KLASIFIKAČNÍ STROMY VE SPOLEHLIVOSTI SOFTWARE

JAN A. STROUHAL
jstrouhal@email.cz
ČVUT, fakulta strojní, katedra matematiky, Praha

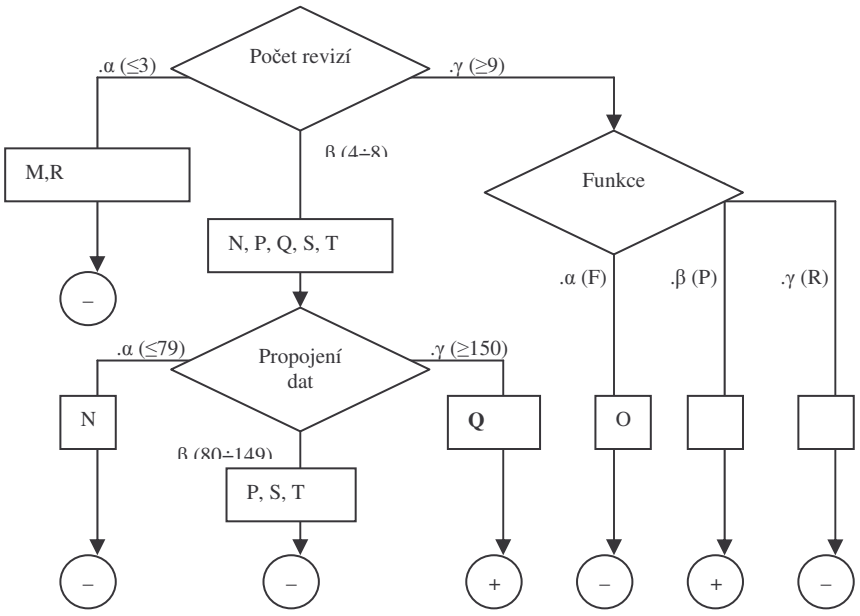
Abstrakt: Příspěvek se zabývá konstrukcí klasifikačních stromů pro použití ve spolehlivosti software. V první části je popsána základní konstrukce, která počítá s expertním počtem i hodnotami hranic metrik, v další části jsou popsány vylepšení této základní konstrukce vyhledáním optimálních hranic metrik. Příspěvek také popisuje rozšíření základní klasifikační škály na více skupin.

KLASIFIKAČNÍ STROMY

Vezmeme-li části programu (moduly), které jsme již v minulosti testovali, můžeme z jejich vlastností (metrik) zkonstruovat klasifikační strom, kterým pak můžeme klasifikovat části programu, u kterých chceme zjistit, zda je máme či nemusíme testovat. Tento zkonstruovaný strom nám může pomoci zkrátit dobu testování nebo dokonce zvýšit spolehlivost, protože se při testování zaměříme více na části, které jsou skutečně potřeba otestovat a odladit.

PŘÍKLAD KLASIFIKAČNÍHO STROMU

Ukážeme názorný příklad klasifikace modulu Q v již vytvořeném stromě. Modul má tyto vlastnosti: počet revizí 7, propojení dat 172 a funkci P (řízení procesu). V prvním kroku jej tedy klasifikujeme do větve β dle metriky počet revizí a následně do větve γ dle metriky propojení dat.



Pomocí klasifikace jsme zjistili, že klasifikovaný modul Q je pravděpodobně náchylný k chybám (skončil v + listu) a měli bychom jej testovat.

METODIKA TVORBY

- Rozklad na moduly
- Ohodnocení do skupin
- Zavedení, výběr a volba metrik
- Volba hranic
- Vhodné dělení

OPTIMALIZACE PŘI TVORBĚ KL. STROMU

Při vytváření klasifikačního stromu předem dané (expertní) hranice metrik nemusí být vždy optimální a mohou způsobovat vytvoření stromu se slepými koncovými listy nebo s koncovými listy, které jsou stále nehomogenní přestože jsou již vyčerpány všechny metriky.

