KLASIFIKACE A REGRESNÍ STROMY

Jaromír Antoch, MFF UK Praha

1. Úvod.

Řada statistických procedur, jež jsou každodenně používány, byla navržena pro datové soubory mající standardní strukturu. To mj. znamená, že všechny proměnné jsou téhož typu, homogenní apod. Bohužel, v současnosti se stále více a více setkáváme s daty (datovými soubory), pro něž je charakteristická:

- vysoká dimensionalita;
- směs datových typů (kategorické, měřitelné);
- nehomogenita (v různých částech výběrového prostoru platí různé vztahy);
- nestandardnost datové struktury (dimenze pozorování se mění objekt od objektu ...) apod.

Databanky sociologů, psychologů i techniků jsou plné příkladů, mnohdy sbíraných po řadu let. Velmi často se navíc jedná o neúplná data, kdy buď ne všechny sledované veličiny byly měřeny u všech objektů (nebo jednoduše chybí), či jsou, více či méně zjevně, nesmyslné.

Společným rysem všech datových souborů tohoto typu je jejich vysoká komplexnost, jež výrazně roste především se zvětšující se dimensionalitou problému. Dimensionalitou zde rozumíme počet pozorovaných proměnných. Je přitom obecným pravidlem, že čím větší je dimensionalita problému, tím řidší a rozptýlenější jsou data. Pravda, v některých situacích můžeme alespoň částečně tento problém vyřešit díky dostatku pozorování. Ale i zde musíme mít vždy na mysli, že ani velké množství pozorování, ani vysoká dimensionalita problému neimplikují bohatost vnitřní struktury. V jiných situacích je naopak situace komplikovaná tím, že nemáme k dispozici dostatek pozorování. Toto je typické např. v řadě medicinských problémů - naštěstí pro lidi, naneštěstí pro statistiky.

Navzdory všem uvedeným "problémům" jsou statistici každodenně svými klienty žádáni mimo jiné o:
- regresní, diskriminační či faktorovou analýzu,
- klasifikaci,
- výběr nejdůležitějších proměnných,
- redukci dimensionality problému,
- odhalení předpovědní struktury dat apod.

Pro řešení těchto problémů byla v literatuře navržena, a je více či méně běžně užívána, řada přístupů. Jedním z méně známých jsou metody založené na binárních stromech a rekurzivním dělení prostoru vysvětlujících proměnných, jež často poskytují zajímavý a velmi ilustrativní pohled na data. Hlavním cílem tohoto článku je upozornit české čtenáře na danou metodiku, již lze zařadit do oblasti neparametrických metod. Je přitom samozřejmé, že popsané postupy nemohou vyřešit všechny naše problémy a nahradit metody klasické. Ostatně si to ani nekladou za cíl. Nicméně, dle autorova přesvědčení, jedná se o velmi pružný nástroj neparametrické statistiky vhodný především v počáteční fázi analýzy dat.

Zatímco druhý odstavec je věnován použití dané metodologie v regresní analýze, třetí odstavec krátce popisuje použití v klasifikaci. Ve čtvrtém odstavci jsou pro ilustraci uvedeny dva příklady. Podrobný výklad celé metodologie postupů založených na binárních stromech nalezne zájemce v monografii Breiman a kol. (1984). Uvedené výsledky byly spočteny pomocí komerčního programového vybavení CART.

2. Regrese

Během posledních dvaceti let bylo věnováno značné úsilí rozvoji neparametrických metod. Velká pozornost přitom byla soustředěna na problematiku regresní analýzy. Pěkný bibliografický přehled o tom podává např. Collomb (1985). Mezi řadou navržených a vyšetřovaných odhadů hrají nejdůležitější roli modifikace odhadu pomocí k_n nejbližších sousedů, resp. jádrové odhady.

Společným rysem těchto neparametrických odhadů je skutečnost, že při jejich konstrukci vycházíme z určitého předem zvoleného způsobu dělení prostoru vysvětlujících proměnných, zatímco vliv naměřených hodnot vysvětlované proměnné se projevuje až druhotně. Ilustrujme si blíže tuto ideu na konstrukci dvou základních typů neparametrických odhadů, tj. odhadu pomocí k_n nejbližších sousedů, resp. jádrového odhadu.

