Egy új nemparaméteres elv predikció konstruálására: szakértők kombinálása

Györfi László

Budapesti Műszaki és Gazdaságtudományi Egyetem
Számítástudományi és Információelméleti Tanszék

1 Bevezetés

Az informatika térhódításának jelenlegi szakaszában felerősödött az igény, hogy a korábban felgyűlt (sokszor igen nagy volumenű) adatsorainkból jó minőségű előrejelzéseket, predikciókat kapjunk a szóbanforgó folyamat, jelenségkör jövőbeni helyzeteivel, viselkedésével kapcsolatban.

Valószínűleg az informatika az egyetlen olyan diszciplína, amelyik nem definiálja általánosan a kutatásának a tárgyát, az információt. Az információval különböző műveleteket végzünk, és minden egyes alkalommal definiáljuk az információnak egy aspektusát. Az információt gyűjtjük, továbbítjuk, tároljuk, kódoljuk, kezeljük, visszakeressük, védjük, tömörítjük, stb. Egy ilyen felsorolásból jellegzetesen kimarad az, hogy miért tesszük mindezt, vagyis az, hogy a keletkezett adatokat feldolgozzuk, kinyerjük a bennük levő értéket.

Az információ feldolgozása alapvetően a matematikai statisztika feladata. A klasszikus statisztika jobbára paraméteres modellekből áll, amelyek arra az esetre vonatkoznak, amikor van valamilyen előzetes ismeretünk az aktuális problémáról, és így feltételezhetjük, hogy az ismeretlen valószínűségi törvény az eloszlások valamely paraméteres osztályához tartozik. Általában viszonylag kevés adatból is konzisztens statisztikai következtetést kaphatunk, továbbá meg tudjuk mondani, hogy előírt pontossághoz mekkora statisztikai minta kell, és adott hibakritériumhoz meghatározhatjuk az optimális eljárást.

A nemparaméteres statisztika esetében nem áll rendelkezésre ilyen előzetes tudás, ezért a következtetések konstruálása nem használhatja a törvény egyetlen speciális jellemzőjét sem, és az alapvető tulajdonságainak (például a konzisztencia) eloszlásfüggetleneknek kell lenniük. A paraméteres problémakörrel ellentétben itt nincs szuper eljárás, nem tudjuk megválaszolni a mintanagyságra vonatkozó kérdéseket, mert tetszőlegesen lassú lehet a konvergencia sebessége, amennyiben a szóbanforgó problémáról nem tudunk semmit, azaz nem élhetünk semmilyen modellfeltevéssel. Ennek a problémának a kezelésére szolgál az automatikus modellválasztás, amely történhet például szakértők kombinálásával.

Megemlíthető továbbá példaként az adatbányászat témaköre, aminek lényege, hogy nagy, korábbi adatsorokban lévő látens összefüggéseket szeretnénk feltárni és hasznosítani. Jól ismert például, hogy a nagy áruházláncok adatbázisaiban az eladásokkal kapcsolatos idősorok elemzésével értékes megállapítások nyerhetők fogyasztási szokásokról.

Az adataink igényesebb felhasználására irányuló törekvésnek tekinthető az univerzális predikció gondolatkörének megjelenése a 90-es években. Egy új információfeldolgozási megközelítésről van szó, ami széles általánosságban alkalmazható előrejelzési feladatokra, és a számítástudomány más eredményeivel ötvözve előrelépést hozhat több területen is. Az univerzális predikció lehetővé teszi, hogy számítógépes statisztikai módszereket alkalmazzunk olyan jelenségek elemzésére is, amelyeknél ez korábban nem látszott lehetségesnek.

A jó minőségű predikció egy igen fontos felhasználási területe a távközlő hálózatok mérésalapú forgalomszabályozása illetve hívásengedélyezése, ahol a hálózat átvitelét, kihasználtságát akarjuk maximalizálni úgy, hogy szigorú minőségi előírásokat kell teljesíteni (pl. ATM, mobil adatátvitel, stb.). Ehhez hasonlóan fel kell készülni arra, hogy a közeli jövőben az Internet forgalmazásnál is bevezetnek szolgáltatásminőséget, ugyanakkor az Internetforgalomra a forgalomelméletben megszokott modellelőírások nem teljesülnek (long range dependence).

A predikcióra eddig használt módszerek általában olyan modellekben működtek, ahol a szóbanforgó sorozat stacionárius és gyengén függő. Való életbeli feladatokban ez gyakran nehezen ellenőrizhető vagy egyáltalán nem teljesül. Gondoljunk csak a tőzsdei folyamatokra, ahol nyilvánvaló a nemstacionaritás. Az ilyen sorozatokat hívja az irodalom individuális sorozatoknak.