OPTIMALIZACE POČTU HRANIC VÝBĚREM NEJLEPŠÍHO POČTU

Vypočteme stromy pro všechny možné počty hranic a pozorujeme maximum E() pro celý strom, počet použitých metrik a pozorujeme homogenitu v koncových listech (poměr homogenních listů ku všem listům).

- Neprovádíme v případech, kdy máme dva (a více) modulů se stejnými hodnotami všech metrik, ale s odlišnou skupinou (☒,☑).
- Pro všechny metriky volíme stejný počet hranic. Na počátku tvorby zvolíme jednu hranici a tento počet postupně zvyšujeme.
- Tvorbu stromů ukončíme ihned, jakmile budeme mít v právě prováděném stromu všechny koncové listy homogenními.
- Dodržujeme podmínku, že metrika nemůže mít více hranic, nežli je počet naměřených odlišných hodnot mínus jedna pro tuto metriku. Pokud nastane tato podmínka, pro tuto metriku již počet hranic nezvyšujeme.
- Pokud již nemáme možnost zvýšit počet hranic u žádné z metrik, tvorbu stromů ukončíme.

Na všech získaných stromech spočítáme *použitelnost* stromu jako počet použitých metrik, dělených násobkem E() a *homogenity* v koncových listech (násobek poměrů vyššího zastoupení skupiny ku všem modulům v listu). Ze všech stromů pak vezmeme ten, který má tuto použitelnost stromu nejvyšší.

OPTIMALIZACE POČTU HRANIC POSTUPNÝM ZVYŠOVÁNÍM

Budeme postupně vytvářet klasifikační strom běžným způsobem s proměnným počtem hranic a budeme zaznamenávat maximum E() pro celý strom a pozorovat homogenitu v koncových listech.

- Neprovádíme v případech, kdy máme dva (a více) modulů se stejnými hodnotami všech metrik, ale s odlišnou skupinou (☒,☑).
- Na počátku tvorby zvolíme jednu hranici pro všechny metriky.
- Tvorbu stromu ukončíme ihned, jakmile budeme mít všechny koncové listy homogenními.
- Pokud bude existovat nehomogenní koncový list, zvýšíme počet hranic o jedničku tam, kde nám vychází nejvyšší E().
- Pokud bude existovat nehomogenní koncový list a ve stejné metrice jako v bezprostředně předchozí tvorbě stromu se nám objeví maximum E(), které je vyšší nebo rovno maximu při bezprostředně předchozí tvorbě stromu, vrátíme počet hranic v této metrice na bezprostředně předchozí. V této metrice již nikdy počet hranic nezvyšujeme a zvyšujeme počet hranic u metriky, která má druhé (popř.další) nejvyšší E().
- Pokud bude existovat nehomogenní koncový list a ve stejné metrice jako v bezprostředně předchozí tvorbě stromu se nám objeví maximum E(), které je nižší než při bezprostředně předchozí tvorbě stromu, zvýšíme počet hranic o jedničku u této metriky jen tehdy, pokud nám nevyšel stejný strom, jinak vrátíme počet hranic v této metrice na bezprostředně předchozí. V této metrice již nikdy počet hranic nezvyšujeme a zvýšíme počet hranic u metriky s druhým (popř.dalším) nejvyšším E().
- Pokud při zvyšování počtu hranic v dané metrice narazíme na počet všech možných (odlišných) hodnot pro danou metriku mínus jedna, počet hranic u této metriky nezvyšujeme a zvyšujeme počet hranic u jiné.
- Nemáme-li již možnost zvýšit počet hranic u žádné z metrik, tvorbu stromu ukončíme.

ROZŠÍŘENÍ KLASIFIKAČNÍ ŠKÁLY

Klasifikace na „pozitivní“ a „negativní“ moduly, uvažovaná v literatuře se v praxi ukazuje jako nedostatečná. Vhodným vylepšením klasifikace je rozšíření klasifikační škály na více skupin podle důležitosti modulu a závažnosti následků jeho selhání. Takovým rozšířením může být například pětistupňová škála (pět skupin):

- netestovat (☒)
- testovat jen namátkově; projít (✖)
- testovat na hlavní funkce, zkráceně (○)
- testovat po určitou dobu (✓)
- důkladně otestovat, dokud nedosáhneme požadované spolehlivosti, případně do stanovené doby (☑)