Sledujme náhodný vektor $(Y,\underline{X})$, kde Y je reálná náhodná veličina (vysvětlovaná proměnná) a $\underline{X} = (X_1,\ldots,X_M)$, $M \geq 1$, je náhodný vektor vysvětlujících proměnných, jež mohou být jak měřitelné, tak kategoriální náhodné veličiny. Předpokládejme dále, že $\forall\, i$, $i=1,\ldots,M$, existuje prostor $\mathcal{X}_i \subseteq R_1$ jenž pokrývá všechny možné hodnoty veličiny X_i. Tzn., že všechny možné hodnoty vektoru $\underline{X}$

padnou do některého prostoru $\mathcal{X} = \mathcal{X}_1 * \ldots * \mathcal{X}_M \subseteq R_M$. Realizace vektoru $(Y,\underline{X})$ budeme značit $(y_i,\underline{x}_i)$, kde $\underline{x}_i=(x_{i1},\ldots,x_{iM})$, $i=1,2,\ldots,n$. Předpokládejme dále, že jednotlivé vysvětlující proměnné měříme, resp. jejich hodnoty zaznamenáváme, vždy v témže pořadí. Naším cílem je odhadnout neznámou regresní křivku $r(.) = E(Y \mid \underline{X} = .\ )$ na základě pozorování $(y_i,\underline{x}_i)$, $i=1,\ldots,n$.

Nejprve popišme konstrukci odhadu $r_n^1(.)$ pomocí k_n nejbližších sousedů v bodě $\underline{x} \in \mathcal{X}$. Nechť $\mathcal{K}_n(\underline{x}) = \{ i \mid \underline{x}_i$ je některý z k_n nejbližších sousedů mezi $\underline{x}_1,\ldots,\underline{x}_n$ k bodu $\underline{x} \}$, kde $k_n, n=1,2,\ldots$ je posloupnost přirozených čísel taková, že $k_n \to \infty$ a $k_n/n \to 0$, $n \to \infty$. Potom

$$(2.1) \qquad r_n^1(\underline{x}) = \sum_{i=1}^{n} y_i\, w_i\, I\,(i \in \mathcal{K}_n(\underline{x})),\ \underline{x} \in \mathcal{X},$$

kde w_i, $i=1,\ldots,n$, jsou váhy. Jedná se tedy o vážený průměr z těch pozorování y_i, pro něž odpovídající $\underline{x}_i$ leží "blízko" bodu v němž odhadujeme. Váhy zpravidla volíme tak, abychom preferovali ta pozorování y_i, pro něž odpovídající $\underline{x}_i$ leží blíže bodu $\underline{x}$ v němž odhadujeme, resp. abychom potlačili vliv odlehlých pozorování (outlierů), mezi y_i, $i \in \mathcal{K}_n(\underline{x})$. Z (2.1) je zřejmé, že odhad počítáme v $\underline{x} \in \mathcal{X}$ z pevného počtu pozorování.

Podobně u jádrového odhadu preferujeme zpravidla ta pozorování y_i (tj. dáváme jim větší váhu), pro něž odpovídající $\underline{x}_i$ leží blíže bodu v němž odhadujeme. Na rozdíl od $r_n^1(\underline{x})$ se však neomezujeme na pevný počet pozorování z nichž odhad počítáme.

Typický jádrový odhad funkce $r(\underline{x})$ lze zapsat ve tvaru

$$(2.2) \qquad r_n^2(\underline{x}) = \frac{\sum_{k=1}^{n} y_i K\left(\frac{\underline{x}-\underline{x}_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{\underline{x}-\underline{x}_i}{h_n}\right)}, \qquad \underline{x} \in \mathcal{X},$$

kde h_n, $n=1,2,\ldots$ je posloupnost nezáporných konstant taková, že $h_n \searrow 0$ pro $n \to \infty$ a $K(.)$ je některá pravděpodobnostní hustota. Bližší podrobnosti viz např. Antoch (1986), Collomb (1985) či Härdle (1988).

Jak je okamžitě vidět z obou uvedených definic, odhady typu (2.1)-(2.2) a jejich modifikace závisí podstatně na naměřených hodnotách vektoru vysvětlujících proměnných $\underline{X}$. Těžkosti s tím spojené jsou zvláště patrné ve vícedimensionálních problémech, kde jsou získané výsledky jen obtížně interpretovatelné a použitelné, především na okraji "oblaku dat". Podíváme-li se podrobněji na robustní verze těchto odhadů, uvidíme podobný jev.

V dalším textu se soustředíme na hlavní cíl tohoto odstavce, tj. na stručný popis konstrukce odhadů regresních křivek pomocí rekurzivního dělení podmnožin prostoru $\mathcal{X}$. Vzhledem k tomu, že tyto odhady lze graficky reprezentovat pomocí binární stromové struktury, budeme je nazývat <u>regresními stromy</u>. Jak již bylo řečeno, systematicky je lze zařadit mezi neparametrické odhady, přesněji řečeno mezi odhady po částech konstantní. Jedná se totiž o odhady rovné konstantě na podmnožinách prostoru $\mathcal{X}$, v kterémžto ohledu připomínají nejjednodušší neparametrický odhad, tzv. regresogram. Nicméně jejich konstrukce je zásadně jiná.

