PŘEDPOVĚĎ PRAVDĚPODOBNOSTI VÝSKYTU METEOROLOGICKÝCH JEVŮ

Martin Dubrovský

Ústav fyziky atmosféry ČSAV, Hradec Králové[1]

1 Úvod

V rámci grantového projektu ČSAV je na našem pracovišti vyvíjen PC systém MESOMAP, jehož součástí je velmi krátkodobá předpověď ($\Delta t \sim 12$ h) pravděpodobnosti výskytu 'významných' meteorologických jevů. Tvar prognostické funkce pro vybraný jev,

$$(1.1) \qquad f(\cdot) = \Pr(Y=1 \,|\, X=\cdot)$$

kde X je r-rozměrný prediktor (vysvětlující proměnná) a Y je binární prediktand (vysvětlovaná proměnná), je hledán pomocí analýzy učebního datového souboru, $T = \{[x,y]_i;\ i=1,..,N\}$. Práce na vývoji prognostických procedur se v zásadě ubírá dvěma směry:

- výběr vhodných prediktorů nesoucích co největší informaci o prediktandu,
- výběr vhodného typu prognostické funkce a způsobu nalezení jejích parametrů.

Předmětem tohoto příspěvku jsou dvě metody odhadu pravděpodobnostní prognostické funkce (1.1) a porovnání jejich úspěšnosti při aplikaci na předpověď pravděpodobnosti výskytu bouřky. První metodou je konstrukce prognostické funkce ve tvaru binárního rozhodovacího stromu (BRS). Zde prezentovaná verze konstrukčního algoritmu vychází z práce ŘEZÁČOVÉ a MOTLA (1990). Algoritmus, který mezitím doznal několika změn, je podrobně popsán v připravované publikaci (DUBROVSKÝ, 199?). Druhou metodou je konstrukce prognostické funkce metodou k-nejbližších sousedů (kNS). Vodítkem při konkretizaci algoritmu mi byl vztah (2.1) v práci ANTOCHA (1988) a vlastní fantazie.

V této práci se předpokládá, že jednotlivé složky prediktorového vektoru $X = (X_1,X_2,...,X_r)$ jsou měřitelné či (v případě rozhodovacího stromu) dichotomické veličiny.

2 Binární rozhodovací strom

Na Obr. 1 je jednoduchá verze binárního rozhodovacího stromu, který vstupnímu prediktorovému vektoru x přiřazuje pravděpodobnost výskytu jevu reprezentovaného binárním prediktandem. Každý průchozí uzel (bifurkaci) stromu lze chápat jako kořen podstromu a zapsat: $t_k = \{b_k(\cdot),\ t_{k<},\ t_{k\geq}\}$, kde $b_k(\cdot)$ je bifurkační funkce a $t_{k<}$, $t_{k\geq}$ jsou dva dceřinné podstromy. Koncové uzly stromu (listy) nesou informaci o odhadu podmíněné pravděpodobnosti výskytu jevu, $\hat{p}_k$. Stavba BRS začíná nalezením bifurkační funkce kořenového uzlu, t_1, s využitím datového souboru $T_1 = T$, a pokračuje rekurzivní bifurkací dceřinných podstromů dokud všechny větve stromu nejsou ukončeny listy.

Algoritmus konstrukce podstromu t_k z učebního souboru T_k:

1.krok: Nalezení bifurkační funkce $b_k(\cdot)$, jež co nejúčinněji (ve smyslu zavedené míry diskriminace) odděluje prvky T_k s výskytem jevu $(Y=1)$ od prvků bez výskytu jevu $(Y=0)$.

2.krok: Test významnosti diskriminace.

3.krok (a): Je-li diskriminace významná, uzel t_k se stává bifurkací s bifurkační funkcí $b_k(\cdot)$, množina T_k je rozdělena na $T_{k<} = \{[x,y]_i \in T_k \,|\, b_k(x_i) < 0\}$ a $T_{k\geq} = \{[x,y]_i \in T_k \,|\, b_k(x_i) \geq 0\}$. Z učebních souborů $T_{k<}$ a $T_{k\geq}$ jsou zkonstruovány dceřinné podstromy $t_{k<}$ a $t_{k\geq}$.

