POUŽITIE NIEKTORÝCH MATEMATICKÝCH METÓD V MEDICÍNE

EUGEN RUŽICKÝ, Bratislava

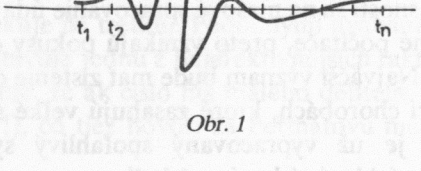
Jedným z kľúčových problémov medicíny je stanoviť diagnózu čo najpresnejšie, najspoľahlivejšie a súčasne čo najjednoduchšie. Ak sa zamyslíme nad tým, čo vlastne znamená stanoviť diagnózu, zistíme, že nejde o nič iné ako o kvantitatívne a kvalitatívne vyhodnotenie hromadných údajov. Tento proces u skúsených lekárov prebieha čiastočne podvedome. V súčasnosti sa na masové spracovanie údajov čoraz častejšie používajú samočinné počítače, preto vznikajú pokusy o automatizované určovanie diagnóz. Najväčší význam bude mať zistenie diagnózy použitím počítačov najmä pri chorobách, ktoré zasahujú veľké skupiny obyvateľstva a pre ktoré je už vypracovaný spoľahlivý systém príznakov. S prihliadnutím na rýchlosť získania výsledkov a na možnosť masového nasadenia možno očakávať dobré výsledky pri odhaľovaní chorôb v štádiách ich zárodku, čo má často rozhodujúci význam pre úspešnú liečbu.

Jednou z možných ciest pre stanovenie diagnózy použitím počítača je postupovať podobne, ako sa vyučuje stanovenie diagnózy na lekárskych fakultách. Študent absolvuje prax v nemocnici a usiluje sa odpozorovať postup skúsených lekárov pri určovaní diagnózy. Ako túto etapu procesu učenia možno spojiť s matematikou a s počítačmi? Ukážeme si to na nasledujúcom príklade. Predpokladajme, že lekár-špecialista vie presne určiť diagnózu choroby. Lekár dostane štatisticky vyváženú skupinu pacientov a rozdelí ju na dve skupiny: zdravých (Z) a chorých (CH). Výsledky svojej práce vloží nejakým spôsobom do počítača. Ako bude určovať diagnózu počítač?

Ak chceme riešiť tento problém, musíme sa na začiatku naučiť vkladať informácie do počítača, t. j. musíme vedieť zakódovať klinický stav pacienta. Najprv vyberieme tie príznaky (teplota, krvný tlak, laboratórne

vyšetrenia atď.), ktoré sú pri určení diagnózy podstatné; to je vlastne práca lekára. Príznaky charakterizujúce chorobu sú dvojaké — kvantitatívne a kvalitatívne. Prvé z nich sa dajú zakódovať v podstate jednoducho ako postupnosti čísel. Kódovať príznaky kvalitatívneho charakteru (napr. EKG, röntgenové vyšetrenie) je podstatne ťažšie. Vyžaduje si to vybudovať matematické teórie, prípadne vybrať zo známych teórií takú, ktorá je najvhodnejšia.

Napríklad pre krivku EKG sú dve možnosti kódovania. Pri prvej z nich zakódujeme funkčné hodnoty v konečnom počte predpísaných bodov. Namiesto krivky EKG budeme brať vektor $(f(t_1), \ldots, f(t_n))$.



Obr. 1

Druhý spôsob spočíva v rozložení funkcie na lineárnu kombináciu vopred zadaných funkcií:

$$f(x) = c_1 g_1(x) + , \ldots, + c_n g_n(x) + \varepsilon(x),$$

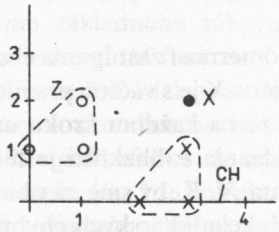
kde $g_1(x), \ldots, g_n(x)$ — sú základné funkcie získané empiricky a $\varepsilon(x)$ je dostatočne malá funkcia nepresnosti. Namiesto krivky EKG teraz budeme uvažovať vektor $(c_1, \ldots, c_n)$. Vektor vznikajúci pri druhej metóde má podstatne menšiu dimenziu ako vektor z prvej metódy.

