A TANANYAG ELŐFELDOLGOZÁSAI A LEGKÖZELEBBI TÁRS MÓDSZERHEZ

Ph.D. tézisek

Ketskeméty László

Debrecen

2003. október

A dolgozat az alakfelismerés témaköréhez tartozó speciális problémákkal foglakozik.

Adott *objektumoknak* egy Ω halmaza, melynek minden eleme egyértelműen véges számú ismert *osztály* valamelyikéhez tartozik. Ha az osztályok számát L-lel jelöljük, az osztályok halmazát megfeleltethetjük az $\mathcal{Y} = \{1, 2, ..., L\}$ halmaznak. Azt, hogy az objektumok melyik osztályba tartoznak, az $Y : \Omega \to \mathcal{Y}$ függvény írja le, melyet *osztályozásnak* nevezünk. Az Y-t nem ismerjük. Az objektumokat egy X függvénnyel egyértelműen le lehet képezni egy $\mathcal{X}$ halmazra. Az $X : \Omega \to \mathcal{X}$ függvényt *alakzatnak* nevezzük, és ismertnek tételezzük fel. Az $\mathcal{X}$ az *alakzattér*, amely matematikai szempontból Ω-nál jobban modellezhető halmaz. $\mathcal{X}$ egy adott d metrikával lehet metrikus tér, lehet lineáris tér, akár véges dimenziós vektortér, de speciális esetben az $\mathbb{R}^p$ Euklideszi tér is. Az alakfelismerés alapfeladata az, hogy egy ismeretlen $\omega \in \Omega$ objektumról megállapítsuk, hogy melyik osztályhoz tartozik. Ehhez meg fogunk adni egy $\phi : \mathcal{X} \to \mathcal{Y}$, ún. *döntésfüggvényt*, amivel ω-hoz a $\phi(X(\omega))$ osztályt fogjuk rendelni. Az alapfeladat tehát regressziós: az ismert X egy függvényével akarjuk jól megbecsülni Y-t. Jelölje az $\mathcal{X}$ nyílt halmazai által generált σ-algebrát $\mathcal{B}$! Legyen az $X : \Omega \to \mathcal{X}$ mérhető leképezés. Tételezzük fel, hogy adott egy μ mérték az $(\mathcal{X}, \mathcal{B})$ mérhető téren, melyre abszolút folytonos X eloszlása, vagyis létezik az $f_X : \mathcal{X} \to \mathbb{R}$ Radon-Nykodim derivált: $\mathbf{P}(X \in B) = \int\limits_B f_X(x)\, d\mu(x)$, minden $B \in \mathcal{B}$-re.

Tekintsük most az $Y = i$ eseménynek az X által generált $\mathcal{G}_X = \sigma(X)$ σ-algebrára vett feltételes valószínűségét:

$$\eta_i(X) = \mathbf{P}(Y = i \mid \mathcal{G}_X).$$

Ekkor az (X, Y) pár együttes eloszlása kifejezhető η_i-vel és f_X-szel:

$$\mathbf{P}(X \in B, Y = i) = \int\limits_{\{X \in B\}} \eta_i(X) d\mathbf{P} = \int\limits_B \eta_i(x) \cdot f_X(x)\, d\mu(x).$$

A képletben szereplő $\eta_i(x)$, $i = 1, 2, ..., L$ függvények alkotját az *a'poszteriori eloszlást*, hiszen $\sum\limits_{i=1}^{L} \eta_i(x) = 1$ minden $x \in \mathcal{X}$ esetén. Adott egy $\underline{\underline{W}} = (w_{ij})$ $L \times L$-es mátrix, a *veszteségmátrix*, ahol w_{ij} annak a rossz döntésnek a veszteségét jellemzi, amit akkor követünk el, ha $\phi(X) = i$, de $Y = j$. A ϕ döntésfüggvény *rizikóján* az

$$R\left(\underline{\underline{W}}, \phi\right) = \sum_{i=1}^{L} \sum_{j=1}^{L} w_{ij} \cdot \mathbf{P}(\phi(X) = i, Y = j) =$$

$$= \sum_{i=1}^{L} \sum_{j=1}^{L} w_{ij} \cdot \int\limits_{\{\phi(X)=i\}} \eta_j(X) d\mathbf{P} = \sum_{i=1}^{L} \sum_{j=1}^{L} w_{ij} \cdot \int\limits_{\{\phi(x)=i\}} \eta_j(x) \cdot f_X(x)\, d\mu(x) =$$

$$= \sum_{i=1}^{L} \int\limits_{\{\phi(x)=i\}} f_X(x) \cdot \sum_{j=1}^{L} w_{ij} \cdot \eta_j(x) d\mu(x)$$