---

[1] Husova 456, 50008 Hradec Králové

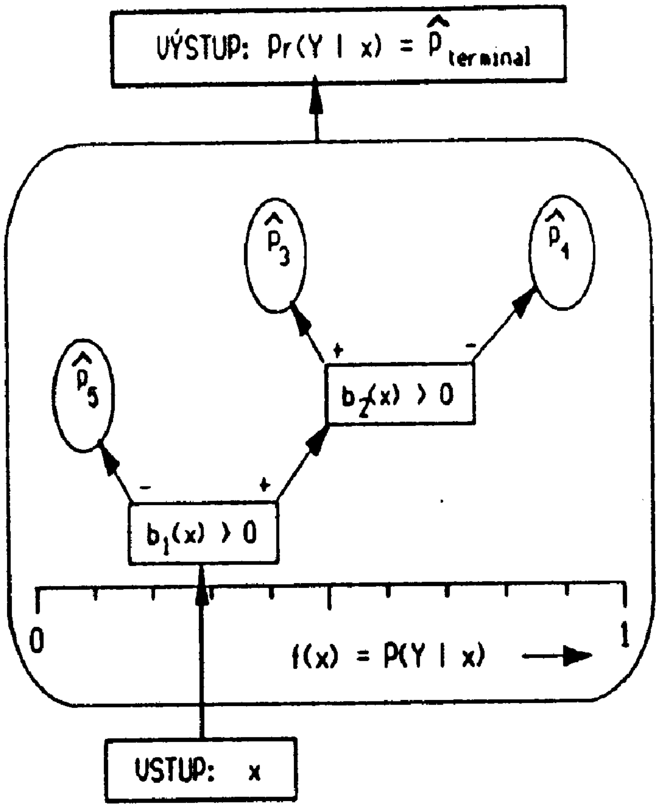
(b) Je-li diskriminace nevýznamná, uzel t_k se stává listem a je mu přiřazena prognostická pravděpodobnost výskytu jevu:

(2.1) $$\hat{P}_k = n_k / N_k$$

kde N_k a n_k je počet prvků v T_k celkem, resp. s $Y=1$.

V tomto obecném schématu konstrukčního algoritmu je třeba specifikovat:

- Míru kvality prognostické funkce a s ní související míru kvality bifurkační funkce
- Metodu štěpení stromu (množinu bifurkačních funkcí)
- Kritérium pro uzavření větve (test významnosti diskriminace)



Obr. 1 Schéma binárního rozhodovacího stromu pro předpověď pravděpodobnosti výskytu jevu. b_k ($k=1,2$) jsou bifurkační funkce průchozích uzlů (bifurkací), $\hat{P}_k$ ($k=3,4,5$) jsou podmíněné pravděpodobnosti výskytu jevu příslušné koncovým uzlům (listům) rozhodovacího stromu.

2.1 Míra kvality prognostické funkce

Je-li ztráta způsobená nepřesnou předpovědí popsána kvadratickou ztrátovou funkcí, $l(y_i, f(x_i)) = (y_i - f(x_i))^2$, kvalitu prognostické funkce $f(\cdot)$ lze vyjádřit střední kvadratickou chybou (MSE):

(2.2) $$e_f^2 = E[(y - f(x))^2]$$

Názornějším vyjádřením kvality prognostické funkce může být relativní skóre úspěšnosti:

(2.3) $$RV(f) = 1 - e_f^2/e_0^2$$

kde e_0^2 je střední kvadratická chyba průměrné klimatické předpovědi $[f_0(x) = n(T)/N(T) = \mathrm{konst}(x)]$ a RV se nazývá *redukce variance.*

Kvalitu bifurkační funkce lze definovat jako pokles MSE vzhledem k bifurkaci podle $b(\cdot)$:

(2.4) $$m(k \mid b) = p_k e_k^2 - (p_{k<} e_{k<}^2 + p_{k\geq} e_{k\geq}^2)$$

kde $p_\bullet$ ($\bullet = k, k<, k\geq$) je pravděpodobnost, že vektor X 'padne' do uzlu $t_\bullet$ a $e_\bullet^2$ je MSE daného uzlu:

(2.5) $$e_\bullet^2 = P_\bullet(1 - \hat{P}_\bullet)^2 + (1 - P_\bullet)\hat{P}_\bullet^{\,2}$$

kde $P_\bullet$ je podmíněná pravděpodobnost výskytu jevu pro daný uzel.