Připomeňme, že při konstrukci regresogramu rozložíme nejprve prostor $\mathcal{X}$ na L disjunktních obdélníků, zpravidla téže velikosti, a na každém z nich odhadneme neznámou regresní křivku konstantou. Zpravidla jako (vážený) průměr z těch pozorování y_i, pro něž odpovídající $\underline{x}_i$ padla do daného obdélníku. Všimněme si, že rozklad prostoru $\mathcal{X}$ na jednotlivé obdélníky opět <u>vůbec nezávisí</u> na naměřených hodnotách vysvětlované proměnné Y.

Při konstrukci binárních stromů je prostor $\mathcal{X}$ také nejprve rozdělen na L disjunktních podmnožin, zatímco v druhé fázi je neznámá regresní křivka odhadnuta na každé podmnožině konstantou. Základní rozdíl od konstrukce regresogramu spočívá v tom, že dělení jednotlivých podmnožin $\mathcal{X}$ (počínaje $\mathcal{X}$ samotným), je realizováno rekurzivně tak, aby se v každém kroku při dělení libovolné podmnožiny od sebe oddělila pozorování s vysokými hodnotami y_i vysvětlované proměnné od pozorování s nízkými hodnotami y_i.

Výsledkem je rozklad prostoru $\mathcal{X}$ na neprázdné podmnožiny $t_1,\ldots,t_L$ takové, že

$$(2.3) \qquad \bigcup_{i=1}^{L} t_i = \mathcal{X}, \quad t_i \cap t_j = \emptyset, \quad 1 \le i,j \le L.$$

Na každé z podmnožin t_i je neznámá regresní křivka odhadnuta konstantou. Výsledný odhad může být zapsán ve tvaru

$$(2.4) \qquad r_n^3(\underline{x}) = \sum_{i=1}^{L} c_i\, I(\underline{x} \in t_i), \quad \underline{x} \in \mathcal{X},$$

kde c_i, $i=1,\ldots,L$, jsou konstanty.

Základní otázky, okolo nichž se točí konstrukce hledaných odhadů, jsou následující:

1) Volba míry kvality odhadu.
2) Stanovení množiny otázek, podle nichž budou jednotlivé podmnožiny prostoru $\mathfrak{X}$ děleny.
3) Odhadnutí tvaru neznámé regresní křivky na jednotlivých podmnožinách.
4) Nalezení pravidla pro výběr optimálního dělení dané podmnožiny.
5) Stanovení pravidla pro ukončení dělení.

Za míru kvality odhadu se zpravidla volí střední čtvercová chyba, resp., chceme-li dosáhnout větší robustnosti procedury, střední absolutní chyba.

Jako odpověď na druhou otázku byla navržena tzv. standardní množina otázek Q, definovaná následovně:

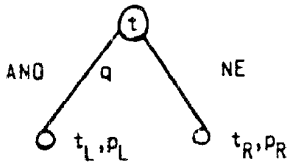
- Každé dělení závisí pouze na hodnotě právě jedné proměnné.
- Je-li X_i měřitelná náhodná veličina, zvolme otázky tvaru $\{\text{Je } X_i \leq c?\}$, $c \in R_1$. Poznamenejme, že vzhledem ke konečnému počtu pozorování jež máme k dispozici lze vždy užít konečný počet konstant c, a tedy i konečný počet otázek. Například lze volit jednotlivá c jako středy intervalů určených po sobě jdoucími pozorováními.
- Je-li X_i kategoriální náhodná veličina nabývající hodnot v množině B, potom Q zahrnuje otázky tvaru $\{\text{Je } X_i \in S?\}$ $S \subseteq B$.

Základní nevýhodou množiny Q, a tím i odpovídajících dělení, je skutečnost, že nepokrývá lineární závislosti, s nimiž se ve statistice běžně pracuje. Proto bylo třeba Q rozšířit jak o otázky pokrývající lineární kombinace měřitelných proměnných, tj. otázky typu $\{\text{Je } \sum a_i X_i \leq c?\}$, $c \in R_1$, tak o otázky pokrývající boolovské kombinace kategoriálních veličin. Přitom je třeba si uvědomit, že volba těchto složitějších otázek podstatně zvyšuje složitost výpočtu, neboť počet všech možných kombinací je s to brzy zahltit i současné superpočítače.

Jak vidíme, všechny uvažované otázky jsou dichotomického typu. Nechť $t \subseteq \mathfrak{X}$. Je zřejmé, že každé otázce $q \in Q$ odpovídá právě jedno dělení t na t_L a t_R takové, že

(2.5) $t_L \cup t_R = t$ a $t_L \cap t_R = \emptyset$.

