Az információelmélet alapfogalmairól

DR. KERPÁN ISTVÁN
KKVMF Híradásipari Intézet

C. E. Shannon alapvetése óta az információelmélet matematikai formája kiforrott, helyének, szerepének a megítélése letisztult. A legelterjedtebb felfogásban az *információ* értelmezését biológiai, társadalmi és technikai rendszerek (nevezzük ezeket gyűjtőnéven „kibernetikai" rendszereknek) vezérlési folyamataihoz kapcsolják. Az *információelmélet* az üzenetforrások és átviteli csatornák statisztikai struktúrájának a leírására alkalmas diszciplína, amelyet a valószínűségelmélet új ágának is tekintenek. Alkalmazási területei közül (mind ez idáig) a kódolással összefüggő problémákat emelik ki. Mindamellett az információelmélet konkrét története fenntartásokat is indukált az elméletnek (mindenekelőtt központi fogalmának, az entrópiának) az értelmezésével kapcsolatban.

Emlékeztetőül és jelöléseink bevezetésére: ha $x_1, x_2, \ldots, x_N$ egy X üzenetforrás lehetséges üzenetei, $p_1, p_2, \ldots, p_N$ pedig az egyező indexű üzenetek valószínűségei, akkor a

$$H(X)=\sum_{i=1}^{N} p_i \log_2 \frac{1}{p_i} \qquad \text{(i)}$$

kifejezés adja meg a forrás — shannon mértékegységben kifejezett — entrópiáját. Az (1)-ben definiált $H(X)$ az

$$I_i=\log_2 \frac{1}{p_i} \qquad (2)$$

mennyiségnek a p_i valószínűségekkel súlyozott átlagával egyenlő. Szokás (2) alapján I_i-t az x_i üzenetre, (1) alapján pedig, $H(X)$-et a teljes forrásra vonatkozó „bizonytalanság" mértékének is tekinteni.

Példa az entrópiával kapcsolatos fenntartásokra: „Az entrópia antropomorf fogalom." (Idézet [6]-ból.) „... a bizonytalanság mértékéül szolgáló I információmennyiség egyértelmű meghatározásához nem tudunk elegendő ‚természetes' szempontot találni..." (L. [5], 110—111. o.)

Ismeretes, hogy a (shannon egységekben kifejezett) $H(X)$ — statisztikusan stacionárius forrás esetében — számszerűen megegyezik a kibocsátott üzenetek kódolásához felhasználandó bináris jelölők (bitek) átlagának az (alkalmas kódolással tetszés szerint megközelíthető, kedvező esetben pedig elérhető) minimumával: Ha x_i üzenet kódolásához n_i bitet használunk, akkor a kibocsátott üzenetek átlagos hossza (n_i-k várható értéke):

$$\bar{n}=\sum_{i=1}^{N} p_i n_i, \qquad (3)$$

és — kedvező esetben —

$$H(X)[\text{shannon}]=(\bar{n}/_{\min})[\text{bit}]. \qquad (4)$$

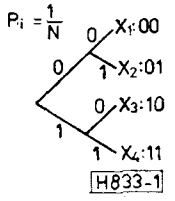
(4)-et — $H(X)$ egyik *értelmezéseként* már [1]-ben megtaláljuk. $(\bar{n})_{\min}$ keresése szuggesztíven *előkészíti* $H(X)$ bevezetését és a fenti értelmezést pl. [3]-ban. Azonban — véleményünk szerint — $(\bar{n})_{\min}$ keresése $H(X)$ (1) kifejezésének *előállítására* is elegendő „természetes" szempontot szolgáltathat („kibernetikai" rendszerek tekintetében). Ezt szeretnénk példázni az alábbiakban.

A jelekkel takarékoskodó, ezért szóelválasztójel nélkül is egyértelmű dekódolást lehetővé tevő bináris kódot kívánunk szerkeszteni az X forrás üzeneteire. Ilyen tulajdonságú kódhoz vezet, ha az üzeneteket egy bináris fa csúcsaival jellemezzük, s a hozzájuk vezető élek (0 vagy 1) szimbólumainak a sorozatával kódoljuk, az ábrák szerint. (A megfontolások lényege nem bináris kód abc-re is alkalmazható. Az így szerkesztett kód egyetlen szavának sem létezik értelmes meghosszabbítása vagy megrövidítése („prefix tulajdonság"). Ez teszi lehetővé a szóelválasztók nélküli egyértelmű dekódolást.