2 Predikció négyzetes költség esetén

A predikció feladatát formálisan a következőképpen fogalmazhatjuk meg. Legyen $y_1, y_2, \ldots$ valós számok egy sorozata és $x_1, x_2, \ldots$ d-dimenziós vektorok egy sorozata. Az $y_1, y_2, \ldots$ sorozatot szeretnék előrejelezni az ő múltjával és $x_1, x_2, \ldots$ sorozattal, vagyis célunk, hogy az $(x_1, \ldots, x_i, y_1, \ldots y_{i-1}) = (x_1^i, y_1^{i-1})$ megfigyelése alapján y_i értékére következtessünk. A prediktor egy $g = \{g_i\}_{i=1}^{\infty}$ függvénysorozat, $g_i(x_1^i, y_1^{i-1})$ jelöli az y_i becslését. n lépés után az empirikus négyzetes hiba az x_1^n, y_1^n sorozatra

$$L_n(g) = \frac{1}{n} \sum_{i=1}^{n} (g_i(x_1^i, y_1^{i-1}) - y_i)^2,$$

vagyis a g prediktort négyzetes költségfüggvénnyel minősítjük, és a prediktor hosszútávú viselkedését a költségek átlagával mérjük. A cél természtesen olyan prediktor konstruálása, amelyre $L_n(g)$ kicsi (Haussler, Kivinen, Warmuth (1998), Merhav, Feder (1998)).

Az univerzális predikció lényege az, hogy feltesszük, hogy K szakértő dolgozik a problémán, azaz adott K előrejelzés, $h^{(1)}, \ldots, h^{(K)}$, és megfigyelhetjük a szakértők $L_n(h^{(k)})$ veszteségeit. Az $n+1$-edik időpontban kombináljuk a szakértőket az eddigi eredményességük alapján. Például úgy, hogy generálunk egy olyan eloszlást a szakértők felett, amelyben egy szakértőnek nagyobb súlyt adunk ha eredményesebb volt, azaz kedvezőbb volt az $L_n(h^{(k)})$. Ezután az eloszlás szerint átlagoljuk a szakértők predikcióit (Cesa-Bianchi et al. (1997), Littlestone, Warmuth (1994), Vovk (1990), Weinberger, Merhav and Feder (1994)).

Ennek az általános problémának egy statikus speciális esete a regressziófüggvénybecslés. Legyen Y valós értékű valószínűségi változó és legyen X d-dimenziós véletlen vektor (megfigyelés). X koordinátái különböző eloszlásúak lehetnek, lehet némelyik diszkrét (például bináris), mások lehetnek abszolút folytonosak. Így nem teszünk fel semmit X eloszlásáról. A regresszióanalízis célja Y becslése, ha X adott, azaz olyan g valós függvényt keresünk, amelyre $g(X)$ "közel" van Y-hoz. Tegyük fel, hogy az analízis fő célja a négyzetes középhiba minimalizálása:

$$\min_g \mathbf{E}\{(g(X) - Y)^2\}.$$

Jól ismert, hogy a minimumot az

$$m(x) \stackrel{\text{def}}{=} \mathbf{E}\{Y|X = x\}$$

regressziófüggvény éri el, ugyanis minden g függvényre

$$\mathbf{E}\{(g(X)-Y)^2\} = \mathbf{E}\{(m(X)-Y)^2\} + \mathbf{E}\{(m(X)-g(X))^2\}$$

A jobb oldal második tagját a g függvény integrált négyzetes hibájának nevezzük és $J(g)$-vel jelöljük:

$$J(g) \stackrel{\text{def}}{=} \mathbf{E}\{(m(X)-g(X))^2\}.$$

A négyzetes középhiba nyilván pontosan akkor lesz közel a minimumhoz, ha $J(g)$ közel van a 0-hoz.

A regresszióbecslés feladatánál adottak az (X,Y)-nak a $D_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ független, azonos eloszlású példányai. Az $m(x)$-nek olyan

$$m_n(x) = m_n(x, D_n)$$

becslését akarjuk megkonstruálni, amelyre $J(m_n)$ kicsi, azaz az m_n regressziófüggvény becslés m-hez való konvergenciáját vizsgáljuk.

Az m_n becslő **gyengén (erősen) univerzálisan konzisztens** ha

$$J(m_n) \to 0 \quad \text{sztochasztikusan (1 valószínűséggel)}$$

(X,Y) minden olyan eloszlására, amelyre $\mathbf{E}\{Y^2\} < \infty$.

Stone (1977) mutatott rá először, hogy léteznek gyengén univerzálisan konzisztens becslők. A következő alakú becslőket vizsgálta:

$$m_n(x) = \sum_{i=1}^{n} W_{ni}(x; X_1,\ldots,X_n) Y_i.$$

Ezek a becslők **lokális átlagoló** tipusúak, mivel a W_{ni} súlyok jellegzetesen nemnegatívak, összegük 1, továbbá W_{ni} "nagy", ha x és X_i "közel" van egymáshoz, egyébként W_{ni} "kicsi".