értéket értjük. A rizikót minimalizáló döntésfüggvényt *Bayes-döntésfüggvénynek* nevezzük, és ϕ^*-gal jelöljük. A minimalizált rizikó az R^* *Bayes-rizikó.*

Belátható, hogy

$$\phi^*(x) = i, \text{ ha } \sum_{j=1}^{L} w_{ij} \cdot \eta_j(x) \leq \sum_{j=1}^{L} w_{kj} \cdot \eta_j(x) \text{ minden } k \neq i\text{-re.}$$

Ha $w_{ij} = 1 - \delta_{ij}$, akkor a Bayes döntésfüggvény egyszerűbben is megadható:

$$\phi^*(x) = i, \text{ ha } \eta_i(x) \geq \eta_k(x) \text{ minden } k \neq i\text{-re.}$$

Ilyenkor a Bayes-rizikó az $L^* = \mathbf{P}\left(\phi^*\left(X\right) \neq Y\right)$ hibás döntési valószínűséget jelenti.

Az (X, Y) pár együttes eloszlásának ismeretében tehát ϕ^* megszerkeszthető. Az alkalmazások többségében azonban ezt nem ismerjük, viszont adott egy

$$\left(X, Y\right), \left(X_1, Y_1\right), \left(X_2, Y_2\right), ..., \left(X_n, Y_n\right), ...$$

független, azonos eloszlású párokból álló sorozat, melyet felhasználhatunk a döntésfüggvény (rekurzív) megadásához. A

$$\mathcal{D}_n = \left\{\left(X_1, Y_1\right), \left(X_2, Y_2\right), ..., \left(X_n, Y_n\right)\right\}$$

halmazt *tananyagnak* nevezzük. X_i az i-edik *tanulópont*, Y_i a megfelelő *tanítás.* Gyakran használjuk a $\mathcal{T}_n = \{X_1, X_2, ..., X_n\}$ *tanulópont-halmaz* elnevezést is, ilyenkor az X_i tanulópont tanítására a $class(X_i)$ jelöléssel fogunk hivatkozni. A tanulóalgoritmusok a $\mathcal{D}_n$ tananyag függvényeként állítják elő a $g_n : \mathcal{X} \to \mathcal{Y}$ döntésfüggvényt.

Tanulóalgoritmusoknál az osztályozás folyamata három lépésből áll.

1^o A tananyag átszerkesztése (előfeldolgozása).

2^o A döntésfüggvény előállítása.

3^o Ismeretlen alakzatvektorok osztályozása.

Az első két lépést csak egyszer kell végrehajtani, míg a harmadik lépésben az osztályozandó alakzatok száma igen nagy is lehet. A tananyag átszerkesztésének és a döntésfüggvény előállításának költségei így eltörpülnek az osztályozás költségeihez képest. Ezért érdemes a tananyagot előfeldolgozás alá vonni, hiszen az osztályozás során a futtatási idő-, számítási művelet- vagy tárolási kapacitás

csökkenés formájában jelentkező megtakarítás az előfeldolgozás költségeit bőségesen fedezni fogja.

A tanulóalgoritmusok g_n döntésfüggvény-sorozatához értelmezhető a

$$\mathbf{P}\left(g_n(X)=i, Y=j \mid \mathcal{D}_n\right)=\mathbf{P}\left(g_n(X)=i, Y=j \mid \mathcal{G}\right)$$

feltételes valószínűség, ahol a feltételben szereplő $\mathcal{G}$ a $\mathcal{D}_n$ tananyag által generált σ-algebra. A g_n döntésfüggvény feltételes rizikóján a:

$$R_n=R\left(\underline{\underline{W}}, g_n\right)=\sum_{i=1}^{L} \sum_{j=1}^{L} w_{ij} \cdot \mathbf{P}\left(g_n\left(X\right)=i, Y=j \mid \mathcal{D}_n\right)$$

valószínűségi változót értjük.