Při odhadu výše uvedených veličin z učebního souboru lze vycházet z odhadu MSE pro daný uzel:

(2.6) $$\hat{e}_\bullet^{\,2} = [n_\bullet(1 - \hat{P}_\bullet)^2 + (N_\bullet - n_\bullet)\hat{P}_\bullet^{\,2}]/N_\bullet$$

Odhady příslušných veličin budou dále značeny stříškou. Je-li příslušná veličina odhadnuta na testovacím souboru, nezávislém na vývojovém souboru (to je ten, na kterém byla nalezena bifurkační funkce a byly odhadnuty prognostické pravděpodobnosti podle 2.1), bude odlišena čarou ($\hat{e}^{2\prime}$, $R\hat{V}'$, $\hat{m}'$, N', n').

2.2 Štěpení stromu

Z množiny 'povolených' bifurkačních funkcí je v každém uzlu stromu hledána ta, která zajišťuje co největší míru diskriminace.

2.2.1 *Typ bifurkační funkce*

Standardní množinou bifurkačních funkcí je množina funkcí typu

$$(2.7) \qquad b(x) = x_{ix} - x^*$$

kde ix je index diskriminačního prediktoru (ix-tá složka prediktorového vektoru) a x^* je jeho kritická hodnota. V každém uzlu je bifurkační funkce volena tak, aby:

$$(2.8) \qquad b_k = \arg\max_{i=1,\dots,r} \left[\hat{m}(k \mid \bar{b}_i) \right] \quad \text{kde} \quad \bar{b}_i = x_i - \bar{x}_i^*$$

Kritické hodnoty $\bar{x}_i^*$ jsou nalezeny jednou ze dvou metod popsaných v ¶2.2.2.

Je-li prediktorový vektor dobře vychovaný (v ideálním případě s normálním rozdělením) a ne příliš rozm... lze hledat bifurkační funkci ve tvaru

$$(2.9) \qquad b_k(x) = \hat{w}_k x - x_k^*$$

kde $\hat{w}_k$ je diskriminační vektor {výpočet viz. např. Pudil et al. (1991), rovnice 7.2-5}. Kritická hodnota x_k^* je vypočtena jako kritická hodnota jednorozměrného prediktoru $z = \hat{w}_k x$ stejným způsobem jako v případě bifurkační funkce jedné proměnné. Aplikace bifurkační funkce typu (2.9) umožňuje rychlejší diskriminaci prediktorového prostoru a ve svém důsledku i vyšší kvalitu odhadu prognostické funkce. Nevýhody vzhledem k bifurkační funkci typu (2.7) jsou: (1) nutný předpoklad o 'vychovanosti' prediktorů, (2) neumožňuje redukci dimenze prediktorového vektoru.

2.2.2 *Kritická hodnota bifurkační funkce*

Nechť X_j je diskriminační prediktor (j-tá složka prediktorového vektoru). Jeho kritická hodnota, x_j^*, je vypočtena podle algoritmu:

Soubor T_k je náhodně rozdělen na dva stejně velké (případně lišící se o jeden prvek) podmnožiny T_k^a and T_k^b. Na každé z těchto podmnožin je nalezena kritická hodnota x_a^*, x_b^*. Toto se třikrát opakuje a výslednou kritickou hodnotou je aritmetický průměr z takto získaných šesti hodnot. Kritická hodnota prediktoru X_j na libovolné množině $T_k^\bullet$ ($\bullet$ = a,b) je určena jednou ze dvou metod:

<u>momentová metoda</u>:

$$(2.10) \qquad x_\bullet^* = \left[\bar{x}_{j0}\, s_{j1} + \bar{x}_{j1}\, s_{j0}\right] / \left[s_{j1} + s_{j0}\right] \; ; \qquad (\bullet = \text{a,b})$$

kde $\bar{x}_{ji}$ a s_{ji} ($i=0,1$) je průměr a směrodatná odchylka diskriminačního prediktoru X_j vypočtená z těch prvků $T_k^\bullet$ pro něž je $y=i$.