A **k-legközelebbi szomszéd becslő** esetében, $W_{ni}(x; X_1,\ldots,X_n) = 1/k$, ha X_i az x k legközelebbi szomszédjának egyike $X_1,\ldots,X_n$ közül, különben $W_{ni} = 0$. Ha

$$k_n \to \infty, \quad k_n/n \to 0,$$

akkor a k_n-legközelebbi szomszéd becslő gyengén univerzálisan konzisztens (Stone (1977)). Ha

$$\lim_{n\to\infty} k_n/\log(n) = \infty, \quad \lim_{n\to\infty} k_n/n = 0,$$

akkor a k_n-legközelebbi szomszéd becslő erősen univerzálisan konzisztens (Devroye, Györfi, Krzyżak and Lugosi (1994)). A konzisztencia feltételei igen

általánosak, ezen belül a konvergenciasebesség nagyon különböző lehet. A k jó választása, az adaptáció tehát egy alapvető kérdés, amely történhet szakértők kombinálásával.

A **hisztogram vagy partíciós becslő** $\mathcal{R}^d$ egy $\mathcal{P}_n = \{A_{n,1}, A_{n,2} \ldots\}$ partíciója esetén az

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K_n(x, X_i)}{\sum_{i=1}^n K_n(x, X_i)},$$

függvény, ahol $K_n(x,u) = \sum_{j=1}^\infty I_{[x \in A_{n,j}, u \in A_{n,j}]}$ (I_A az A esemény indikátorfüggvényét jelöli). Gyenge és erős univerzális konzisztenciára vonatkozó eredményeket bizonyított Devroye, Györfi (1983) és Györfi (1991).

A regresszió függvényt az

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)},$$

magfüggvényes becslővel is lehet becsülni ahol $h = h_n > 0$ egy n-től függő simító tényező, K egy abszolút integrálható függvény (a magfüggvény), és $K_h(x) = K(x/h)$. A magfüggvényes becslés gyenge és erős konzisztenciáját bizonyította Devroye és Wagner (1980), Spiegelman és Sacks (1980), Devroye és Krzyżak (1989).

Az eddig ismertetett regresszióbecslési módszerek a lokális átlagolás elvén alapultak. Létezik egy másik, hasonlóan természetes alapelv, az empirikus hibaminimalizálás, amely szintén elvezethet univerzálisan konzisztens becslésekhez. Választunk egy $\mathcal{F}_n$ függvényosztályt, és a regressziócslés ebből az osztályból veszi az értékeit. Az $\mathcal{F}_n$ kiválasztásakor vagy az m regressziófüggvényről szerzett ismereteinket vesszük figyelembe, vagy $\mathcal{F}_n$ olyan függvényekből áll, amelyek számítógéppel bizonyos számítási bonyolultsággal realizálhatók. Az **empirikus hibaminimalizálás** alapötlete, hogy az

$$\frac{1}{n} \sum_{j=1}^n |f(X_j) - Y_j|^2$$

empirikus hibát minimalizáljuk, és azt a függvényt válasszuk ki $\mathcal{F}_n$ függvényosztályból, amelynek az empirikus hibája minimális (Györfi, Kohler, Krzyżak, Walk (2002)).

Az univerzális predikció problémájának másik forrása az idősorok, vagyis a **függő megfigyelés**ek esete. Itt az adatok a $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ sorozat, amelyről feltesszük, hogy stacionárius és ergodikus. A feladat minden n időpontban a

$$\min_g \mathbf{E}\{(g(X_{n+1}, D_n) - Y_{n+1})^2\}$$

minimalizálás. A legjobb prediktor a

$$\mathbf{E}\{Y_{n+1}|X_{n+1}, D_n\}$$

feltételes várható érték, amely nem tanulható meg olyan értelemben, hogy nincs olyan g_n predikciósorozat, melyre

$$\lim_{n\to\infty}(g_n(X_{n+1}, D_n) - \mathbf{E}\{Y_{n+1}|X_{n+1}, D_n\}) = 0$$

egy valószínűséggel minden stacionárius és ergodikus adatsorozatra.

Ha az adatok egy Gauss-folyamat, akkor ez a feltételes várható érték, vagyis a prediktor lineáris, és a probléma lényegesen egyszerűbb. Sajnos az újabb idősoraink jellegzetesen messze nem Gauss-folyamatok, sőt ha a megszokott lineáris előrejelzéseket használjuk, akkor gyakran használhatatlan eredményeket kapunk. Ilyen kellemetlen tapasztalatokra számíthatunk, ha az emberi tényezőnek szerepe van az illető idősor keletkezésében, mint például a pénzügyi folyamatok esetén.