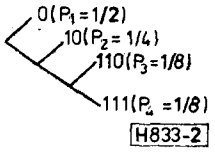
Bináris jelölőink gazdaságos kihasználása szükségessé teszi, hogy minden újabb jelölő pozícióval (a lefektetve ábrázolt fák minden emeletével) az eredeti X halmazt, majd annak részhalmazait mindig két részre osszuk. Az üzenetek alcsoportok közötti szétosztásának a módjától függ a fa maximális emeletszáma és a kódszavak hossza (vö. az *1.* és *2. ábrákat!*).

Ha minden üzenet egyforma valószínűségű $\left(p_1=\frac{1}{N}=\text{konstans}\right)$, a legrövidebb *átlagos* szóhosszúságot nyilvánvalóan akkor kapjuk, ha az osztályozási lépések során az üzenetek számát, s ezzel egyben az üzenetvalószínűségek összegét is felezzük. Az üzenetszám felezését akkor tudjuk következetesen keresztülvinni, ha N 2-nek egész kitevőjű hatványa. Ekkor $n_i=\bar{n}=\log_2 N$.

Előadásként elhangzott a KKVMF VII. tudományos ülésszakán



1. ábra



2. ábra

Ha az üzenetek nem egyforma valószínűségűek, választanunk kell: az üzenetek számát vagy az üzenetek valószínűségeinek az összegét felezzük? Utóbbi mellett kell döntenünk akkor, ha a kibocsátott üzenetek átlagos hosszát kívánjuk minimalizálni. Ehhez ugyanis a nagyobb valószínűségű, azaz a gyakoribb üzeneteket célszerű rövidebben kódolni. Ha $p_1 > p_3+p_4$, akkor a 2. ábra megoldása kisebb átlagos jelölőhosszhoz vezet, mint az 1. ábráé. A bináris jelölők leggazdaságosabb kihasználása ez esetben akkor lehetséges, ha az üzenetvalószínűségek olyan eloszlást mutatnak, hogy összegük mindaddig felezhető, ameddig az egyes üzenetekhez eljutunk. Ez azt jelenti, hogy az $\frac{1}{p}$ mennyiségek bármelyik p_i-re kettőnek valamelyik pozitív egész kitevőjű hatványai (Emellett természetesen $\sum_{i=1}^{N} p_i = 1$ is teljesül.) (További példa: 3. ábra.)

A valószínűség-összegek felezgetésével nyert kód a *Shannon—Fano-féle kód*. Ez optimális esetben (ha az üzenetvalószínűségek eloszlására tett kikötések teljesülnek)

$$n_i = \log_2 \frac{1}{p_i} \tag{5}$$

szóhosszúságokat eredményez. Ennek átlaga, $(\bar{n})_{\min}$ igy éppen az (1)-ben definiált $H(X)$.

Az említett esetben a Shannon—Fano kódolással minden üzenetet olyan szóhosszal kódolunk, mintha egy $\frac{1}{p_i} = N_i$ számosságú, egyenletes valószínűség eloszlású üzenetforráshoz tartozna, amelynél

$$n_i = \log_2 N_i = \log_2 \frac{1}{p_i}.$$

Ha az optimális eset feltételei nem teljesülnek, akkor (egy lehetséges megoldásként) fogjuk össze és kódoljuk az X forrás M egymást követő üzenetéből álló *blokkokat*. Ha M elegendően nagy, akkor minden blokkban p_1M darab x_1 üzenet, p_2M darab x_2 üzenet, ... p_NM darab x_N üzenet lesz (az esetek elhanyagolhatóan kicsiny hányadától eltekintve). Egy ilyen „tipikus" blokk előfordulási valószínűsége (ha az üzenetek egymástól függetlenek):

$$q = p_1^{p_1M} \cdot p^{p_2M} \ldots p_N^{p_NM}. \tag{6}$$

Mindegyik tipikus blokk egyforma valószínűségű, s valószínűségeik összege (elhanyagolhatóan kicsi különbséggel): $\Sigma q = Kq = 1$. Így az egyforma valószínűségű tipikus üzenetek száma: $K = \frac{1}{q}$. Az $\frac{1}{q}$ darab, egyforma valószínűségű üzenet kódolásához optimális esetben (ha $\frac{1}{q}\sqrt{2}$-nek egész kitevőjű hatványa) $\log_2 \frac{1}{q}$, egyébként pedig legfeljebb $1+\log_2 \frac{1}{q}$ bináris jelölő szükséges. Így az M üzenetet tartalmazó blokk kódolásához átlagosan szükséges bináris jelölőszám