<u>maximální metoda</u> je založena na nalezení hodnoty prediktoru, pro níž míra diskriminace nabývá maxima:

$$(2.11) \qquad x_\bullet^* = \arg\max_{x^* \in R^1} \left[\hat{m}(k \mid b(x) = x_j - x^*)\right] \; ; \qquad (\bullet = \text{a,b})$$

Účinnost momentové i maximální metody v závislosti na velikosti učebního souboru, apriorní pravděpodobnosti výskytu jevu a tvaru distribuční funkce diskriminačního prediktoru byla testována jednak metodou Monte Carlo na náhodně generovaných souborech, jednak na reálných datech (viz. DUBROVSKÝ, 199?). Zde jen uvedu, že v případě dobře vychovaného diskriminačního prediktoru, a dobře (resp. špatně) separovatelného souboru ((ve smyslu diskriminační míry zavedené vztahem (2.4), s použitím bifurkační funkce typu (2.7)) je výhodnější použití momentové (resp. maximální) metody. V případě špatně vychovaných prediktorů je maximální metoda výhodnější vzhledem k své robustnosti.

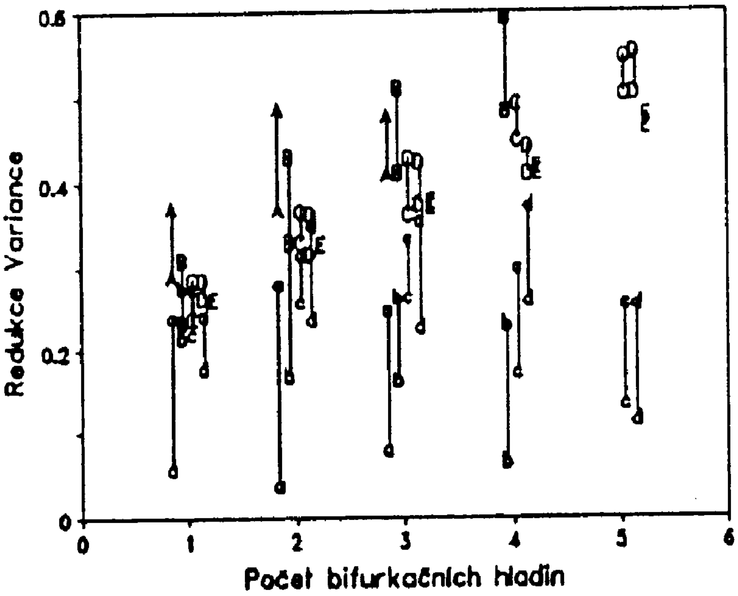
2.3 Kritérium pro uzavření větve

Je-li kvalita prognostické funkce odvozena od kvadratické ztrátové funkce, existuje nebezpečí, že štěpením stromu dojde k poklesu místo k růstu RV. Proto je třeba před každým štěpením stromu testovat významnost diskriminace. Testovací statistikou může být míra diskriminace vypočtená buď na vývojovém nebo nezávislém testovacím souboru, který je štěpen paralelně s vývojovým souborem. V případě statistiky vypočtené na vývojovém souboru je kritériem pro akceptování bifurkace:

(2.12) $$\hat{m}(k\,|\,b_k) \geq m^*(N_k,n_k,r)$$

přičemž síla testu klesá s rostoucím počtem prediktorů (r), z nichž je diskriminační prediktor vybrán. V případě statistiky vypočtené na nezávislém datovém souboru je síla testu nezávislá na r. Vzhledem k tomu, že výběrová míra diskriminace vypočtená na nezávislém testovacím souboru je nestranným odhadem její skutečné hodnoty, zdá se, že vhodným kritériem minimalizujícím pravděpodobnost špatného rozhodnutí (přijetí bifurkace když $m(k\,|\,b_k) < 0$ nebo zamítnutí bifurkace když $m(k\,|\,b_k) > 0$) je splnění podmínky:

(2.13) $$\hat{m}'(k\,|\,b_k) \geq 0$$



Obr. 2 Závislost redukce variance na výšce stromu (počet bifurkačních hladin). Malými [velkými] písmeny jsou vyznačeny intervaly $\langle \mathrm{avg}(R\hat{V}) \pm s(R\hat{V}) \rangle$ [$\langle \mathrm{avg}(R\hat{V}') \pm s(R\hat{V}') \rangle$], kde avg($\cdot$), s($\cdot$) jsou průměr a směrodatná odchylka z deseti simulací. $[N,N']$ = [100,500] (intervaly A–A, a–a); [200,500] (B,b); [500,500] (C,c); [700,300] (D,d); [1000,–] (E,–). Prediktand = $C_{<12-24>}$.