Ha $\lim_{n\to\infty} \mathbf{E}R_n = R^*$, akkor a $\{g_n\}$ döntésfüggvény-sorozat, illetve a kapcsolatos tanulóalgoritmus *gyengén konzisztens.*

Az alkalmazásokban az egyik legnépszerűbb tanulóalgoritmus a *legközelebbi szomszéd módszer* (*vagy legközelebbi társ módszer*), amelynél a döntésfüggvény-sorozat definíciója.:

$$g_n\left(x\right)=Y_j, \text{ ha } d\left(x, X_j\right)=\min_{\alpha=1,2,\ldots,n} d(x, X_\alpha).$$

Általános feltételek mellett megmutatható, hogyha létezik $\lim_{n\to\infty} \mathbf{E}\left(R\left(\underline{\underline{W}}, g_n\right)\right) = R_{NN}$, akkor $R^* \leq R_{NN} \leq 2R^*$, vagyis $R^*=0$ esetben a legközelebbi társ tanulóalgoritmus gyengén konzisztens lesz.

A dolgozat kizárólag a legközelebbi társ módszerrel foglalkozik. Az az alapfeladat, hogy egy $x \in \mathcal{X}$ metrikus ponthoz megtaláljuk egy $\mathcal{T}_n \subset \mathcal{X}$ véges elemszámú halmaz $t_{NN}(x)$-szel jelölt legközelebbi elemét. Ismeretes (pl. DEVROYE (1981)), hogy metrikus térben, az alakzat általános eloszlása mellett a legközelebbi társ módszer a mintaelemszám végtelenhez tartása esetén olyan ϕ_n döntésfüggvény-sorozatot produkál, melyek osztályozási hibája az elvileg legjobb (Bayes) hiba kétszeresénél nem lehet nagyobb. Egyszerűsége és ez a tulajdonság teszi azt, hogy a legközelebbi társ módszer az alkalmazásokban igen népszerű. Hátránya az, hogy mindig csak véges elemszámú tananyag áll a rendelkezésre, mellyel az osztályozás igen költséges, hisz az osztályozandó ponthoz mindegyik tanulóponttól vett távolságot ki kell számolni. Egyes alkalmazásokban a távolság kiszámítása nagyon hosszadalmas és költséges. Gondoljunk pl. a meteorológiai hosszútávú analógiás előrejelzésnek a problémájára. A pillanatnyi időjárási helyzetet megelőző féléves időszakot kell a meteorológiai archívumban tárolt féléves hosszúságú időszakaival összehasonlítani. Ki kell választani azt az évet, amelyben a félévnyi lefolyás leginkább hasonlít a most vizsgálandó időszak féléves előzményéhez. Ezt az évet egy nagyon komplikált metrikafüggvény segítségével, a legközelebbi társ módszerrel határozzák meg. Ennek a legközelebbi társ időjárási helyzetnek a folytatódása képezi az előrejelzés

alapját. Az alkalmazásokban tehát felvetődik az a kérdés, hogy miképpen lehetne csökkenteni a tananyag elemszámát, hogy az osztályozás pontossága ne csökkenjen jelentős mértékben. (Hasonlatos a probléma a kecske és a káposzta dillemájához...) A disszertáció azokat a lehetőségeket járja körül, hogy mit lehet tenni ennek az ügynek az érdekében. Milyen átalakítások, szerkesztési műveletek szükségesek ahhoz, hogy az osztályozás költségeit leszorítsuk miközben nem romlik a pontosság. A szerkesztési műveletek vonatkozhatnak egyrészt a $\mathcal{D}_n$ tananyag tanulópontjaira, a tanításokra és a távolságot definiáló d metrikafüggvényre egyaránt. Megvizsgáljuk azt a kérdést is, hogy egyáltalán szükség van-e minden tanulóponttal való távolság kiszámítására, vagy kevesebb lépéssel is eljuthatunk a legközelebbi társhoz.

A dolgozat három fejezetre tagozódik.

Az első fejezet témája a $\mathcal{D}_n$ **tananyag szerkesztése**. Alapvetően négy szerkesztéstípust tárgyalunk.

