Faragó T., 1979:
Makroszinoptikus helyzetek statisztikus-algoritmikus felismerése.
Időjárás, 83:1, 20-26. o.

Makroszinoptikus helyzetek statisztikus-algoritmikus felismerése

Faragó Tibor

The Statistical- Algorithmical Recognition of Macrosynoptic Situations. The identification of various objects, especially the elements of a given system of macrosynoptic codes is a typical statistical task that may be solved by methods treated in the theory of the statistical pattern recognition. The concrete problem, which is in our case the recognition of the codes of the Hess- Brezowsky's system provided pressure fields upon a highly bounded area, may be so memory- and time-consuming that we ought to reduce even the “learning” phase of the classification algorithm. Such a procedure and its application to the above- mentioned macrosynoptic codes are dealt with in the article.

*

Статистическое-алгоритмическое распознавание макросиноптических ситуаций. Идентификация различных объектов, как например макросиноптических кодов является такой типичной статистической задачей, которая может быть решена методами теории статистического распознавания образов. Конкретная задача, в нашем случае задача распознавания кодов системы Гесс—Брезовски на основании полей давления сильно ограниченной территории, может нуждаться в таком большой объеме машинной памяти и вычислений, что уже фаза «обучения» алгоритма классификадии должна быть сокращена. Одна такая процедура и ее применение к вышеупомянутым кодам представлены в статье.

*

A rendszerezés a kutatás egyik alapvető szakasza. Különösen fontos ez a tevékenység az olyan, bonyolult jelenségekkel foglalkozó területen, mint amilyen a szinoptikus meteorológia. A különféle léptékű légköri folyamatok osztály - besorolási elvei mind közvetlenebb tapasztalati úton, mind pedig általánosabb számítógépes eljárások (algoritmusok) útján alkalmazhatók. A makroszinoptikus helyzetek rendszerezésének sorában tipikus az európai szinoptikai körzetre vonatkozó, Hess és Brezowsky (1952) által bevezetett kódsorozat. Példaként említhetünk egy Délkelet-Európára mértékadóbb osztályozást is (Topor, 1954). Jól ismert a Magyarország-orientált Péczely-féle rendszer is (Péczely, 1957). Többek között Vangengejm és Girsz (1970) dolgoztak ki az egész északi hemiszférára vonatkozó, már egyértelműen a hosszútávú előrejelzések céljait szolgáló cirkulációs kódokat. Utóbb a matematikai statisztika módszereit is alkalmazták objektívebb osztályok előállítására (Grúza és Ranykov, 1970; Jakovleva, 1970; Gulyás, 1977), illetve az analógia elvének meghatározására (Zverjev és Pegy, 1960; Bagrov, 1964; Pegy, 1970). Az ilyen osztályozások lehetséges táv- prognosztikai alkalmazásának egyes kérdéseivel is foglalkozott Kaba, Faragó és Gulyás (1975), valamint Kuba és Faragó (1976).

A különféleképpen származtatott makroszinoptikus kódok alkalmazása azonban nehézségekbe ütközhet, hiszen bizonyos helyzetek egyértelmű azonosítása nem mindig lehetséges. Következésképpen a felismerési algoritmusok is csak statisztikai jellegűek lehetnek. Az adott statisztikai feladat, a *Hess–Brezowsky*-féle kódok szerinti besorolás 29 osztály megkülönböztetését jelenti szélsőségesen egyszerűsített, de így is 35-dimenziós alakzatokkal, ami kizárja, hogy közvetlenebb – például sűrűségfüggvény-becslő, vagy más explicit diszkrimináló függvényeket felhasználó – módszert alkalmazzunk. Szinte az egyedüli, kézenfekvő kiutat az analógiás (angol terminológiával Nearest Neighbour, a továbbiakban NN-) eljárások kínálják.

Analógiás felismerő-osztályozási eljárások

Az analógiák keresésén alapuló eljárásoknak igen gazdag matematikai háttere van és meglehetősen elterjedtek a meteorológiai kutatásokban is. Statisztikai sajátosságaival, konvergenciájával, alkalmazhatóságának elveivel behatóan foglalkozott többek között *Cover* és *Hart* (1967), *Cover* (1968) és *Wagner* (1971). Az NN-eljárást a meteorológiában legtöbbször közép- és hosszútávú előrejelzések céljaira alkalmazzák. Az analógiakeresés ezen túlmenően alkalmas lehet szűkebb értelemben vett osztályozási problémák megoldására is. Egy ilyen típusú feladat a makroszinoptikus helyzetek felismerése.