Důsledek zanedbání testu významnosti diskriminace je demonstrován na Obr. 2, kde je znázorněn vývoj redukce variance v závislosti na postupném, nekontrolovaném štěpení stromu. Z obrázku je patrná jednak rostoucí divergence větví $R\hat{V}$ a $R\hat{V}'$ s klesajícím N, jednak existence optimální velikosti stromu, která roste (stejně jako maximum hodnoty $R\hat{V}'$) s rostoucí velikostí vývojového souboru, přičemž překročení optimální velikosti stromu je doprovázeno poklesem kvality stromu ('Bugaboo' efekt).

3 Metoda k nejbližších sousedů

Odhad prognostické funkce metodou k nejbližších sousedů je váženým aritmetickým průměrem z těch pozorování, pro něž odpovídající x_i je mezi k 'nejbližšími' sousedy bodu, v němž odhadujeme (Antoch, 1988):

(3.1) $$f(x) = \sum_{j=1,\dots,k} q_{ix_j}\, y_{ix_j} \Big/ \sum_{j=1,\dots,k} q_{ix_j}$$

kde q_{ix_j} je váha ix_j-tého vektoru, ix_j je index j-tého nejbližšího souseda vektoru x vybraného z učebního souboru $\{[x_i,y_i];\ i=1,\dots,N\}$ tak, že $\delta(x,x_{ix_{j-1}}) \leq \delta(x,x_{ix_j}) \leq \delta(x,x_{ix_{j+1}})$. Definice vzdálenosti $\delta(x,x_{ix_j})$, definice váhy q_{ix_j} a počet sousedů k je předmětem optimalizace, která probíhá opět s využitím metody Monte Carlo. Nejchoulostivějším prvkem se zdá být definice vzdálenosti. Jako nejvýhodnější se ukazují dva typy vzdáleností:

(a) Mahalanobisova vzdálenost:

(3.2) $$\delta(x,x_i) = (x-x_i)^t\,\Sigma^{-1}(x-x_i)$$

kde Σ je kovarianční matice vektoru X. Pokus o zohlednění vzájemné vazby jednotlivých složek vektoru X s prediktandem vedl k modifikaci předešlé definice:

(3.3) $$\delta(x,x_i) = (x_k-x_{i_k})\,c_k\,\Sigma^{-1}_{kl}\,c_l\,(x_l-x_{i_l})$$; (k, l jsou sčítací indexy)

kde $c_\cdot$ ($\cdot = k,l$) je váha vyjadřující vazbu mezi prediktandem a $\cdot$-tou složkou vektoru X. Testování různě definovaných vah c však zatím nevedlo k jednoznačnému zlepšení.

(b) Vzdálenost projekcí dvou bodů na nejlepší diskriminační přímku:

$$\delta(x,x_j) = \left|\hat{w}.(x - x_j)\right| \tag{3.4}$$

kde $\hat{w}$ je diskriminační vektor určen stejným způsobem jako pro (2.9).

Pro volbu váhy q se ukázaly jako přibližně stejně dobré: (a) $q_{ix_j} = 1$; (b) $q_{ix_j} = [1/\delta(x,x_{ix_j})]^{\frac{1}{2}}$; (c) $q_{ix_j} = (1/j)^{\frac{1}{2}}$

Ke zrobustnění metody lze složky vektoru x_i transformovat dvěma způsoby: (a) $x_{ij} \to R_j(x_i)$; (b) $x_{ij} \to F^{-1}[R_j(x_i)/(N+1)]$, kde $R_j(x_i)$ je pořadí vektoru x_i v posloupnosti tvořené prvky učebního souboru uspořádanými podle j-té složky vektoru, N je počet individuí učebního souboru a F je distribuční funkce normálního rozdělení.