Általánosan a vágyunk az

$$L^* = \lim_{n\to\infty} \min_g \mathbf{E}\{(g(X_{n+1}, D_n) - Y_{n+1})^2\}$$

optimum elérése, amely az előzőek miatt nem lehetséges. Ugyanakkor létezik Cesáro-konzisztens predikciósorozat g_n, azaz amelyre

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(g_i(X_{i+1}, D_i) - Y_{i+1})^2 = L^*$$

egy valószínűséggel minden stacionárius és ergodikus adatsorozatra. Az ilyen predikciósorozatot **univerzálisan konzisztens**nek nevezzük, és vázlatosan megmutatjuk, hogyan konstruálható univerzálisan konzisztens prediktor szakértők kombinálásával.

1. lemma: *(Kivinen, Warmuth (1999)). Jelölje $h_1, h_2, \ldots$ predikcióknak egy családját (szakértőket) és $\{q_k\}$ egy valószínűségeloszlást. Tegyük fel, hogy $|h_k(y_1^{n-1})| \le B$ és $|y_n| \le B$. Legyen*

$$w_{n,k} = q_k e^{-(n-1)L_{n-1}(h_k)/c},$$

ahol $c = 8B^2$, és

$$v_{n,k} = \frac{w_{n,k}}{\sum_{i=1}^{\infty} w_{n,i}}.$$

Ha a kombinált g predikciót úgy definiáljuk, hogy

$$g_n(y_1^{n-1}) = \sum_{k=1}^{\infty} v_{n,k} h_k(y_1^{n-1}),$$

akkor

$$L_n(g) \le \inf_k \left(L_n(h_k) - \frac{c \ln q_k}{n} \right).$$

Ennek a lemmának az egyik speciális esete az, amikor véges sok, K szakértő (prediktor) van és $\{q_k\}$ az egyenletes eloszlás. Ekkor

$$L_n(g) \le \min_k L_n(h_k) + \frac{c \ln K}{n},$$

vagyis a kombinált szakértő hibája alig nagyobb, mint a legjobb szakértő hibája.

Bizonyítás. A bizonyítás elemi. Legyen $W_1 = 1$ és $W_t = \sum_{k=1}^{\infty} w_{t,k}$, ha $t > 1$. Mivel

$$\begin{aligned}
W_{t+1} &= \sum_{k=1}^{\infty} w_{t+1,k} \\
&= \sum_{k=1}^{\infty} q_k e^{-tL_t(h_k)/c} \\
&= \sum_{k=1}^{\infty} q_k e^{-(t-1)L_{t-1}(h_k)/c - \left(y_t - h_k(y_1^{t-1})\right)^2/c} \\
&= \sum_{k=1}^{\infty} w_{t,k} e^{-\left(y_t - h_k(y_1^{t-1})\right)^2/c} \\
&= W_t \sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - h_k(y_1^{t-1})\right)^2/c},
\end{aligned}$$

ezért

$$-c \ln \frac{W_{t+1}}{W_t} = -c \ln \left(\sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - h_k(y_1^{t-1})\right)^2/c} \right).$$

$c = 8B^2$ miatt az $F(z) = e^{-(y_t - z)^2/c}$ függvény konkáv a $[-B, B]$ intervallumon, tehát a Jensen egyenlőtlenség miatt

$$\begin{aligned}
&\exp\left(\frac{-1}{c} \left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - h_k(y_1^{t-1}) \right) \right]^2 \right) \\
\ge\ &\sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - h_k(y_1^{t-1})\right)^2/c},
\end{aligned}$$

ezért $t > 1$ esetén

$$\left[\sum_{k=1}^{\infty} v_{t,k}\left(y_t - h_k(y_1^{t-1})\right)\right]^2 \le -c \ln \frac{W_{t+1}}{W_t}.$$

Következésképp

$$\begin{aligned}
nL_n(g) &= \sum_{t=1}^{n}\left(y_t - g(y_1^{t-1})\right)^2 \\
&= \sum_{t=1}^{n}\left[\sum_{k=1}^{\infty} v_{t,k}\left(y_t - h_k(y_1^{t-1})\right)\right]^2 \\
&\le -c\sum_{t=1}^{n} \ln \frac{W_{t+1}}{W_t} \\
&= -c \ln W_{n+1} \\
&= -c \ln\left(\sum_{k=1}^{\infty} w_{n+1,k}\right) \\
&= -c \ln\left(\sum_{k=1}^{\infty} q_k e^{-nL_n(h_k)/c}\right) \\
&\le -c \ln\left(\sup_k q_k e^{-nL_n(h_k)/c}\right) \\
&= \inf_k \left(-c \ln q_k + nL_n(h_k)\right) \ .
\end{aligned}$$