Pro jednoduchost se dohodněme, že do t_L budeme vždy zařazovat ta pozorování z t, pro něž je odpověď na otázku q ANO, zatímco do t_R ostatní pozorování z t. Tuto situaci si můžeme schematicky znázornit pomocí



kde p_L (resp. p_R) značí podíl těch pozorování z t, jež padnou do t_L (resp. t_R).

Tvar odhadu, tj. hodnoty konstant c_i v (2.4) záleží dle očekávání podstatně na výběru míry kvality odhadu. Mějme k dispozici některý rozklad množiny $\mathfrak{X}$ tvaru (2.3). Zvolíme-li střední kvadratickou chybu, lze snadno ukázat, že optimální c_i jsou tvaru

(2.6) $$c_i = \frac{1}{\text{card } \delta_{t_i}} \sum_{\delta_{t_i}} y_j \quad , i=1,\ldots,L,$$

kde $\delta_{t_i} = \{ (y_j, \underline{x}_j) \mid \underline{x}_j \in t_i \}$. Analogicky, zvolíme-li za kritérium střední absolutní chybu, optimální c_i jsou tvaru

(2.7) $$c_i = \operatorname{med}_{\delta_{t_i}} y_j \quad , i=1,\ldots,L.$$

Obraťme se nyní k otázce číslo 4. Jak jsme již uvedli, jedním ze základních kamenů celého postupu je rekurzivní dělení jednotlivých podmnožin prostoru $\mathfrak{X}$. Nechť $t \subseteq \mathfrak{X}$, podívejme se, jak při dané množině otázek Q a zvolené míře kvality odhadu lze nalézt optimální rozklad t splňující (2.5). Pro ilustraci zvolme za míru kvality střední čtvercovou chybu. Jak jsme si již ukázali, ke každé otázce $q \in Q$ existuje některý rozklad splňující (2.5) a podle (2.6) k němu můžeme spočíst odpovídající hodnotu konstant c_L a c_R generujících odhad. Pomocí nich lze dále odhadnout střední čtvercovou chybu odpovídající tomuto odhadu, tj. odhadu generovanému rozkladem množiny t odpovídajícím otázce q. Její hodnotu si označme r(t,q). Optimálním dělením množiny t rozumíme dělení odpovídající otázce $q^* \in Q$ poskytující nejmenší střední kvadratickou chybu mezi všemi otázkami $q \in Q$, tj. pro niž $r(t,q^*) = \min_{q \in Q} r(t,q)$.

Je-li minima dosaženo pro více dělení, můžeme za optimální vybrat libovolné z nich.

Zbývá nám odpovědět na otázku číslo pět, která tvoří úhelný kámen celé metodologie. Začneme-li dělit prostor $\mathfrak{X}$, rekurzivním používáním předchozího postupu můžeme pokračovat tak dlouho, až v každé podmnožině zbývá právě jedno pozorování. Potom $c_i = y_i$, $i=1,\ldots,n$, což je však prakticky naprosto nepřijatelný odhad. Přirozené řešení, tj. zastavit dělení v okamžiku, kdy se výrazně zpomalil pokles zvolené celkové míry kvality odhadu se neosvědčilo ze dvou důvodů. Za prvé neexistuje rozumné pravidlo umožňující stanovit dolní hranici poklesu míry kvality odhadu. Zvolíme-li přísnou hranici, jsou výsledné odhady příliš "chudé" a mají malou vypovídací schopnost. Naopak, je-li hranice příliš volná, výsledné odhady jsou příliš rozsáhlé a nepřehledné. Za druhé, malý pokles míry kvality v jednom kroku nám nic neříká o možném velkém poklesu míry kvality v krocích následujících. Po mnoha pokusech bylo v metodologii CART přistoupeno k následujícímu postupu.

V prvním kroku je prostor $\mathfrak{X}$ rozložen posloupností rekurzivních dělení, popsaných v předchozím bodě, na velice mnoho co možná nejmenších podmnožin. Máme-li k dispozici dostatečně velkou paměť počítače, každá podmnožina výsledného rozkladu T_{max} bude obsahovat pouze jedno pozorování. V druhém kroku je aplikován algoritmus kolapsující (rekombinující) tento počáteční rozklad postupně zpět až do $\mathfrak{X}$. Při kolapsování se používá téže míry kvality odhadů, jako při konstrukci, tj. střední čtvercové (absolutní) chyby, modifikované však o člen penalizující nás za příliš rozsáhlé rozklady. Výsledkem kolapsování je posloupnost do sebe vnořených rozkladů prostoru $\mathfrak{X}$, počínající T_{max} a končící samotným prostorem $\mathfrak{X}$. Z této množiny je třeba vybrat řešení optimální (uvědomme si, že každý z rozkladů je jedním z možných řešení naší úlohy).