4 Aplikace na reálná data

Učební datový soubor sestává z 1058 pozorování čtyřrozměrného prediktorového vektoru [X = (*ADED2, KMOD, SICP, FI*), kde jednotlivé složky vektoru jsou indexy instability odvozené z aerologické sondáže Praha - Libuš] a čtyř různě definovaných binárních prediktandů (A, B, $C_{<12\text{-}18>}$, $C_{<12\text{-}24>}$) charakterizujících výskyt bouřky na daném území, v daném časovém intervalu. Pro každý ze čtyř prediktandů byla hledána prognostická funkce (1.1) jednak ve tvaru binárního stromu s bifurkační funkcí typu buď (2.7) nebo (2.9), jednak metodou k nejbližších sousedů se vzdáleností definovanou vztahem buď (3.2) nebo (3.4). Prediktorem je buď kompletní vektor X = (*ADED2, KMOD, SICP, FI*), nebo samotný Faustův index (*FI*) — index s nejvyšší individuální prediktivní schopností. Mírou úspěšnosti je redukce variance vypočtená na nezávislém datovém souboru, $R\vartheta'$. Pro každou kombinaci vstupních podmínek bylo provedeno 10 nezávislých simulací, z nichž byla spočtena průměrná hodnota a směrodatná odchylka $R\vartheta'$. Variabilita hodnoty $R\vartheta'$ je dána náhodným způsobem rozdělení vstupních dat na vývojový a testovací soubor, v případě konstrukce stromu má vliv též neurčitost při výpočtu diskriminační hodnoty.

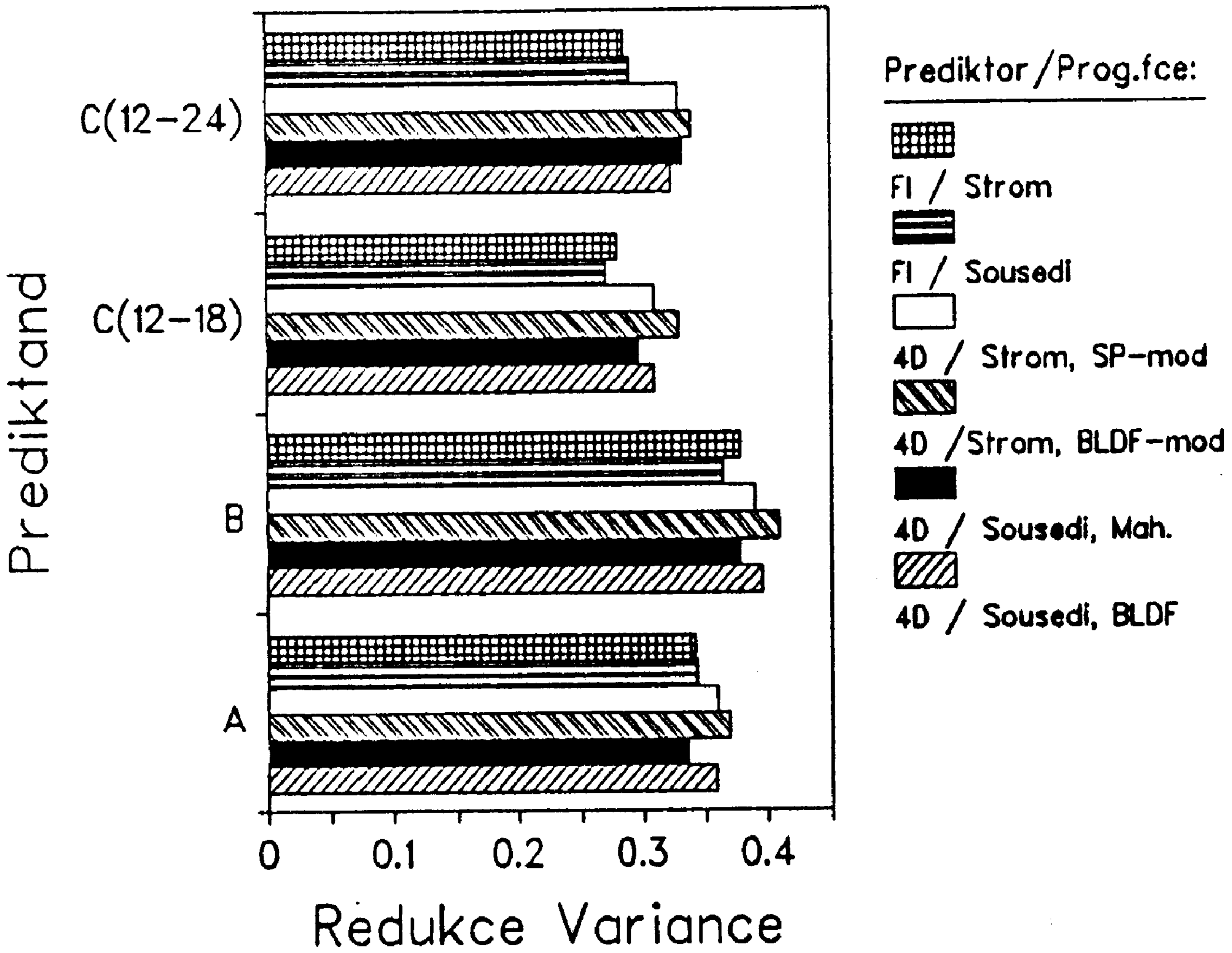
Tabulka. Výsledky 10 konstrukcí binárního rozhodovacího stromu. $R\vartheta'$ je redukce variance vypočtená na testovacím souboru ($N' = N = 529$), L je počet listů stromu. avg(·) a s(·) je aritmetický průměr a směrodatná odchylka vypočtené z 10 hodnot.