Térjünk vissza a függő megfigyelések problémájára, vagyis amikor $(X_1, Y_1), \ldots, (X_n, Y_n)$ stacionárius és ergodikus. Tegyük fel, hogy $|Y_0| \le B$. Egy elemi prediktort (szakértőt) jelöljön $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$. Legyen G_ℓ az $\mathcal{R}^d$ ill. H_ℓ az $\mathcal{R}$ kvantálója. Rögzített k, ℓ esetén legyen I_n azon $k < i < n$ időpontok halmaza, amikor van k hosszú illeszkedés a kvantált sorozatra:

$$G_\ell(x_{i-k}^{i}) = G_\ell(x_{n-k}^{n})$$

és

$$H_\ell(y_{i-k}^{i-1}) = H_\ell(y_{n-k}^{n-1}).$$

Akkor a szakértő predikciója az i illeszkedési időpontokhoz tartozó y_i-k átlaga:

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = \frac{\sum_{i \in I_n} y_i}{|I_n|}.$$

Egyenként ezek a szakértők nem univerzálisan konzisztensek, ugyanis kicsi k esetén a becslés torzítása nagy, másrészt nagy k esetén a szórás lesz

nagy, mert kevés az illeszkedés. Ugyanez mondható el a kvantálókra. A probléma az, hogy az adatok függvényében hogyan választható meg k, ℓ. Ennek a feladatnak egy lehetséges megoldása a szakértők kombinálása.

A kombinált predikciót az előző lemma segítségével konstruáljuk, azaz választunk egy $\{q_{k,\ell}\}$ valószínűségeloszlást (k,ℓ)-en, és $c = 8B^2$ esetén legyen

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/c}$$

és

$$v_{t,k,\ell} = \frac{w_{t,k,\ell}}{\sum_{i,j=1}^{\infty} w_{t,i,j}} \ .$$

Ekkor a kombinált predikció

$$g_t(x_1^t, y_1^{t-1}) = \sum_{k,\ell=1}^{\infty} v_{t,k,\ell} h^{(k,\ell)}(x_1^t, y_1^{t-1}) \ .$$

1. tétel: *(Györfi, Lugosi (2001)). Ha a G_ℓ és H_ℓ kvantálók "aszimptotikusan finomak", és $\mathbf{P}\{Y_i \in [-B,B]\} = 1$, akkor a kombinált g prediktor univerzálisan konzisztens.*

3 Alakfelismerés: $0-1$ költség

Az alakfelismerés általános feladatában az y_i $\{1, 2, \dots M\}$ értékű. Az i időpontban a döntőnek (osztályozónak, preditornak) tippelni kell y_i-re, ha rendelkezésre áll a múlt (x_1^i, y_1^{i-1}).

n lépés után az empirikus hiba az x_1^n, y_1^n sorozatra

$$L_n(g) = \frac{1}{n} \sum_{i=1}^{n} I_{\{g(x_1^i, y_1^{i-1}) \neq y_i\}},$$

vagyis a költség a $0-1$ költség, és $L_n(g)$ a hibák relatív gyakorisága.

Az alakfelismerés statikus feladatában az $\{1, 2, \dots M\}$ lehetséges értékű Y valószínűségi változó (címke, osztály) értékét kell eldönteni adott X d-dimenziós véletlen vektor (megfigyelés) alapján. A döntés vagy osztályozási szabály egy

$$g : \mathcal{R}^d \to \{1, 2, \dots M\}$$

döntésfüggvény. Az osztályozási szabály minőségét az

$$L(g) \stackrel{\text{def}}{=} \mathbf{P}\{g(X) \neq Y\}.$$

hibavalószínűség méri. A lehető legjobb (legkisebb) hibavalószínűséget a Bayes-döntés adja:

$$g^*(x) \stackrel{\text{def}}{=} i, \text{ ha } \mathbf{P}\{Y = i|X = x\} = \max_j \mathbf{P}\{Y = j|X = x\}.$$

A Bayes-döntés hibáját, $L(g^*)$-ot, Bayes-hibának nevezik, és L^*-gal is jelölik. A Bayes-döntéshez ismernünk kellene Y-nak X feltétel melletti feltételes eloszlásait. Az alakfelismerés feladatánál ezek ismeretlenek, de rendelkezésünkre állnak a $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ adatok, melyek (X, Y) független és azonos eloszlású példányai. Egy olyan adatoktól függő

$$g_n(x) = g_n(x, D_n)$$

osztályozási szabályt akarunk megkonstruálni, amelyre

$$L_n \stackrel{\text{def}}{=} L(g_n) = \mathbf{P}\{g_n(X) \neq Y|D_n\}$$

közel van L^*-hoz.