Tekintsük át mindenekelőtt az eljárás elvi alapjait. Legyenek a kérdéses objektumok $\bar{x}$ számvektorok az n-dimenziós vektortérből és legyen adott egy mérőszám, mely bármelyik két ilyen objektum „közelségét" (hasonlóságát) méri: $\varrho(\bar{x}_1, \bar{x}_2)$. Következésképpen $\varrho\,(\bar{x}_1, \bar{x}_2) \geqq 0$, illetve $\varrho\,(\bar{x}, \bar{x}) = 0$. Ha most adva vannak a véletlentől függő objektumok (valószínűségi vektorváltozók): $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_N$ és mindegyikhez hozzárendelten valamilyen (valószínűségi) paraméter (kategória, osztály) $\vartheta_1, \vartheta_2, \ldots, \vartheta_N \in \Theta$, akkor az NN-eljárás értelmében egy ismeretlen paraméterű $\bar{x}$ megfigyelt vektorhoz annak az $\bar{x}_k$ objektumnak a ϑ_k paraméterét rendeljük hozzá, amelyik a legközelebb van $\bar{x}$-hez:

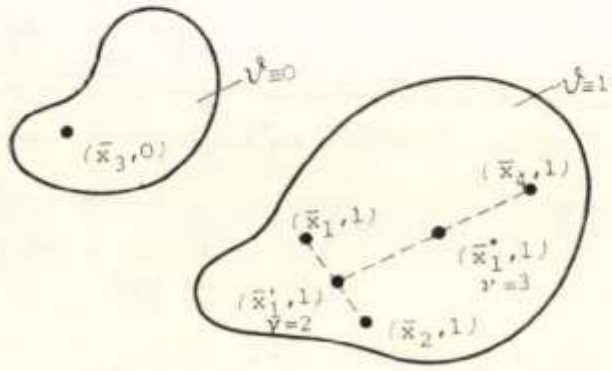
$$\varrho\,(\bar{x}, \bar{x}_k) \leqq \varrho\,(\bar{x}, \bar{x}_i), \quad i = 1, 2, \ldots, N.$$

Az analógiakeresés alapjául szolgáló $T_N = \{(\bar{x}_1, \vartheta_1), (\bar{x}_2, \vartheta_2), \ldots, (\bar{x}_N, \vartheta_N)\}$ tananyagot a legegyszerűbben egy megfigyelési sorozat elemeiből állíthatjuk elő. A feladat annál bonyolultabb, minél nagyobb a vektorok dimenziószáma (n) és minél több lehetséges eleme van a Θ paramétertérnek. A számításigény N növekedésével exponenciálisan nő, hiszen az $\bar{x}_k$ „legközelebbi társ" kikeresése érdekében miután kiszámítottuk az összes $\varrho\,(\bar{x}, \bar{x}_i)$, $i = 1, 2, \ldots, N$ mérőszámot, ezek közül ki kell keresni a legkisebbet. A feladat egyszerűsítése érdekében a közvetlenül megfigyelt sorozatot többféleképpen is próbálták úgy redukálni (*Hart*, 1968; *Gates*, 1972), hogy utólag, bizonyos teszteléseket követően elhagytak annak elemei közül. Ez a redukció – bár csak egyszer kell végrehajtani egy feladatban – igen számításigényes. Ennek kiküszöbölésére a következő módosítást vezetjük be a fentebb leírt klasszikus NN-eljáráshoz képest: a rendre megfigyelt $(\bar{x}_i, \vartheta_i)$ párok közül csak azokat vesszük fel a tananyagba, amelyeket az előző tananyag alapján az eljárás rosszul osztályozott. Tehát ebben a módosított NN-módszerben (MNN) az első lépésben $T_1 = \{(\bar{x}_1, \vartheta_1)\}$ illetve az $(i+1)$-ik lépésben (α) $T_{i+1} = T_i$, ha $\bar{x}_{i+1}$ legköze-