FUNKCE VÝBĚRU METRIK PRO PĚTISTUPŇOVOU ŠKÁLU

Pro pětistupňovou škálu se změní funkce výběru, protože nyní již neuvažujeme pouze rozdělení na pozitivní a negativní moduly, ale rozdělujeme moduly do pěti skupin.Funkce výběru metrik pro pětistupňovou škálu (F je v rozsahu [0,1]) je:

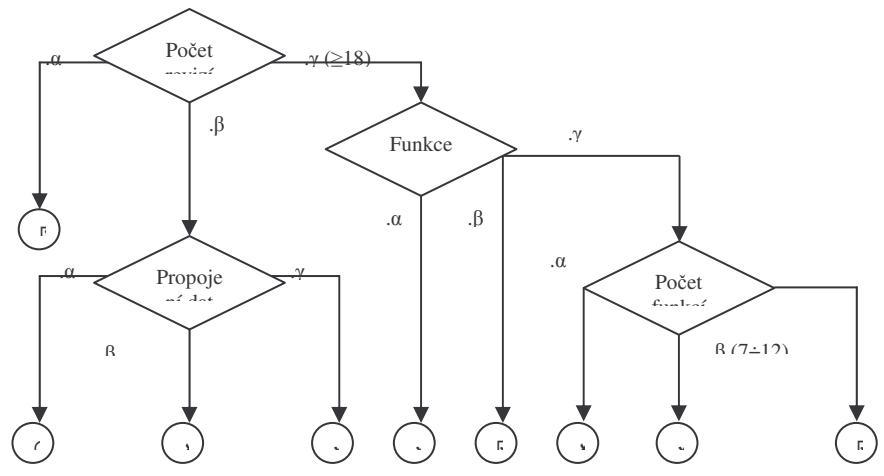
$$F(a_i;b_i;c_i;d_i;e_i) = -\frac{a_i}{s_i}\log_5\frac{a_i}{s_i} - \frac{b_i}{s_i}\log_5\frac{b_i}{s_i} - \frac{c_i}{s_i}\log_5\frac{c_i}{s_i} - \frac{d_i}{s_i}\log_5\frac{d_i}{s_i} - \frac{e_i}{s_i}\log_5\frac{e_i}{s_i},$$

kde a_i jsou moduly prvního stupně škály (méně náchylné k chybám – netestovat), b_i je počet modulů druhého stupně škály v i-tém uzlu, c_i je počet modulů třetího stupně škály v i-tém uzlu, d_i je počet modulů čtvrtého stupně škály v i-tém uzlu a e_i je počet modulů pátého stupně škály v i-tém uzlu. Součet s_i je součet všech modulů v i-tém uzlu $a_i+b_i+c_i+d_i+e_i$. Čím je *F()* menší, tím větší je homogenita v uzlu.

KRITÉRIUM TERMINACE PRO PĚTISTUPŇOVOU ŠKÁLU

Nejjednodušším kritériem terminace je nulaprocentní tolerance, požadující, že všechny moduly patří do určité skupiny, což se pro pětistupňovou škálu těžko dá zajistit, toto kritérium je příliš přísné.
Je lepší použít mírnější kritérium, požadující, že například polovina a více modulů patří do určité skupiny a zbytek patří do skupin sousedících. Například šest modulů patří do skupiny 2, dva do skupiny 1 (sousedící) a tři do skupiny 3 (také sousedící). Při vyhodnocování lze s úspěchem využívat i fuzzy matematického aparátu.

PŘÍKLAD SESTAVENÉHO KLASIF. STROMU PRO PĚTISTUPŇOVOU ŠKÁLU.



ZÁVĚR

V příspěvku jsem se zabýval konstrukcí klasifikačních stromů pro použití ve spolehlivosti software. V první části jsem popsal základní konstrukci, která počítá s expertním počtem i hodnotami hranic metrik, v další části jsem popsal vylepšení této základní konstrukce vyhledáním optimálních hranic metrik. V příspěvku jsem také popsal rozšíření základní klasifikační škály na více skupin.