$$\log_2 \frac{1}{p_1^{p_1M} \cdot p_2^{p_2M} \ldots p_N^{p_NM}} =$$

$$= M \sum_{i=1}^{N} p_i \log_2 \frac{1}{p_i} = M\, H(X)$$

és $1+MH(X)$ között van. Az egy-egy x_i üzenetre jutó átlagos jelölőszám pedig ennek M-ed részére:

$$H(X) \leqq \bar{n} < H(X) + \frac{1}{M}. \tag{7}$$

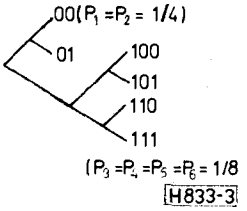
Ha M elég nagy és $\frac{1}{M} \ll H(X)$, akkor egyszerűen [egyezően [4]-gyel]

$$\bar{n} = (\bar{n})_{\min} = H(X).$$

Ha egyetlen kikötésünk az egyértelmű, megbízható kódoláshoz-dekodóláshoz szükséges átlagos jelölőszám minimalizálása és (mint az eddigiekben is) eltekintünk a zajtól, akkor fentiek elegendő természetes szempontot adhatnak arra, hogy $H(X)$-et $(\bar{n})_{\min}$ mértékeként bevezessük és technikai rendszerek tervezésében figyelembe vegyük.

A fenti feltételekkel megadhatók olyan modellek is (biológiai analógiákra), amelyekre a (4) összefüggést kielégítő kód spontán folyamatok nyomán is realizálódhat: önreprodukáló rendszerek ingerspecifikus reakciói tartalmazzák a megkülönböztetés és azonosítás (a dekódolás) mozzanatát. Feltételezhetjük, hogy az ilyen rendszerek véletlen mutánsai közül azoknak a fennmaradási esélye lehet nagyobb, amelyek többféle üzenetsorozatot képesek megkülönböztetni, viszonylag kevesebb jelölő felhasználásával. Ezen az alapon megkezdődhet a kiválogatódás és az optimálás kódhoz tartó evolúció. (Az élő minták információelméleti-kibernetikai hatását Neumann és Wiener munkássága példázhatja.)

Más helyen (l. [8]) kísérletet tettünk az információelmélet további fogalmainak (a zajos csatorna vesztesége és kapacitása) tisztán kódolási megfontolásokon és egy reális (elemeiben Shannontól származó) hírközlési modellen alapuló bevezetésére. Megjegyezzük: hasonló modell zárt szabályozási körök alapján



3. ábra

is értelmezhető. A problémák ilyenfajta megközelítése (az axiomatikus és egzakt matematikai tárgyalás *kiegészítéseként*) talán elméletileg sem érdektelen (pl. a bevezetőben említett fenntartások eloszlatásában). Gyakorlati haszna pedig az lehet, hogy segíti az elmélet eredményeit szélesebb körben munkaeszközzé válni. Mindkét vonatkozásban látunk még tennivalót.

IRODALOM

[1] *C. E. Shannon—W. Weaver:* The Mathematical Theory Of Communication. THE UNIVERSITY OF ILLINOIS PRESS, URBANA, 1949.

[2] *A. M. Jaglom—J. M. Jaglom—A. Ja. Hincsin:* Az információelmélet matematikai alapjai. Műszaki Könyvkiadó, Budapest 1959.

[3] *Rényi Alfréd:* Napló az információelméletről. Gondolat Kiadó, Budapest, 1976.

[4] *Fazlollah M. Reza:* Bevezetés az információelméletbe. Műszaki Könyvkiadó, Budapest, 1966.

[5] *Fényes Imre:* Entrópia. Gondolat Kiadó, 1962.

[6] Információ és entrópia. Vitaismertetés, Természet Világa, 1978. 6. sz.

[7] *R. A. Bones:* Dictionary of Telecommunications. London, NEWNES — BUTTERWORTHS 1970.

[8] *N. Wiener:* Válogatott tanulmányok. Gondolat Kiadó, 1974.

[9] *Neumann János:* A számológép és az agy. Gondolat Kiadó, 1972.

[10] *Kerpán István:* A hírközlő csatorna kapacitása. Híradástechnika, 1982. 6. sz.