Most is kétfajta univerzális konzisztenciáról beszélhetünk: A g_n osztályozási szabály **gyengén univerzálisan konzisztens**, ha

$$\mathbf{E}L_n = \mathbf{P}\{g_n(X) \neq Y\} \to L^*$$

(X, Y) minden eloszlására. A g_n osztályozási szabály **erősen univerzálisan konzisztens**, ha

$$\mathbf{P}\{g_n(X) \neq Y|D_n\} \to L^* \quad 1 \text{ valószínűséggel}$$

(X, Y) minden eloszlására.

A regressziófüggvénybecsléshez hasonlóan itt is két egyszerű alapelv fogalmazható meg. Az egyik a **lokális többségi döntés** elve:
A k**-legközelebbi szomszéd szabály** a

$$g_n(x) = \arg\max_{j \in \{1,2,\ldots M\}} \sum_{i=1}^{n} W_{n,i}(x) I_{\{Y_i = j\}}$$

döntésfüggvény, ahol $W_{n,i}$ az előző szakaszban definiált legközelebbi szomszéd súly.
A **partíciós szabály** a

$$g_n(x) = \arg\max_{j \in \{1,2,\ldots M\}} \sum_{i=1}^{n} I_{\{X_i \in A_n(x)\}} I_{\{Y_i = j\}}$$

döntésfüggvény, ahol $A_n(x)$ a $\mathcal{P}_n$ azon cellája, amelybe x beleesik.
A **magfüggvényes szabály** a

$$g_n(x) = \arg\max_{j\in\{1,2,\dots M\}} \sum_{i=1}^{n} K_{h_n}(X_i - x) I_{\{Y_i=j\}}$$

döntésfüggvény.

Az alakfelismerési módszerek másik nagy csoportja, az **empirikus hibaminimalizálás**on alapuló módszerek. Egy g döntésfüggvény empirikus hibája az

$$L_n(g) = \frac{1}{n}\sum_{i=1}^{n} I_{\{g(X_i)\neq Y_i\}}$$

mennyiség. Az empirikus hibaminimalizálás módszere a döntésfüggvények egy $\mathcal{F}_n$ családja esetén a

$$g_n(x) = \arg\min_{g\in\mathcal{F}_n} L_n(g)$$

döntésfüggvényt választja (Devroye, Györfi and Lugosi (1996)).

Visszatérve az univerzális predikció problémájához $0-1$ költségfüggvény esetén megjegyezzük, hogy itt más a kombinálás, mégpedig egy randomizálás a szakértők felett:

2. lemma: *(Cesa-Bianchi (1997)). Jelölje $h_1,\dots,h_K$ predikcióknak egy véges családját. Legyen*

$$w_{n,k} = e^{-(n-1)L_{n-1}(h_k)}$$

és

$$v_{n,k} = \frac{w_{n,k}}{\sum_{i=1}^{K} w_{n,i}}.$$

Ha a kombinált, randomizált g predikciót úgy definiáljuk, hogy rögzített n esetén a $v_{n,k}$ eloszlás szerint választunk egy U_n véletlen számot, és

$$g(y_1^{n-1}, U_n) = h_{U_n}(y_1^{n-1}),$$

akkor

$$\mathbf{E}L_n(g) \le \min_k L_n(h_k) + \sqrt{\frac{\ln K}{2n}}.$$

Ezzel a kombinálási technikával oldható meg az alakfelismerés feladata függő megfigyelések esetén (Györfi, Lugosi and Morvai (1999)).

4 Portfolioválasztás: logaritmikus költség

A tőzsdei portfolioválasztás is egy előrejelzési probléma, ahol a részvények értékének változását akarjuk megjósolni. Tegyük fel, hogy d darab részvény van, és a tőkénket minden nap elején szabadon újraoszthatjuk a részvények között.

Egy nap végén az egyes részvények eredményességét az $x = (x^{(1)}, \dots x^{(d)})$ hozamvektor adja meg, amelynek j-edik komponense $x^{(j)}$ a j-edik részvénybe fektetett egységnyi tőke értéke a nap végén. A befektetésünket a $b = (b^{(1)}, \dots b^{(d)})$ portfoliovektorral adjuk meg, amelynek j-edik komponense $b^{(j)}$ azt mondja meg, hogy a j-edik részvénybe tőkénk hanyadrészét fektetjük be. Ekkor S_0 kezdő tőke esetén a tőke egy nap múlva

$$S_1 = S_0 \sum_{j=1}^{d} b^{(j)} x^{(j)} = S_0(b, x),$$

ahol $(\cdot, \cdot)$ a skalárszorzatot jelöli.