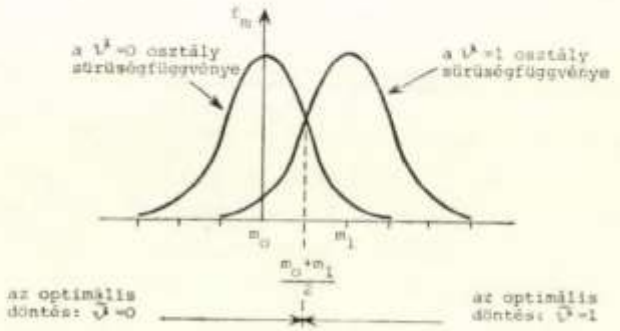
lebbi társa a T_i tananyagból $\bar{x}_k$ és $\vartheta_{i+1} = \vartheta_k$, máskülönben ($\beta$) a tananyag kibővül az új párral: $T_{i+1} = T_i \cup \{(x_{i+1}, \vartheta_{i+1})\}$. Megadjuk a módszer még egy változatát: a szelektív NN-módszert (SNN). Ekkor az (α) esetben annyival változtatjuk a tanagyagot, hogy x_k-t módosítjuk:

$$\bar{x}_k = \frac{\nu_k \bar{x}_k + \bar{x}_{i+1}}{\nu_k + 1}, \qquad \vartheta'_k = \vartheta_k,$$

ahol ν_k azt mutatja meg, ez az $\bar{x}_k$ objektum az eddigi iterációk során hányszor volt valamelyik új elem legközelebbi társa. Következésképpen ezt is változtatni kell az (α) esetben: $\nu'_k = \nu_k + 1$. Az SNN eljárásban tehát a tananyag



1. ábra. Jól szeparált osztályok esetén az adott halmazba kerülő első minta marad a tananyag egyetlen erre vonatkozó eleme, mely fokozatosan tart a megfelelő eloszlás (feltételes) várható értékéhez



2. ábra. Két normális eloszlású (1 szórású, $m_0 = 0$, ill. $m_1 = 2$ várható értékű) változó esetén az optimális döntés két tartományát az $(m_0 + m_1)/2$ pont határolja el

elemei bolyonganak, aminek értelme abban rejlik, hogy ezek szekvenciálisan közelítenek az azonos paraméterű környezetük „súlypontjához" (feltételes várható értékéhez). Míg az MNN-módszer konvergenciatétele lényegében egybevág a klasszikus NN-módszerével (*Cover* és *Hart*, 1967), addig az SNN-módszer ettől alapjaiban különbözik és elméletileg csupán csak az állítható, hogy jól elkülönülő osztályok esetén a nagy számok törvénye közvetlenül alkalmazható. A folyamat lényegét az *1. ábra* szemlélteti.

Szólnunk kell a statisztikai osztályozási eljárások jóságáról is. Mikor az új $\bar{x}$ objektum ismeretlen ϑ paraméterét az $\bar{x}_k$ legközelebbi társ ϑ_k paraméterével becsüljük, akkor egy döntést hozunk, melynek egy általunk előírt $W(\vartheta, \vartheta_k) \geqq 0$ rizikója van. Ennek várható értéke az átlagos rizikó $R_N(\vartheta) = EW(\vartheta, \vartheta_k)$. Ha a tananyagot minden határon túl kiterjesztjük ($N \to \infty$), akkor az aszimptotikus rizikó $R_\vartheta(\vartheta) = \lim_{N\to\infty} R_N(\vartheta)$ már a konkrét tananyagtól függetlenül az eljárás minőségét adja meg. A legoptimálisabb – legkisebb R^* aszimptotikus rizikójú – statisztikai eljáráshoz viszonyítva például $W'(\vartheta, \vartheta_k) = \{0$, ha $\vartheta = \vartheta_k$; 1, ha $\vartheta \neq \vartheta_k\}$ esetén $R' = R^*(1 - R^*)$, illetve $W''(\vartheta, \vartheta_k) = (\vartheta - \vartheta_k)^2$ esetén $R'' = 2R^*$.