V zásadě existují dva postupy pro výběr optimálního řešení. Máme-li k dispozici dostatek pozorování, použijeme metodu testového souboru. Tzn., že pozorování rozdělíme na dvě skupiny. Zatímco pomocí jedné z nich budujeme hledaný odhad, druhou skupinu použijeme pro testování jeho kvality. Nemáme-li k dispozici dostatek pozorování, je zpravidla výhodnější metoda tzv. křížového ověřování (cross-validation), kde všechna data slouží jak pro konstrukci hledaného odhadu, tak k ověřování jeho kvality. Tento postup však vyžaduje jak víc paměti počítače, tak spotřebuje více času, neboť v každém kroku je třeba projít řadu pomocných řešení.

3. Klasifikace

Jak jsem se zmínil již v úvodu, vlastní metodologie byla původně vybudována pro potřeby klasifikace. Proto i monografie Breiman a kol. (1984) nejprve podrobně rozebírá tuto situaci a použití v regresi je pojednáno pouze jako speciální případ. Hlavní důvod, proč jsem ve svém příspěvku pořadí výkladu obrátil je ten, že zatímco užití "stromů" v klasifikaci je v té či oné formě více méně běžně užíváno, pro regresi toto neplatí. Vzhledem k existenci řady analogií s předchozím odstavcem zde budu ve výkladu postupovat mnohem rychleji.

Uvažujme situaci analogickou předchozímu odstavci. Tzn., že na jednotlivých objektech měříme M znaků, a to vždy v tom samém pořadí. Nechť náhodná veličina X_i, $i=1,\ldots,M$, popisující chování znaku i, může být buď měřitelného nebo kategoriálního typu a předpokládejme, že může nabýt hodnot z prostoru $\mathfrak{X}_i \subseteq R_1$. Položme $\underline{X} = (X_1,\ldots,X_M)$, potom všechny možné hodnoty vektoru $\underline{X}$ padnou do prostoru $\mathfrak{X} = \mathfrak{X}_1 * \ldots * \mathfrak{X}_M \subseteq R_M$. Předpokládejme dále, že každý objekt může být zařazen do právě jedné z J tříd. Jednotlivé třídy si označme identifikátory z množiny $C = \{1,\ldots,J\}$.

Podobně jako v jiných klasifikačních problémech je naším cílem nalézt klasifikátor (klasifikační pravidlo), tj. systematickou cestu umožňující předpovídat, do které třídy z C ten který objekt patří. Při konstrukci klasifikátoru mějme vždy na paměti především následující dva hlavní cíle:

- nalézt co nejpřesnější klasifikátor;
- odkrýt předpovědní strukturu problému.

Připomeňme přitom, že na libovolný klasifikátor se lze dívat dvěma navzájem ekvivalentními způsoby:

- jako na některou funkci $d: \mathfrak{X} \mapsto C$, takže $\forall\, \underline{x} \in \mathfrak{X}$ $d(\underline{x})$ je rovno jedné a právě jedné hodnotě z C, resp.;
- jako na některý rozklad prostoru $\mathfrak{X}$ na J navzájem disjunktních podmnožin A_j takových, že $\forall \underline{x} \in A_j$ $d(\underline{x}) = j$. Tzn., že $A_j = \{\underline{x} \in \mathfrak{X} \mid d(\underline{x}) = j\}$ a $\bigcup_{j=1} A_j = \mathfrak{X}$, $A_i \cap A_j = \emptyset$ pokud $i \neq j$. Jednotlivé podmnožiny A_j přitom nemusí být nutně souvislé.

Jak je zvykem, naši minulou zkušenost potřebnou pro konstrukci klasifikátoru soustředíme v tzv. učebním souboru, tj. množině měření znaků $X_1,\ldots,X_M$ na n individuích (objektech) spolu s informací o jejich skutečné klasifikaci. Jinými slovy, učebním souborem pro nás bude množina dvojic $\mathcal{L} = \{(\underline{x}_1, j_1),\ldots, (\underline{x}_n, j_n)\}$, $\underline{x}_i \in \mathcal{X}$, $j_i \in C$, i=1,...,n, kde jak $\underline{x}_i$ tak c_i známe.

Naším cílem je zkonstruovat klasifikátor typu binárního stromu, tj. klasifikátor, při jehož použití výsledná klasifikace bude záležet na zodpovězení konečného počtu dichotomických otázek. Jak již bylo uvedeno, při jeho konstrukci užijeme metodologii popsanou v předchozím odstavci. Je zřejmé, že bude třeba zodpovědět následující otázky:

1) Volba míry kvality klasifikátoru a způsob jejího odhadu.