Hosszú idejű befektetések esetén jelölje az x_i vektor az i-edik nap végén a hozamvektort, azaz az x_i vektor j-edik komponense azt mondja meg, hogy a j-edik részvénybe fektetett egységnyi tőke mennyit ér az i-edik nap végén. Legyen kezdetben S_0 tőkénk. Első nap a portfolionkat egy b_1 vektorral adjuk meg, ahol b_1 komponensei nemnegatívak és az összegük 1. Az első nap végén a tőkénk

$$S_1 = S_0 \cdot (b_1, x_1).$$

S_1-et a második nap elején újból befektetjük egy b_2 portfolio szerint, ahol b_2 már függhet x_1-től. Ekkor

$$S_2 = S_0 \cdot (b_1, x_1) \cdot (b_2(x_1), x_2).$$

Ezt folytatva tőkénk az n-edik nap végén

$$S_n = S_0 \prod_{i=1}^{n} (b_i(x_1^{i-1}), x_i) = S_0 e^{\sum_{i=1}^{n} \ln(b_i(x_1^{i-1}), x_i)} = S_0 e^{n L_n(b^*)}.$$

Tehát a $b^* = \{b(x_1^{i-1})\}$ portfolio (prediktor) esetén annál jobban gyarapodik a pénzünk, minél nagyobb az $L_n(b^*) = \frac{1}{n} \sum_{i=1}^{n} \ln(b_i(x_1^{i-1}), x_i)$ átlagos kamatszint. Ezzel indokolható a logaritmikus költségfüggvény, bár ebben az esetben a költségfüggvény elnevezés nem annyira szerencsés, hiszen most $L_n(b^*)$-t nyilván maximalizálni akarjuk (Cover (1991), Cover and Thomas (1991)).

Az előző szakaszok alapján a szakértők kombinálása itt igen egyszerűen levezethető. Jelölje $h_1, h_2, \dots$ portfolióknak egy családját (szakértőket) és

$\{q_k\}$ egy valószínűségeloszlást. Legyen

$$w_{n,k} = q_k e^{(n-1)L_{n-1}(h_k)}$$

és

$$v_{n,k} = \frac{w_{n,k}}{\sum_{i=1}^{\infty} w_{n,i}}.$$

A kombinált g portfoliót úgy definiáljuk, hogy

$$g_n(x_1^{n-1}) = \sum_{k=1}^{\infty} v_{n,k} h_k(x_1^{n-1}).$$

Egyszerűen belátható, hogy ez a kombinálás egy elbonyolított formája annak az eljárásnak, amikor az elején az S_0 kezdő tőkémet a $\{q_k\}$ eloszlás szerint szétosztom a szakértők között, és "hagyom őket dolgozni", azaz az n-edik időpontban a pénzem összesen:

$$\begin{aligned}
S_n(g) &= S_0 e^{nL_n(g)} \\
&= S_0 e^{\sum_{i=1}^{n} \ln(g(x_1^{i-1}),x_i)} \\
&= S_0 \prod_{i=1}^{n} (g(x_1^{i-1}), x_i) \\
&= S_0 \prod_{i=1}^{n} \frac{\sum_{k=1}^{\infty} w_{i,k}(h_k(x_1^{i-1}), x_i)}{\sum_{k=1}^{\infty} w_{i,k}} \\
&= S_0 \prod_{i=1}^{n} \frac{\sum_{k=1}^{\infty} q_k e^{(i-1)L_{i-1}(h_k)}(h_k(x_1^{i-1}), x_i)}{\sum_{k=1}^{\infty} q_k e^{(i-1)L_{i-1}(h_k)}} \\
&= S_0 \prod_{i=1}^{n} \frac{\sum_{k=1}^{\infty} q_k e^{iL_i(h_k)}}{\sum_{k=1}^{\infty} q_k e^{(i-1)L_{i-1}(h_k)}} \\
&= S_0 \sum_{k=1}^{\infty} q_k e^{nL_n(h_k)} \\
&= \sum_k q_k S_n(h_k),
\end{aligned}$$

ahol $S_n(h_k)$ a h_k portfolio értéke S_0 kezdő tőke esetén (Cover and Ordentlich (1996), Opper and Haussler (1997)).

Ebből következik egy alsó becslés a kombinált portfolio kamatszintjére, ugyanis $S_0 = 1$ esetén

$$L_n(g) = \frac{1}{n} \ln S_n(g)$$

$$
\begin{aligned}
&= \frac{1}{n} \ln \left(\sum_k q_k S_n(h_k) \right) \\
&\geq \frac{1}{n} \ln \left(\max_k q_k S_n(h_k) \right) \\
&= \frac{1}{n} \max_k \left(\ln q_k + \ln S_n(h_k) \right) \\
&= \max_k \left(L_n(h_k) + \frac{\ln q_k}{n} \right).
\end{aligned}
$$

Ha véges sok K portfolionk van, és a q_k eloszlás egyenletes, akkor ez a korlát

$$L_n(g) \geq \max_k L_n(h_k) - \frac{\ln K}{n}.$$

Ha $K = 20$ szakértőt dolgoztatunk egy évig ($n = 365$), akkor kamatszinten legfeljebb 0.008 a veszteségünk a legjobb kamatszinthez képest.