A három eljárás összehasonlítása végett válasszunk egy egyszerű feladatot. Legyen $n = 1$ és a minták származzanak két statisztikai sokaságból, melyek mindegyike egydimenziós normális eloszlás 1 szórással, de eltérő várható értékkel. A várható érték $\vartheta = 0$ esetén $m_0 = 0$, $\vartheta = 1$ esetén pedig $m_1 = 2$. Mindhárom esetben $N = 100$ mintával „tanítjuk" a rendszert, majd 100 mintával

teszteljük. A rizikófüggvény $W'(\vartheta, \vartheta_k) = \{0$, ha $\vartheta = \vartheta_k$; 1, ha $\vartheta \neq \vartheta_k\}$, következésképpen az átlagos rizikó azt adja meg, milyen valószínűséggel téveszt az eljárás (ez a hibavalószínűség): $EW'(\vartheta, \vartheta_k) = P\{\vartheta \neq \vartheta_k\}$. Az elméletileg optimális esetben $x \leqq (m_0 + m_1)/2$ mellett az x paraméterére vonatkozó döntésünk 0, ellenkező esetben viszont 1, mint a *2. ábra* mutatja. (Vagyis aszerint döntünk, melyik osztály sűrűségfüggvénye a nagyobb.) Ekkor a minimális átlagos (elméleti) rizikó:

$$R^* = P\{\vartheta \neq \vartheta_k\} = P\{\vartheta_k = 1/\vartheta = 0\} P\{\vartheta_k = 0\} + P\{\vartheta_k = 0/\vartheta = 1\} P\{\vartheta = 1\} = 0{,}5 \int_1^\infty f_0(x)\, dx + 0{,}5 \int_{-\infty} f_2(x)\, dx = 0{,}16$$

I. TÁBLÁZAT

Eljárás	100 minta után a tananyag mérete	100 mintás teszt alapján	
		a hibavalószínűség	az optimális eljárás hibavalószínűsége
NN	100	0,20	0,13
MNN	32	0,18	0,13
SNN	28	0,18	0,13
Elméletileg optimális eljárás			0,16

feltéve, hogy mindkét osztály egyforma valószínűséggel figyelhető meg: $P\{\vartheta = 0\} = P\{\vartheta = 1\} = 0{,}5$. Itt $f_m(x)$ az 1 szórású, m várható értékű normális eloszlás sűrűségfüggvénye. Az eljárások eredményeit az *I. táblázat* adja meg.

Láthatóan tekintélyes mértékben redukálódott az osztályozásban részt vevő mintaelemek száma és ezzel a legközelebbi társak kikereséséhez szükséges számítási igény úgy, hogy emellett az empirikus hibavalószínűség nem növekedett. Az optimális eljárás hibavalószínűségét is párhuzamosan becsülve, nyilvánvalóan az egészen jól közelíti az elméletit, tehát a módszerek konvergenciája – ebben a példában – elfogadhatónak látszik.

Makroszinoptikus kódok felismerése

Olyan absztrakt paramétertér esetén, mint amilyet például makroszinoptikus kódok alkotnak, a felismerő algoritmusok minőségét sokkal nehezebb értékelni. A probléma úgy hidalható át, hogy vagy szinoptikai ismeretek birtokában tapasztalati úton kerül bevezetésre egy W rizikófüggvény, miként azt a bracknelli Meteorológiai Hivatal munkatársai tették (*Kaba* és *Faragó*, 1976), ahol tulajdonképpen a rizikófüggvény fordítottjáról, egy nyereségfüggvényről van szó), vagy valamilyen statisztikai módszerrel származtathatunk egy ilyen függvényt.

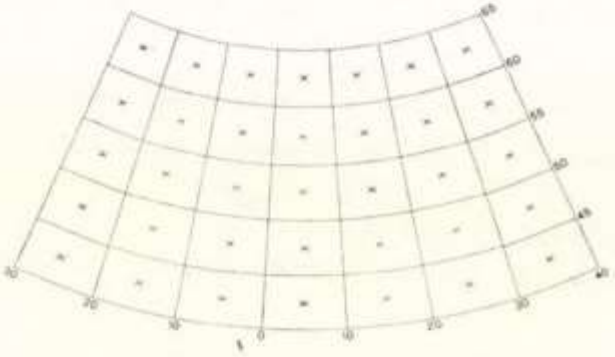
Mindenekelőtt definiáljuk a mintateret. A *Hess-Brezowsky* kódok elsősorban a tengerszinti légnyomáskép szinoptikus elemzésével, valamint – kisebb mértékben – az 500 mb-os abszolút topográfia alkalmazásával származ-

tathatók. Az algoritmusban a tengerszinti légnyomásmezőre szorítkozunk, melyet egy minimális 5×7-es rács pontjaiban adunk meg. A rács elhelyezkedését a *3. ábra* mutatja be. A mintavektorok tehát 35-dimenziósak. Két-két ilyen mező összehasonlítására a