odhad diskrim. hodnoty:		momentová metoda			maximální metoda		
prediktor/bif.fce[1]:		FI/sp	4D/sp	4D/bldf	FI/sp	4D/sp	4D/bldf
Y= A	avg(Rϑ')=	0.34	0.36	0.37	0.35	0.35	0.38
	s(Rϑ')=	0.03	0.04	0.03	0.03	0.03	0.02
	avg(L)=	12	12	9	11	10	8
	s(L)=	3	3	2	2	3	2
Y= B	avg(Rϑ')=	0.38	0.39	0.41	0.38	0.38	0.40
	s(Rϑ')=	0.03	0.05	0.04	0.02	0.02	0.03
	avg(L)=	9	9	9	10	8	8
	s(L)=	2	3	2	2	2	3
Y= C<12-18>	avg(Rϑ')=	0.28	0.31	0.33	0.29	0.30	0.31
	s(Rϑ')=	0.04	0.04	0.03	0.03	0.05	0.03
	avg(L)=	8	9	7	9	9	7
	s(L)=	3	2	2	2	2	2
Y= C<12-24>	avg(Rϑ')=	0.29	0.33	0.34	0.31	0.34	0.33
	s(Rϑ')=	0.03	0.04	0.04	0.03	0.04	0.04
	avg(L)=	7	9	7	8	8	8
	s(L)=	2	3	2	2	1	3

1) FI/sp: x = FI, bifurkační funkce typu (2.7); 4D/sp: X = (*ADED2, KMOD, SICP, FI*), bifurkační funkce typu (2.7); 4D/bldf: x = (*ADED2, KMOD, SICP, FI*), bifurkační funkce typu (2.9).

V Tabulce jsou výsledky simulací stavby stromu, z nichž vyplývá:

- úspěšnost stromů využívajících vícerozměrný prediktor je vyšší než při použití jediného indexu (*FI*). Zlepšení je ovšem malé, což lze přičíst na vrub značné vzájemné korelovanosti jednotlivých indexů instability.
- použití bifurkačních funkcí ve tvaru nejlepší lineární diskriminační funkce (2.9) dává nevýrazně (vzhledem k variabilitě



Obr. 3 Skóre úspěšnosti [Redukce Variance = avg($R\hat{V}'$), kde $R\hat{V}'$ je redukce variance vypočtená na testovacím souboru, avg($\cdot$) značí průměr z 10 simulací] Binárního rozhodovacího stromu (Strom) a metody k nejbližších sousedů (Sousedi) aplikované na 4-rozměrný prediktor 4D = (*ADED2, KMOD, SICP, FI*) a samotný Faustův index (*FI*). Strom je staven buď v sp-módu (bifurkační funkce typu (2.7)) nebo bldf-módu (bifurkační funkce typu (2.9)), odhad diskriminační hodnoty je momentovou metodou. Odhad metodou *k*NS: $k = 50$, $q_i = (1/\delta_i)^{\frac{1}{2}}$, vzdálenost δ_i je buď podle (3.2) (Mah.) nebo podle vztahu (3.4) (bldf).

hodnot $R\hat{V}'$) vyšší skóre úspěšnosti než při použití bifurkačních funkcí jedné proměnné (2.7). Toto lze opět vysvětlit značnou vzájemnou závislostí jednotlivých prediktorů. Za povšimnutí stojí nižší počet bifurkací u stromů využívajících bifurkační funkce typu *bldf*, což je v souladu s očekávaným rychlejším vyčerpáním informace obsažené ve vícerozměrném prediktoru.