Stacionárius és ergodikus tőzsdei folyamat esetén Algoet (1992) adott aszimptotikusan optimális empirikus portfoliostratégiát.

Irodalom

[1] Algoet, P. (1992). Universal schemes for prediction, gambling, and portfolio selection. *Annals of Probability*, 20:901–941.

[2] Cesa-Bianchi, N. (1997). Analysis of two gradient-based algorithms for on-line regression. in *Proc. 10th Ann. Conf. Computational Learning Theory*, New York ACM Press, pp. 163-170.

[3] Cesa-Bianchi, N., Freund, Y.,Helmbold, D. P., Haussler, D., Schapire, R. and Warmuth, M. K. (1997). How to use expert advice. *Journal of the ACM*, 44(3):427–485.

[4] Cover, T. (1991). Universal Portfolios. *Mathematical Finance*, 1, 1–29.

[5] Cover, T. and Ordentlich, E. (1996). Universal Portfolios with Side Information. *IEEE Transactions on Information Theory*, 42(2), 348–363.

[6] Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley and Sons.

[7] Devroye, L. and Györfi, L. (1983). Distribution-free exponential upper bound on the L_1 error of partitioning estimates of a regression function.

In *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, Konecny F., Mogyoródi, J. and Wetz, W. Eds., pp. 67-76, Budapest, Akadémiai Kiadó.

[8] Devroye, L., Györfi, L., Krzyżak, A. and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22, pp. 1371–1385.

[9] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer Verlag.

[10] Devroye, L. and Krzyżak, A.(1989). An equivalence theorem for L1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, 23, pp. 71–82.

[11] Devroye, L. and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, 8, pp. 231–239.

[12] Györfi, L. (1991). Universal consistencies of regression estimate for unbounded regression functions. In *Nonparametric functional estimation and related topics, ed. G. Roussas,* pp. 329–338. Dordrecht: Kluwer Academic Publishers.

[13] Györfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer Verlag.

[14] Györfi, L. and Lugosi, G. (2001) "Strategies for sequential prediction of stationary time series", in *Modeling Uncertainty: An Examination of its Theory, Methods and Applications*, M. Dror, P. L'Ecuyer, F. Szidarovszky (Eds.), Kluwer Academic Publisher.

[15] Györfi, L., Lugosi, G. and Morvai, G. (1999). A simple randomized algorithm for consistent sequential prediction of ergodic time series. *IEEE Transactions on Information Theory*, 45:2642–2650.

[16] Haussler, D., Kivinen, J. and Warmuth, M. (1998). Sequential Prediction of Individual Sequences Under General Loss Functions. *IEEE Transactions on Information Theory*, 44, 1906–1925.

[17] Kivinen, J. and Warmuth, M. K.. (1999). Averaging expert predictions. In H. U. Simon P. Fischer, editor, *Computational Learning Theory: Proceedings of the Fourth European Conference, EuroCOLT'99*, pages 153–167. Springer, Berlin. Lecture Notes in Artificial Intelligence 1572.

[18] Littlestone, N. and Warmuth, M. K. (1994). The weighted majority algorithm. *Information and Computation*, 108:212–261.

[19] Merhav, N. and Feder, M. (1998). Universal prediction. *IEEE Transactions on Information Theory*, 44:2124–2147.

[20] Morvai, G., Yakowitz, S. and Algoet, P. (1997). Weakly convergent stationary time series. *IEEE Transactions on Information Theory*, 43:483–498.

[21] Morvai, G., Yakowitz, S. and Györfi, L. (1996). Nonparametric inference for ergodic, stationary time series. *Annals of Statistics*, 24:370–379.

[22] Opper, M. and Haussler, D. (1997). Worst Case Prediction over Sequences under Log Loss. In: *The Mathematics of Information Coding, Extraction, and Distribution*. Springer Verlag.

[23] Spiegelman, C. and Sacks, J. (1980). Consistent window estimation in nonparametric regression. *Annals of Statistics*, 8, pp. 240–246.

[24] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, 5, pp. 595–645.

[25] Vovk, V. G. (1990). Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 372–383. Association of Computing Machinery, New York.

[26] Weinberger, M., Merhav, N. and Feder, M. (1994). Optimal Sequential Probability Assignment for Individual Sequences. *IEEE Transactions on Information Theory*, 40, 384–396.