2) Stanovení množiny otázek podle nichž budou jednotlivé podmnožiny prostoru $\mathcal{X}$ děleny.

3) Výběr pravidla přiřazujícího index třídy pro podmnožiny jež nebudou dále děleny.

4) Nalezení pravidla pro výběr optimálního dělení dané podmnožiny.

5) Stanovení pravidla pro ukončení dělení.

Za míru kvality klasifikátoru se zpravidla volí pravděpodobnost špatné klasifikace, resp. riziko klasifikátoru. Způsob jejich odhadu je otázka spíše technické povahy.

Množina otázek, podle nichž budou jednotlivé podmnožiny prostoru $\mathcal{X}$ děleny, se užívá přesně táž jako v regresním případě.

Výběr pravidla přiřazujícího index třídy je opět spíše technické povahy. Praxe ukazuje, že jeho výběr nehraje v celé konstrukci klasifikátoru zdaleka tak klíčovou roli jako odpověď na poslední otázku.

Základní idea při dělení některé množiny $t \subseteq \mathcal{X}$ na podmnožiny t_L a t_R splňující (2.5) spočívá v rozdělení objektů z t do t_L a t_R tak, aby data v obou podmnožinách byla "čistší", tzn. homogenější z hlediska klasifikace než v t. Jinými slovy to znamená, že množiny t_L a t_R by měly umožnit klasifikaci objektů, jež do nich padnou nejméně tak přesně jako t, resp. pokud možno přesněji. Technické provedení je analogické regresnímu případu. Rozdíl je pouze ve zvolené míře kvality klasifikace.

I v případě klasifikace se ukazuje být nejdůležitější odpověď na poslední, pátou otázku. Nejvíce se i zde osvědčila myšlenka zkonstruovat nejprve co nejrozsáhlejší klasifikátor, zpětně jej zkolapsovat (zrekombinovat) v posloupnost do sebe vnořených jednodušších klasifikátorů a z nich teprve vybrat optimální řešení. Základní myšlenky procedury pro konstrukci počátečního klasifikátoru a jeho kolapsování přitom zůstávají naprosto tytéž jako v regresním případě. Jediný rozdíl spočívá v nahrazení střední čtvercové (absolutní) chyby vhodným odhadem umožňujícím odhadnout pravděpodobnost špatné klasifikace, resp. riziko klasifikátoru.

Vedle hlavního cíle, tj. umožnění klasifikace (resp. predikce v regresním případě) je třeba zdůraznit, že popisovaná metodologie je schopna zvládnout mnohem více. Mezi jiným nabízí možnost:

- vyrovnat se s chybějícími pozorováními;
- stanovit důležitost sledovaných proměnných;
- nalézt nejlepší kompetitivní(náhradní) otázky v každém rozhodovacím kroku;
- snížit podstatně dimensionalitu problému pro potřeby klasifikace (resp. predikce);
- zvládnout klasifikaci i v situacích, kdy se mění dimensionalita sledovaných objektů případ od případu.

4. PŘÍKLADY

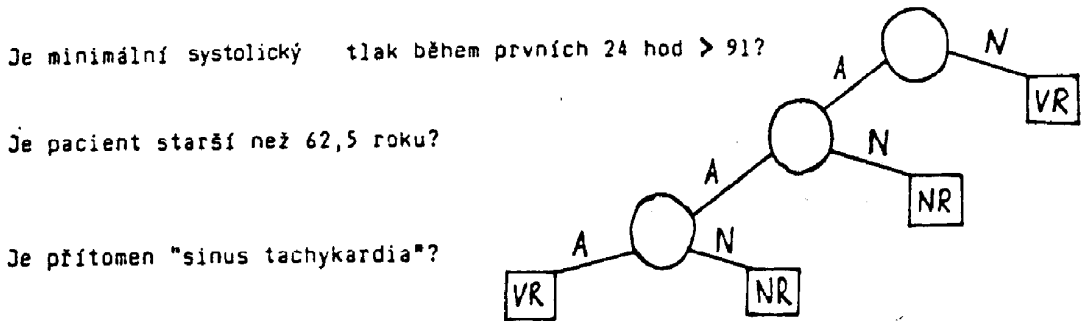
Příklad 4.1

V nemocnici Kalifornské university v San Diegu byli sledováni pacienti po infarktu. Z pacientů, kteří byli přivezeni do nemocnice s danou diagnózou byla vybrána skupina těch, kteří přežili alespoň 24 hodin. U nich bylo zjišťováno 19 charakteristik z následujících skupin otázek:

- anamnéza pacienta,
- výsledky EKG,
- charakteristiky měření sledujících uvolňování enzymů,
- historie charakteristické bolesti na prsou.