Na Obr. 3 je porovnána úspěšnost odhadu prognostické funkce metodou stromu a k nejbližších sousedů. Z obrázku lze vyčíst:

- Úspěšnost obou metod aplikovaných na 4-rozměrný prediktor je vyšší než při použití nejsilnějšího jednorozměrného prediktoru, nicméně toto zvýšení je vzhledem ke značné vzájemné závislosti mezi jednotlivými indexy nevýrazné.
- Úspěšnost obou metod (BRS vs. *k*NS) se v rámci chyby neliší
- Využitím nejlepší lineární diskriminační funkce (*bldf*) lze u obou metod (bifurkační funkce typu (2.9) při konstrukci stromu; definice vzdálenosti vztahem (3.4) u metody *k*NS)) sledovat určité zlepšení, což však může být důsledek toho, že použité prediktory jsou poměrně hodné (s trochou dobré vůle s normálním rozdělením).

5 Závěr

Kvalita prognostické funkce nalezené metodou analýzy učebního souboru je dána jednak kvalitou prediktorů (tedy mírou informace o prediktandu nesené prediktorem), jednak schopností použité statistické metody extrahovat z učebního souboru co nejoptimálnější prognostickou funkci. Vzhledem k tomu, že úspěšnost obou statistických metod testovaných v této práci vychází přibližně stejně, pro konečné rozhodnutí o výběru metody mohou hrát roli následující 'druhotná' kritéria:

- *rychlost*: konstrukce prognostické funkce ve tvaru stromu zabírá určitý časový interval, použití hotového stromu na konkrétní prediktorový vektor je však velice rychlé. Naproti tomu, v případě odhadu metodou *k* nejbližších sousedů je při každé aplikaci prognostické funkce vyhledání sousedů časově náročné.
- *náročnost na paměť počítače*: Odhad prognostické funkce metodou sousedů vyžaduje při každé aplikaci přístup k celému datovému souboru, kdežto prognostickou funkci ve tvaru stromu lze v paměti počítače uschovat mnohem úsporněji.
- *sebezdokonalování*: chceme-li, aby se prognostická funkce postupně zdokonalovala (tak jak se rozrůstá učební soubor), v případě *k*NS je toho dosaženo prostým přidáním nového pozorování k učebnímu souboru. V případě BRS je třeba při rozšíření učebního souboru konstruovat nový strom.
- metoda *k*NS se hůře vyrovnává s *rostoucí dimenzí*: v práci připravované pro publikaci (Dubrovský, 199?) je pro konstrukci stromu použit 50-ti rozměrný prediktorový vektor, s čímž se algoritmus konstrukce stromu s využitím bifurkačních funkcí typu (2.7) s přehledem vypořádal. Navíc algoritmus konstrukce BRS v sobě zahrnuje mechanismus pro redukci dimenze prediktorového vektoru. Pro *k*NS nastávají při rostoucí dimenzi prediktoru značné potíže vzhledem k nutnosti pracovat s kovarianční maticí - v případě obou použitých definic vzdálenosti.
- metoda *k*NS se hůře vyrovnává s kategoriálními prediktory

Závěrem bych chtěl dodat, že kombinace prediktorů použitá v této práci byla vybrána tak trochu subjektivně na základě vyhodnocení frekvence výskytu prediktorů v dolních uzlech stromů zkonstruovaných z učebního souboru zahrnujícího pozorované hodnoty 50 různých prediktorů. V další fázi práce se počítá jednak se zahrnutím nových druhů prediktorů do prediktorového vektoru, jednak s dalším zdokonalováním v této práci uvedených statistických metod, případně vyzkoušení jiných metod.

Literatura

ANTOCH J., 1988: Klasifikace a regresní stromy. *Sborník ROBUST 88*, JČSMF, str.0-6.

DUBROVSKÝ M., 199?: The binary decision tree: the growing algorithm and application to thunderstorm forecasting. Nabídnuto k publikaci v *Studia geophysica et geodaetica*.

PUDIL P., NOVOVIČOVÁ J. and BLÁHA S., 1991: Statistical approach to pattern recognition (Theory and practical solution by means of PREDITAS system). Supplement to the Journal *Kybernetika* 27, 78pp.

ŘEZÁČOVÁ and MOTL, 1990: The use of the simple 1D steady-state convective cloud model in the decision tree for determining the probability of thunderstorm occurrence. *Studia geoph. et geod.* 34, 147-166.