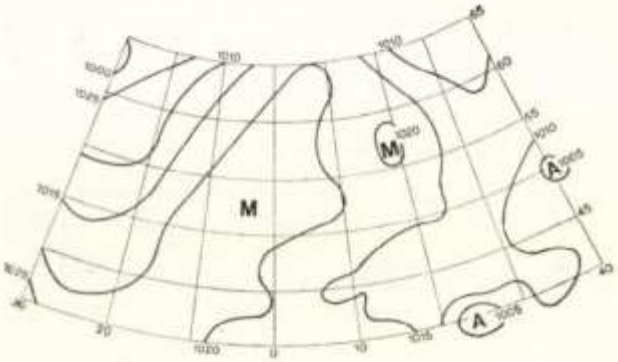
$$\varrho(\bar{x}', \bar{x}'') = \frac{1}{35} \sum_{j=1}^{35} \left| (x_j' - \mu') - (x_j'' - \mu'') \right|$$

függvényt (metrikát) alkalmazzuk, ahol μ a mező átlagát jelöli:

$$\mu = \frac{1}{35} \sum_{j=1}^{35} x_j,$$



3. ábra: A tengerszint légnyomás mező megadására alkalmazott rács. A rácspontokat „x" jelölik



4. ábra: Az 1977. június 23. 00 GMT időjárás helyzete: észak-keleti anticiklonális (NEA) típus

x_j, $j = 1, 2, \ldots, 35$ az $\bar{x}$ vektor komponensei. A napi *Hess-Brezowsky*-kódokat rendszeresen megjelentetik Offenbachban (Die Grosswetterlagen Europas) egy-egy éves összesítőben. Mintavételezéssel az 1970–1977-es évek anyagából megbecsültük a kódok a priori eloszlását, azaz a $P_1 = P\{\vartheta = 1\}, P_2 = P\{\vartheta = 2\}$ $\ldots, P_{28} = P\{\vartheta = 28\}, P_{30} = P\{\vartheta = 30\}$ valószínűségeket. Itt $\Theta = \{1, 2, \ldots, 28, 30\}$, a bracknelli jelölési rendszer szerint. (A $\vartheta = 29$ kód a szinoptikusan osztályozhatatlan mezőket jelölte, de ennek alkalmazásától itt eltekintettünk.) A minták alapján az egyes osztályok közepei (feltételes várható értékei) is becsülhetők; jelölje ezeket $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{28}, \bar{y}_{30}$. E közepek páronkénti „távolságaival" adjuk meg a rizikófüggvényt:

$$W(\vartheta, \hat{\vartheta}) = \varrho(\bar{y}_\vartheta, \bar{y}_{\hat{\vartheta}}).$$

(Megjegyzendő, hogy a bracknelli és az itt bevezetett függvény – az angol verzióban „scoring" – természetszerűleg egymástól eltérő, de mindamellett bizonyos fokú egyezés közöttük fennáll; egyetlenegy példával illusztrálva: az első soraik között a korrelációs együttható $-0{,}47$.)

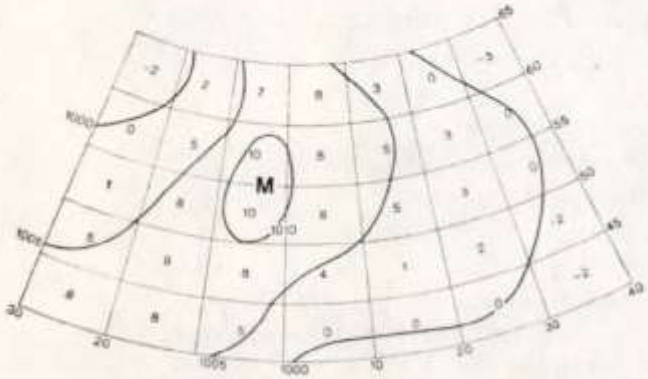
Ebben a konkrét sokdimenziós feladatban természetszerűleg nem adhatjuk meg az optimális eljárás átlagos rizikóját, ehelyett a későbbi tesztelés eredményeit egy triviális felismerési módszer minőségéhez hasonlíthatjuk. E módszer értelmében a paraméter ϑ becslését is a fent említett a priori eloszlással sorsoljuk ki. Jelölje ezt az eloszlást $P_0 = (P_1, P_2, \ldots, P_{28}, P_{30})$, akkor e triviális eljárás átlagos elméleti rizikója:

$$EW(\vartheta, \hat{\vartheta}) = \iint \varrho(\bar{y}_\vartheta, \bar{y}_{\hat{\vartheta}})\, dP(\vartheta, \hat{\vartheta}) = \int [\int \varrho(\bar{y}_\vartheta, \bar{y}_{\hat{\vartheta}})\, P_0(d\vartheta)]\, P_0(d\hat{\vartheta}) =$$
$$= \sum_i \left(\sum_j \varrho(\bar{y}_i, \bar{y}_j) P_j \right) P_i = 6{,}8.$$

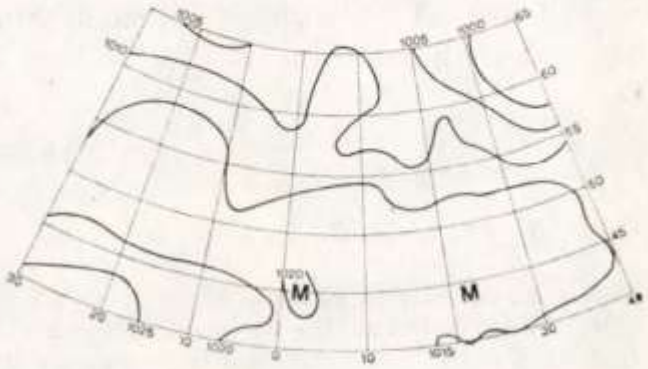
Ehhez fogjuk viszonyítani az SNN-eljárás empirikus rizikóját.

Az algoritmus P_0-eloszlású $n = 174$ mintából „tanul", azaz közvetlenül ennyi esetet figyelünk meg. A fokozatos szelekció eredményeképpen a tananyag végül 93 elemet tartalmaz. Az eljárás indításánál a korábban már említett 29 átlagmezőt (y_0) alkalmazzuk, a mintavétel csak ezt követően válik véletlenszerűvé. A tananyag csökkenése még ebben a bonyolultabb feladatban is tetemes. Az így képződött tananyag felismerőképességét 66 mintával teszteltük; ennek átlagos rizikója 2,9 volt. Az egyes szituációk osztályozásának elemzése is jól alátámasztotta ezt a triviális döntési eljárás viszonyítási szintjéhez képest igen jó eredményt.

Az algoritmikus osztályozás két konkrét példáját mutatjuk be. Az 1977.

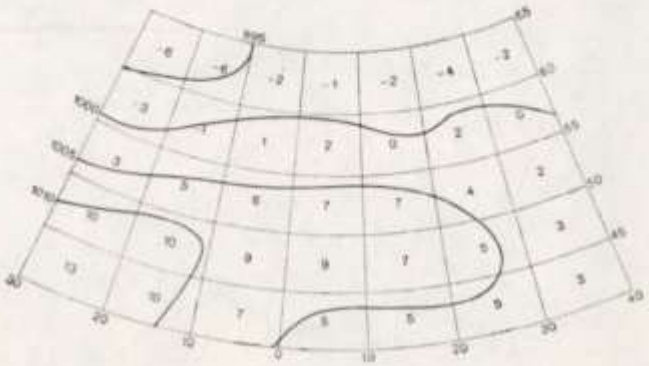


5. ábra: Az osztályozó tananyag egy északkeleti anticiklonális (NEA) eleme (a rácspontadatok az 1000 mb-hoz vett anomáliák)



6. ábra: Az 1977. június 25. 00 GMT időjárási helyzete: észak-nyugati anticiklonális (NWA) típus

június 23-i éjféli (OO GMT) helyzet *(4. ábra)* a *Hess-Brezowsky*-féle makroszinoptikus osztályozási rendszerben északkeleti anticiklonális besorolást kapott. Markáns helyzetről lévén szó, ez a megfeleltetés meglehetősen egyértelmű, annak ellenére, hogy a középpontjával a Brit-szigetek feletti anticiklon magja a megelőző napok helyzeteivel egybevetve folyamatosan elsimult. Az SNN-eljárással nyert osztályozó tananyagból a fenti helyzethez kapott legközelebbi társat az *5. ábra* mutatja be. Ez a mező a tanulási fázis során több minta szekvenciális átlagaként képződött (nevezetesen $\nu = 6$) és ennek megfelelően izobárjai az átlagosnál ritkábbak, mindamellett az adott esetben az osztálybasorolás tökéletesnek bizonyult. A vázolt időjárási helyzet 1977. június 25-re úgy módosult, hogy az említett anticiklon északról méginkább visszahúzódott, Közép-Európa időjárására már nem ennek előoldala volt a meghatározó, hanem egy majdnem tisztán zónális irányítottságú helyzet alakult ki *(6. ábra)*. Az Izland térségében elhelyezkedő ciklon hatására a zonalitás és egyben az időjárás is módosul: a *Hess-Brezowsky*-féle osztályozás ennek megfelelően észak-