Pacienti, kteří přežili méně než 30 dnů (a více než 24 hodin), byli klasifikování jako vysoce riziková, zatímco pacienti, kteří přežili více než 30 dní, byli klasifikováni jako málo riziková. Cílem bylo nalézt klasifikátor umožňující klasifikovat nově příchozí na vysoce rizikové, resp. málo rizikové. K dispozici bylo 215 kompletních měření pacientů, z nichž 37 zemřelo dříve než 30 dní po infarktu, zatímco 178 přežilo déle.

Výsledné řešení vypadá následovně (užilo se desetinásobně křížové ověřování a Giniho kriterium pro dělení podmnožin):



Jednoduchost řešení, jež je skutečně tvaru binárního stromu, je až zarážející. Výsledek byl proto porovnán s jinými metodami, např. logistickou regresí. Ukázalo se přitom, že CART poskytuje z hlediska kvality klasifikace naprosto srovnatelné řešení s výsledky mnohem více sofistikovaných postupů, navíc však velmi jednoduché a snadno interpretovatelné. Toto je důležité především při použití v "terénní praxi". Ověření na datech z Birghimské nemocnice výsledky potvrdilo.

Příklad 4.2

Harrison a Rubinfeld (1978) sledovali vliv znečištění ovzduší na cenu domů v Bostonské oblasti. Data s sestávala z měření 14 proměnných u každého z 506 domovních bloků a stala se zvláště populární poté, co byla užita jako jeden z hlavních ilustrativních příkladů v monografii Belsley a kol. (1980), tzv. Boston housing data. Míra znečištění ovzduší byla posuzována podle naměřeného množství oxidu dusíku. U každého bloku byly sledovány následující veličiny:

y - medián hodnot domů v bloku v 10^3 US$ (MV);
x_1 - úroveň kriminality (CRIM);
x_2 - procento zastavěné půdy (ZN);
x_3 - procento velkoobchodů (INDUS);
x_4 - 1 jestliže blok je ve čtvrti River Side, 0 jinak (CHAS);
x_5 - koncentrace oxidu dusíku v pphm (NOX);
x_6 - průměrný počet pokojů (RM);
x_7 - procento domů vystavěných před rokem 1940 (AGE);
x_8 - vážená vzdálenost do center zaměstnání (DIS);
x_9 - přístupnost na okružní dálnici (RAD);
x_{10} - úroková míra (TAX);
x_{11} - průměrný počet žáků na učitele (P/T);
x_{12} - procento černochů (B);
x_{13} - procento "chudé" společnosti žijící v bloku (LSTAT).

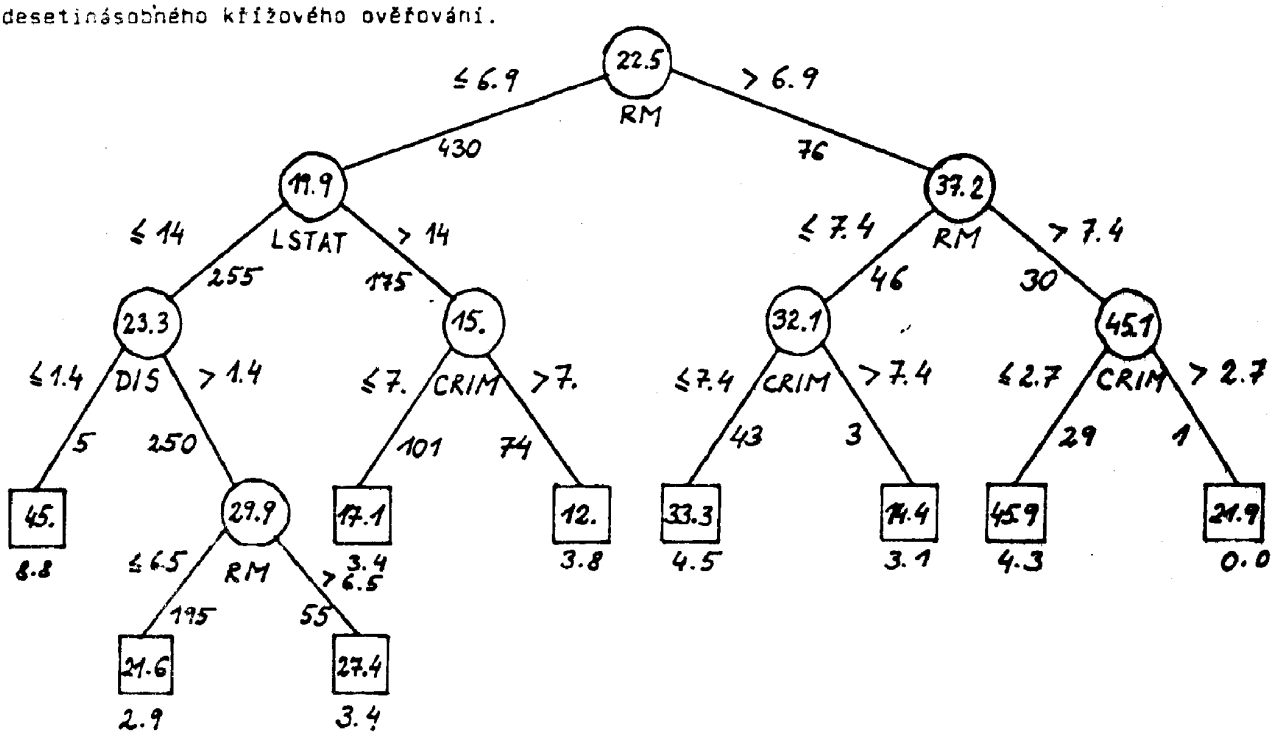
Cílem, jak již bylo zmíněno, bylo odhalit vliv znečištění ovzduší oxidem dusíku na ceny domů v Bostonské oblasti a zároveň odhalit předpovědní strukturu dat.

