

Távközlési Kutató Intézet

Az orvosdiagnosztika segítése tanuló algoritmusokkal

Györfi László és Csibi Sándor

ÁTTEKINTÉS

Intézetünkben 1969 óta foglalkozunk tanuló-felismerő problémákkal. Ez egyrészt matematikai kutatást jelent, másrészt keressük azokat az alkalmazási területeket, ahol ezek a módszerek hatékonyan használhatók. Ezek közül az alkalmazási lehetőségek közül számunkra az egyik legfontosabb az orvosdiagnosztika.

Az orvos a diagnózis megállapítása során alternatív lehetőségek közül választ, dönt arról, hogy adott alternatív hipotézisek közül melyik teljesül. Egy diagnózis megalkotása bizonyos elemi döntések sorozata. Ha számítógépes tanulásról esik szó, akkor triviálisan adódik az az ötlet, hogy használjuk arra a számítógépet, hogy tároljuk az orvosi irodalomban az arra az esetre vonatkozó döntési szabályt, és így automatizáljunk egy ilyen elemi döntést. Természetesen, osztályozási szabályok ilyenfajta közvetlen gépre másolása esetén is felmerül a kérdés, hogy a gépre vitt szabályt klinikai esetek tömegén ne csak ellenőrizzük, hanem tovább is fejlesszük.

Az alakzátfelismerési módszerek - ezeken belül pedig a tanuló algoritmusok - ahhoz nyújtanak segítséget, hogy az összegyűjtött tapasztalati anyagnak - esetünkben például kórlapok sorozatának - a felhasználásával konstruáljunk egy ilyen elemi döntést. Ettől az elemi döntéstől azt várjuk el, hogy a döntéshez tartozó tévedés valószínűsége a lehető legkisebb legyen. A tanuló algoritmusok egyszerűen alkalmazható módszerek arra, hogy az empirikus adatokat feldolgozzuk úgy, hogy a feldolgozás eredménye döntési szabály legyen.

Matematikailag ez azt jelenti, hogy egy  $x$  valószínűségi vektorváltozó (méréssorozat) esetén döntünk arról, hogy  $H_+$  vagy  $H_-$  alternatív hipotézis következett-e be, és a döntés valószínűsége minimális legyen. Ez a döntés - ami a Bayes-döntés speciális esete - a következő: tegyük fel, hogy  $x$  minden lehetséges értéke esetén ismerjük a  $H_+$  és  $H_-$  bekövetkezésének a valószínűségét. Ha a  $H_+$ -hoz tartozó a nagyobb, akkor a  $H_+$  bekövetkezésére szavazunk, egyébként  $H_-$ -ra. Ezeket a feltételes valószínűségeket azonban gyakorlati feladatoknál nem ismerjük, viszont rendelkezésre áll egy minta,

az ugynevezett tananyag. Legyen  $(x_1, p_1), (x_2, p_2) \dots$  a kórlapokról kigyűjtött sorozat, ahol  $x_n$  az  $n$ -edik kórlapon lévő mérések vektora,  $p_n$  pedig az  $n$ -es tanítás, amely  $+1$ , ha ekkor a  $H_+$  következtetett be, egyébként  $-1$ . A feladat két részből áll:

1. A minta segítségével konstruálni egy döntést (ezt a szakaszt nevezik tanulásnak).

2. A kapott döntést alkalmazni.

Ennek megfelelően a programok is két részből állnak:

1. Tanulás

2. Osztályozás vagy ellenőrzés

Ellenőrzésnél ismert diagnózissal bíró eseteken vizsgáljuk a kapott döntési módszer jószágát, becsüljük a tévedés valószínűségét, álnegatív, álpozitív döntés valószínűségét.

### Algoritmusok

Röviden áttekintjük azokat a módszereket, melyek egyrészt irodalomból dolgoztunk fel, másik része intézetben belüli eredmény.