Vráťme sa k podstate problému určovania diagnózy. Máme teda zadaný základný súbor zdravých (Z) a chorých (CH) pacientov. Vidíme, že každý lekársky príznak nemoci sa nahrádza postupnosťou čísel. Celý súbor lekárskych príznakov sa takto nahradí konečnou postupnosťou čísel, ktoré budeme nazývať matematickými príznakmi. Takže môžeme každého pacienta interpretovať ako bod v euklidovskom priestore.

Pre názornosť si ukážeme konkrétny príklad. Majme šiestich pacientov a dva príznaky, zapísaných v *tabuľke 1*.

1. pacient zdravý ..... $A(1) = (1, 2)$
2. pacient zdravý ..... $A(2) = (1, 1)$
3. pacient zdravý ..... $A(3) = (0, 1)$
4. pacient chorý ..... $A(4) = (3, 1)$
5. pacient chorý ..... $A(5) = (3, 0)$
6. pacient chorý ..... $A(6) = (2, 0)$

Každý pacient predstavuje bod v rovine, ako na *obr. 2*.



Obr. 2

Každého nového pacienta môžeme geometricky interpretovať ako bod v priestore príznakov a pýtať sa na jeho zaradenie do skupiny Z alebo CH. Pýtame sa, kam bude zaradený pacient X s príznakmi (3, 2)? Zdá sa, že je bližšie k skupine chorých pacientov ako k skupine zdravých. Túto intuíciu matematicky spresníme. Vzdialenosť bodov v rovine budeme počítať podľa známeho vzorca:

$$d(A, B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

Ďalej budeme postupovať podľa nasledujúceho algoritmu.

1. Spočítame vzdialenosti bodu X od každého bodu $A(i)$, $i = 1, \ldots, 6$ a usporiadame ich podľa veľkosti (*tab. 2*).

Tabuľka 2

$d(X, A)$	1	2	2	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{10}$
i	4	1	5	2	6	3

2. Určíme si pravidlo n-blízkosti. K bodu X vyberieme n najbližších bodov $A(i)$ a určíme pomernosť zastúpenia bodov zo skupiny CH. Pri rovnosti vzdialeností k n-tému bodu musíme pribrať ešte tie body, ktoré majú rovnakú vzdialenosť ako n-tý bod.

V našom prípade

Tabuľka 3

n	1	2	3	4	5	6
pomer	1	2/3	2/3	3/5	3/5	1/2

Z príkladu vidíme, že pomernosť zastúpenia chorých pacientov je väčšia ako zdravých, a preto pacient X je s väčšou pravdepodobnosťou chorý.

Je načase pripomenúť, že na každom kroku určovania diagnózy môže nastať omyl. Dôvod zavádzania n-blízkosti je ten, že zmenšuje možnosť mylnej klasifikácie pacienta X. Keby sme sa obmedzili len na 1-blízkosť, je nebezpečenstvo reprodukcie lekárskych chybných diagnóz. Na druhej strane aj n-blízkosť s narastajúcim n tiež stráca spoľahlivosť určovania diagnózy, lebo pri vyváženom základnom súbore (t. j. počet Z = počet CH) pre n blízke k počtu všetkých pacientov je pomernosť blízka 1/2. Ďalšou úlohou je preto vybrať optimálne n.

Tieto úvahy sa opierali o hypotézu, že dvaja pacienti s málo odlišnými príznakmi budú mať rovnakú diagnózu. Tento matematický model je všeobecne prijatý a do určitej miery simuluje postup určovania diagnózy lekárom.