Po řadě transformací Harrison a Rubinfeld navrhli pro popis dat následující model (získaný metodou nejmenších čtverců):

$$\log(MV) = a_1 + a_2(RM)^2 + a_3(AGE) + a_4\log(DIS) + a_5\log(RAD) + a_6(TAX) + a_7(P/T) + a_8(B-63)^2 +$$
$$+ a_9\log(LSTAT) + a_{10}(TAX) + a_{11}(ZN) + a_{12}(INDUS) + a_{13}(CHAS) + a_{14}(NOX)^b + \varepsilon \ ,$$

kde $\underline{a} = (a_1,\ldots,a_{14})$ a b jsou parametry, jež je třeba odhadnout. Dosáhli přitom $r^2_{Y,\hat{Y}} \approx 0{,}81$.

Podívejme se nyní, jaký výsledek poskytla metoda CART při použití střední čtvercové chyby a desetinásobného křížového ověřování.



Jak lze interpretovat předchozí obrázek. Číslo uvnitř každého uzlu je průměr z hodnot závislé proměnné y pro všechny případy (objekty), jež padnou do daného uzlu. Např. pro všech 506 objektů je průměrná cena 2250 US. Pod každým nekoncovým uzlem je uvedena otázka, podle které je daný uzel dělen. Pod každým koncovým uzlem je místo toho uvedena směrodatná odchylka hodnot závisle proměnné v daném uzlu. Čísla u každé čáry vycházející z uzlu je počet případů (objektů) zařazených do levého (resp. pravého) synovského uzlu. Tzn., že z 506 bloků jich 430 mělo průměrný počet pokojů ≤ 6.9, zatímco u 76 tomu bylo naopak.

Chceme-li použít výsledný regresní strom pro predikci, je postupně potřeba odpovídat na otázky jež nám klade. Predikovaná hodnota pak je dána jako průměr z hodnot závisle proměnné pro odpovídající koncový uzel, tj. jako průměrná hodnota cen objektů, jež byly do téhož koncového uzlu zařazeny. Tedy např., je-li v našem konkrétním případě RM $\leq$ 6.9 & LSTAT $>$ 14 & CRIM $>$ 7.0, predikovaná hodnota bytu v takovémto bloku je 12.000 US.

Je zajímavé si všimnout, že z 13 vysvětlujících proměnných se ve výsledném řešení objevují pouze čtyři, tj. RM, LSTAT, DIS a CRIM. Proměnná charakterizující znečištění ovzduší, tj. NOX, se v něm nevyskytuje. Je však jednou z nejlepších náhradních veličin pro proměnné CRIM a LSTAT.

Celkově lze říci, že řešení je opět tvaru binárního stromu a odhaluje především výpovědní strukturu dat. Zároveň dobře vystihuje vzájemné vazby mezi proměnnými, zatímco význam pro predikci je až druhotný.

Literatura:

Antoch J.(1986), Neparametrické odhady regresních křivek. Sborník ROBUST 86, 1-20, JČSMF.

Breiman L., Friedman J.H., Olshen R.A. a Stone Ch.J.(1984), Classification and regression trees. Wadsworth, California.

CART (1986). California statistical Software, Inc., Lafayette, California.

Collomb G. (1985). Non-parametric regression: an up-to-date bibliography. Statistics, 16, 309-324.

Härdle, W.(1988), Applied non-parametric regression. Rukopis.

Harrison D. a Rubinfeld D.L.(1978), Hedonic prices and the demand for clean air. J.Envir.Econ. and Management, 5, 81-102.