1. Nearest Neighbour algoritmus. Ez a döntési szabály igen szemléletes. Egy alkalmazását mutatta be Gulyás Ottó egy EKG feladaton. Ha rendelkezésre áll egy  $(x_1, p_1) \dots (x_n, p_n)$  minta és  $x$  mérés esetén dönteni akarunk, akkor megkeressük az  $x_n$  pontot  $x_1 \dots x_n$ -ből, mely  $x$ -hez legközelebb van. (Ezután az  $x$ -hez tartozó döntés eredménye a legközelebbi társ  $(x'_n)$  tartozó tanítás.) Ennek egy speciális esete az, amikor  $x$  mérés már szerepel valamelyik kórlapon, vagyis a legközelebbi társ vele azonos, ekkor nyilván senkit sem kell különösebben meggyőzni arról, hogy elfogadja ezt a döntést. Cover és Hart mutatta meg, hogy ez a módszer szoros kapcsolatban áll a Bayes döntéssel. A módszer hátránya, hogy számítógépes realizálás esetén igen memória- és időigényes.

2. Bizonyos speciális feladatokkal közel állunk ahhoz, hogy a legjobb döntés hibavalószínűsége  $0$ , vagyis az elméletileg lehetséges legjobb döntés tökéletes is. Ezeknek a feladatoknak megoldására ajánlhatók a potenciálfüggvényes algoritmusok ilyen - vagy ezekhez közelálló - kidolgozott változatai. A korrekció megválasztására az elmélet megalapozói ilyenfajta célokra 4 lehetőséget tárgyalnak. A programban a felhasználó - a feladat adottságainak és pontosabb céljainak megfelelően - 9 korrekció közül választhat.

3. Harmadikként említjük azokat a módszereket, amelyek nullától erősen különböző hibavalószínűség mellett is jól alkalmaz-

hatók. Itt lényegében arról van szó, hogy becsüljük a hipotézisekhez tartozó a posteriori valószínűségek különbségét - illetve az ezektől le származtatott döntéshüggvényeket - és ezek segítségével realizáljuk a Bayes döntést. E módszerek egyik klasszikus példája a Parzen-féle sűrűségfüggvény-beclés. Ez a módszer ugyan nagy minta esetén a legjobb döntést eredményezné, de a kiszámítandó függvényértékek száma a minta nagyságával lineárisan növekszik. Ezért alkalmazzuk többször azt a módszert, mely sűrűségfüggvények egy rögzített tagszámú sorfejtését becsüli. E téren is - az ismert eljárásokon túl - hatékony új eljárásokkal is rendelkezünk. Ennek alkalmazásáról beszéltünk a fertilitás prognózisával kapcsolatban.

4. A tanulási feladatok egy részénél alkalmazható az a feltevés, hogy a két feltételes eloszlás normális. Ennek a klasszikus feladatnak a megoldását is programban realizáltuk.

5. Az eddig említett módszereken kívül még a diszkriminációanalízis programját készítettük el.

A programokat CII-10010-es számítógépre írtuk meg ALGOL nyelven. E programok az OKI-TKI EKG mintarendszer alkalmazási programrendszerek részét képezik.

Nem tudjuk sikerült-e érzékeltetni azt, hogy a lehetséges módszereknek milyen széles köre áll rendelkezésre, hogy az orvos munkáját megkönnyítse, és az orvosi kutatás metodikáját kibővítsse.

A programok azonosítói:

POT 1

POT 2

NN

BAYES

GAUSS

PARZEN

CM

LK 4

I r o d a l o m

- T. Cover Hart "The nearest neighbour pattern classification"  
IEEE IT-13. 1967.
- S. Csibi "Simple and compound processes in iterative machine  
learning" CISM Summer Course, Udine, Italy, 1972.
- S. Csibi "Stability and complexity of learning algorithms"  
CISM Summer Course, Udine, Italy, 1972.
- O. Gulyás "On extended potential function type learning algorithms  
and their convergence rate" Problems of Control and Informa-  
tion Theory, Vol. 1 No 1 1972.
- L. Györfi "Potenciálfüggvényes algoritmusok" Egyetemi doktori érte-  
kezés, 1972.
- L. Györfi "Estimation of probability density and optimal decision  
function in RKHS". Proceedings of European Meeting of  
Statisticians, Budapest, 1972. (megjelenés alatt)
- E. Parzen "On estimation of probability density function and mode"  
Ann. Math. Stat. vol. 33. Sept. 1962.