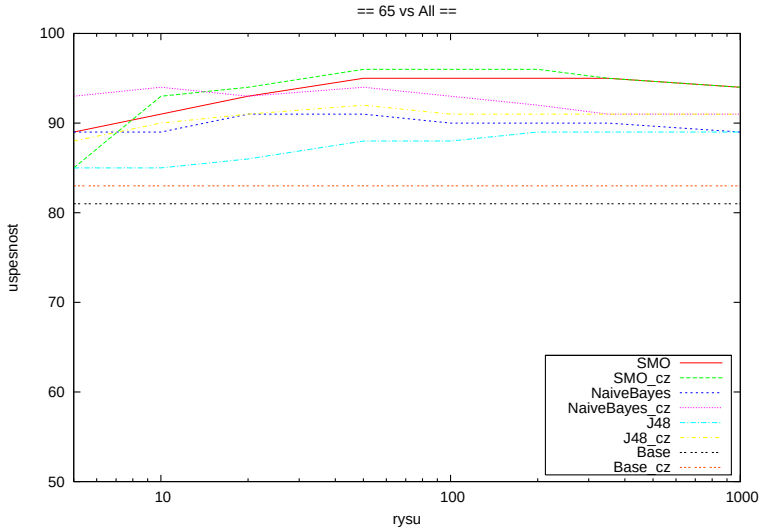
Obrázek 2. ‚Epidemie' versus ostatní



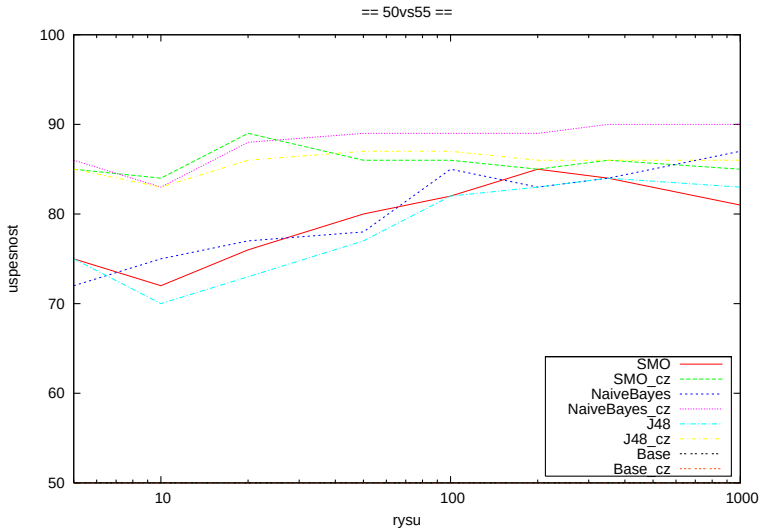
Obrázek 3. ‚Vlhkostní režim' versus ostatní



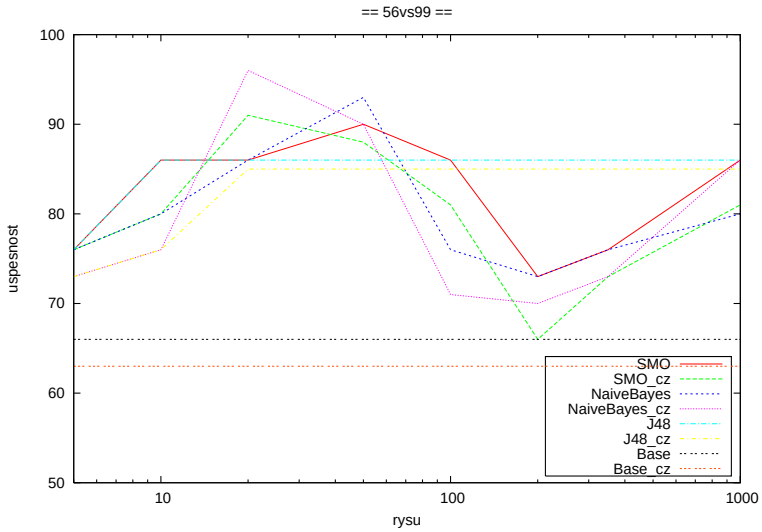
Obrázek 4. ‚Bioindikátory' versus ostatní



Obrázek 5. ‚Ceny' versus ostatní



Obrázek 6. ‚Stav, výskyt živočichů' versus ‚Úroda'



Obrázek 7. ‚Polní práce' versus ‚Nezařazené'