Teraz uvedený diskrétny model n-blízkosti zovšeobecníme. Z príkladu vidíme, že algoritmus n-blízkosti nám pre každý bod roviny určil hodnotu miery výskytu chorých pacientov. Predpokladajme, že máme vopred zadanú funkciu p na základnom súbore bodov (pacientov) v priestore k. V našom príklade pre skupinu Z hodnotu 0 a pre skupinu CH hodnotu 1. Pýtame sa, ako túto funkciu aproximovať spojitou funkciou definovanou na priestore k.

V predchádzajúcom príklade sme pri určovaní hodnoty $p(X)$ brali v okolí bodu X všetky funkčné hodnoty bodov A s rovnakou váhou. To však nie je opodstatnené, pretože tie body, ktoré sú bližšie k skúmanému bodu, mali by mať väčšiu dôležitosť pri určení funkčnej hodnoty $p(X)$.

Preto budeme určovať diagnózu nasledovne:

$$p(x) = \frac{\sum_i f(\mathrm{d}(X, A(i)).p(A(i))}{\sum_i f(\mathrm{d}(X, A(i)))}$$

kde $p(A(i))$ sú vopred zadané a funkciu f volíme tak, aby bola klesajúca pri rastúcej vzdialenosti (napr. môže to byť funkcia $e^{-\mathrm{d}^2}$, $\frac{1}{1+\mathrm{d}^2}$). Otázka znie, do akej miery funkcia $p(X)$ aproximuje zadané hodnoty? Odpoveď môžeme získať testovaním základného súboru pacientov vzhľadom na funkciu f. Pritom sa snažíme vybrať funkciu f optimálne.

Teraz, keď sme zovšeobecnili model, nastoľujú sa nové otázky. Kedy budeme klasifikovať pacienta ako chorého alebo zdravého podľa aproximujúcej funkcie $p(x)$? Predpokladajme, že aproximujúca funkcia $(p(x)$ nadobúda hodnoty v intervale $\langle 0, 1\rangle$. Potrebujeme vybrať dve hodnoty p_1, p_2 z intervalu $(0, 1)$ a diagnózu určíme podľa tohto pravidla:

$$p(x) \in \begin{cases} \langle 0, p_1) & \text{— pacient je zdravý,} \\ \langle p_1, p_2\rangle & \text{— pacient je rizikový,} \\ (p_2, 1\rangle & \text{— pacient je chorý.} \end{cases}$$

Výber hodnôt p_1, p_2 nie je ľubovoľný, ale závisí od choroby a od základného súboru pacientov.

V praxi sa uskutočňuje testovanie základného súboru pacientov spolu s optimalizáciou na funkciu f. Výsledkom je potom určenie prahových hodnôt p_1, p_2 a tiež spoľahlivosť danej metódy.

Zatiaľ sme uvažovali, že všetky príznaky pri určovaní diagnózy majú rovnakú váhu. Zo skúseností lekárov však vieme, že nie všetky príznaky sú rovnako podstatné. Isté zlepšenie môžeme dosiahnuť zmenou metriky v priestore príznakov R. Túto zmenu uskutočníme vhodnou deformáciou v priestore tak, aby sa zvýraznil rozdiel medzi skupinou zdravých a chorých pacientov.

Ak testovaný základný súbor pacientov dáva pre určitú chorobu výsledky s veľkou presnosťou, potom sa môžeme pýtať, či množina Z a množina CH sa nedá oddeliť nadrovinou v priestore príznakov. Predpokladajme, že konvexné obaly množín Z a CH sú disjunktné. Geomet-

rická podstata vyhľadania takejto nadroviny spočíva v tom, že sa nájdu také body z konvexných obalov Z a CH, ktoré určujú vzdialenosť týchto množín. Pre tieto dva body sa zostrojí nadrovina, ktorá je kolmá na ich spojnicu a prechádza stredom týchto bodov. Programová realizácia sa uskutočňuje v postupnom hľadaní najbližších bodov na simplexoch z množín Z a CH. Analyticky nadrovinu môžeme zapísať v tvare

$$\boldsymbol{w}.x + w_0 = 0$$

kde $\boldsymbol{w}$ je vektor kolmý na nadrovinu a w_0 je absolútny člen. Pre oddeľujúcu nadrovinu musí platiť:

$$\boldsymbol{w}.x + w_0 < 0, \text{ ak bod } X \text{ je zo skupiny } Z$$

a

$$\boldsymbol{w}.x + w_0 > 0, \text{ ak bod } X \text{ je zo skupiny } CH$$

To nám uľahčuje situáciu pri určení diagnózy pre pacienta X, ktorého príznaky poznáme. Stačí zistiť, či hodnota $\boldsymbol{w}.x + w_0$ je z intervalu $(-\infty, -\varepsilon)$, $\langle -\varepsilon, \varepsilon \rangle$, $(\varepsilon, +\infty)$ a podľa toho klasifikovať, či je pacient zdravý, rizikový, resp. chorý.

Doposiaľ sme študovali prípad, že základný súbor pacientov bol rozdelený na dve skupiny Z a CH. Pri lekárskej praxi sa môže stať, že lekár nevie stanoviť diagnózu podľa nameraných príznakov. Dokonca aj v takomto prípade existujú matematické metódy, podľa ktorých sa dajú vytvoriť skupiny pacientov s podobným druhom ochorenia. Táto metóda môže slúžiť na orientáciu lekára pri určovaní nedostatočne prebádanej choroby.

Nech je daný základný súbor pacientov bez vyčlenených skupín Z a CH. Ak príznaky rozlišujú druh ochorenia, potom body zodpovedajúcich pacientov v priestore príznakov budú vytvárať zhluky. Zaujíma nás klasifikácia týchto zhlukov.

Opíšeme jeden jednoduchý algoritmus určovania zhlukov.

1. Majme zadanú množinu bodov $A(1), \ldots, A(n)$ v priestore. Zvolíme si prvý bod $A(1)$ za centrum z prvého zhluku. Vypočítame vzdialenosť druhého bodu $A(2)$ od centra z. Ak táto vzdialenosť prekročí vopred zadanú hranicu t, potom tento bod zvolíme za centrum z ďalšieho zhluku. Ak neprekročí hranicu t, potom pridáme bod $A(2)$ k zhluku z. Ďalej postupujeme iteratívne podľa bodu dva.

2. Máme prebrané body A(1), ..., $A(k)$ a vytvorené zhluky s centrami $z_1, \ldots, z_i$. Ak všetky vzdialenosti bodu $A(k+1)$ od centrov $z_1, \ldots, z_i$ sú väčšie ako hranica t, potom tento bod zvolíme za ďalšie centrum. V inom prípade bod $A(k+1)$ pridáme k tomu zhluku, ku ktorému je najbližšie vzhľadom na centrum.

Pri používaní metód rozpoznávania obrazcov treba mať stále na pamäti, že výsledok metódy závisí od presnosti a výberu štatistickej postupnosti základného súboru pacientov. Od matematickej metódy môžeme očakávať správne výsledky len vtedy, keď použijeme na začiatku správne údaje. Niektoré chyby v štatistickej postupnosti dokážeme objaviť jej analýzou, ale to len za predpokladu, že väčšina ostatných údajov je správna. Prednosťou algoritmov rozpoznávania obrazcov je nielen to, že dokážeme určiť diagnózu nového pacienta na základe skúseností, ktoré sme získali zo štatistickej postupnosti, ale dokážeme oceniť dôležitosť prípadne rozhodnúť o úplnosti meraných údajov.

Matematicko-fyzikálna fakulta má vypracované programy na rozpoznávanie obrazcov a zhlukovú analýzu. Tieto modely sa dajú využiť aj v iných oblastiach, ako napríklad pri rozpoznávaní symbolov, pri automatickom riadení atómových elektrární, pri spracovaní šekov v bankách a pri rozpoznávaní máp v kartografii.

Literatúra

[1] Nejmark, Ju. I.: Raspoznavanie obrazov i medicinskaja diagnostika, MOSKVA 1972.
[2] Gonzales, R. C.: Pattern recognition principles, LONDON 1974.