7. ábra: Az osztályozó tananyag egy nyugati anticiklonális (WA) eleme (a rácspontadatok az 1000 mb-hoz vett anomáliák)

nyugati anticiklonális lesz. Az algoritmus ezt a helyzetet a *7. ábrán* látható legközelebbi társ kategóriájának megfelelően nyugati anticiklonális típusúként ismerte fel. A fentiek alapján nyilvánvaló, hogy e döntés szinoptikai szempontból is elég közel áll a valósághoz, amit többek között az is igazol, hogy e két kód bracknelli scoringja majdnem a maximális értékű.

A kapott teszteredmények és az itt bemutatott esettanulmányok is az eljárás alkalmazhatóságát igazolják; a klasszikus NN-módszerrel szemben a memóriaigény lényegesen kisebb, a felismerőképesség már viszonylagos kis minta után is számottevő. Mindamellett az ilyen és hasonló feladatok esetében a statisztikai osztályozási algoritmusok célszerű alkalmazása a „döntéselőkészítés", mely utóbb lehetővé teszi a szinoptikus esetleges beavatkozását a tényleges döntés kialakításakor.

A szerző köszönetét fejezi ki *Salamon Lászlónénak* a szükséges számítástechnikai munkákban vállalt lelkes tevékenységéért.

IRODALOM

Cover, T. M., Hart, P. E., 1967: Nearest neighbour pattern classification. IEEE Trans. on Inform. Theory, 1.

Cover, T. M., 1968: Estimation by the NN rule. IEEE Trans. on Inform. Theory, 1.

Gates, G. W., 1972: The reduced NN-rule. IEEE Trans. on Inform. Theory, 3.

Gulyás, O., 1977: Az analógia fogalma és felhasználása típusok képzésére. Időjárás, 1.

Hess, P., Brezowsky, H., 1952: Katalog der Grosswetterlagen Europas. Berichte des Deutschen Wetterdienstes in der US-Zone, Bad Kissingen, 33.

Hart, P. E., 1968: The condensed NN-rule. IEEE Trans. on Inform. Theory, 3.

Kaba M., Faragó T., Gulyás O., 1975: Az analógia elvén alapuló prognosztikai módszerek matematikai modellje. Időjárás, 3.

Péczely Gy., 1957: Grosswetterlagen in Ungarn. Orsz. Meteorológiai Szolgálat Kisebb Kiadványai, 30., Budapest.

Topor, N., 1954: A távelőrejelzés kutatása a Román Népköztársaságban. Időjárás, 6..

Wada, H., Kitahara, E., 1971: A proposal for classification of 500 mb patterns over the northern hemisphere. J. Meteo. Soc. Japan, 12.

Багров, Н. А., 1964: Индекс аналогичности векторых полей. Тр. ЦИП, 123.

Гирс, А. А., 1960: Основы долгосрочных прогнозов погоды. Гидрометеоиздат, Ленинград.

Груза, Г. В.—Раньков, Е. Я., 1970: О принципах автоматической классификации метеорологических объектов. Метеорология и Гидрология, 2.

Зверев, Н. И.—*Педь, Д. А.*, 1960: Определение аналогичности полей метеорологических элементов при помощи электронной считающей машины «Погода». Метеорология и Гидрология, 10.

Каба, М.—Фарago Т., 1976: Статистический прогноз месячных средних температур и оценка Байесова риска. *Időjárás*, 6.

Педь, Д. А., 1970: О критериях аналогичности гидрометеорологических полей. Тр. ГМЦ, 64.

Яковлева, Н. И., 1970: Применение статистических главных компонентов для целей объективной классификации метеорологических ситуаций и полей. Метеорология и Гидрология, 2.