1. *A $\mathcal{D}_n$ ritkítása.* Ha a $\mathcal{D}_n$ tananyag olyan $\mathcal{D}_m^*$ részhalmazát állítjuk elő, mellyel a $\mathcal{D}_n$ tanulópontjai $\mathcal{D}_m^*$-gal pontosan osztályozhatók lesznek, *ritkításról* beszélünk. pontosabban fogalmazva olyan $\mathcal{T}_m^* \subset \mathcal{T}_n$ részhalmaz előállítása a cél, amely rendelkezik az alábbi két tulajdonsággal:

a.) $\mathcal{T}_m^*$ pontjai pontosan osztályozzák a $\mathcal{T}_n \setminus \mathcal{T}_m^*$ tanulópontokat, azaz $\mathcal{T}_m^*$ *konzisztens.*

b.) $\mathcal{T}_m^*$ minimális elemszámú konzisztens részhalmaza $\mathcal{T}_n$-nek.

Összefoglaljuk az irodalomban javasolt ritkító algoritmusokat. A tananyag ritkításának témakörében az első eredmény HART (1968) kondenzált legközelebbi társ módszerének (CNN) megjelenése volt. HART iteratív algoritmusa olyan konzisztens $\mathcal{T}_m^* \subset \mathcal{T}_n$ halmazt állít elő, ami függ a feldolgozás sorrendjétől, és általában nem minimális elemszámú. A CNN-módszert többen módosították. Legélesebb eredmény RITTER et.al (1975) nevéhez fűződik. Közölt algoritmusuk minimális elemszámú konzisztens részhalmazt állít elő. A közölt algoritmus bonyolult, nehezen implementálható. Érdekes TOMEK (1976) megoldása, amely vektortérben Euklideszi-metrikánál alkalmazható. Nem bonyolult, iterációs lépések sorozatában törekszik eljutni a konzisztens részhalmazhoz, hanem felállít egy kritériumot a konzisztens halmaz elemeire. A feltételt kielégítő *idegen pontpárok* halmaza ugyan még nem konzisztens halmaz, de ebből kiindulva a CNN-módszerrel kevesebb iterációval lehet konzisztens halmazt konstruálni. A disszertációban újdonságként megadjuk TOMEK idegen-pontpárokat szerkesztő algoritmusának általános metrikus térben is alkalmazható változatát.

2. *A $\mathcal{D}_n$ tömörítése.* Ekkor a $\mathcal{D}_n$ alakzatpontjait helyettesítőkkel váltjuk fel, megkonstruálva egy kisebb elemszámú az osztályozásnál tananyagként felhasználható $\mathcal{P}_m$ *prototípus-halmazt.* $\mathcal{P}_m$ tanulópontjai nem feltétlenül vannak benne $\mathcal{D}_n$-ben, hanem a metrikus tér alkalmasan kiválasztott új pontjai is hozzátartozhatnak $\mathcal{P}_m$-hez. Tehát a $\mathcal{P}_m = \{(z_1, l_1), (z_2, l_2) ..., (z_m, l_m)\}, z_i \in \mathcal{X}$

, $l_i \in \mathcal{Y}$ a $\mathcal{D}_n$ *prototípusa*, ha tananyag, $m << n$ és $\mathcal{P}_m$ pontosan osztályozza $\mathcal{D}_n$-t. Általában $\mathcal{P}_m \nsubseteq \mathcal{D}_n$, sőt $\mathcal{P}_m \cap \mathcal{D}_n = \emptyset$ is elképzelhető.

A dolgozatban áttekintjük az irodalomban javasolt megoldásokat. A tömörítési algoritmusok egyik része speciális metrikus térben alkalmazható. Kihasználják, hogy $\mathcal{X}$ elemei pl. vektorok, és a prototípus pontokat a tanulópontok átlagaival definiálják. A másik lehetséges megoldás az, amikor a tanításokat változtatják meg, és a kapott halmazt ritkítják. A disszertációban a szerző által közölt tömörítő eljárás általános metrikus térben alkalmazható. Feltétel az, hogy rendelkezésre álljon egy $\mathcal{H}_k \subset \mathcal{X}$ segédpont-halmaz is. A prototípushalmaz szerkesztése úgy történik, hogy $\mathcal{H}_k$-ből kiválasztjuk azokat az elemeket, amelyekkel $\mathcal{D}_n$ tanulópontjai közül egyszerre többet is ki tudunk váltani.

3. A $\mathcal{D}_n$ tananyag átdefiniálása. A harmadik átszerkesztési lehetőség a tananyag osztályainak átdefiniálása. Ezen egy $f : \{1, 2, ..., L\} \to \{1, 2, ..., K\}$ transzformáció megadását értjük, amikor is $t \in \mathcal{T}_n$ esetén $class(t)$ helyett $f(class(t))$-t tekintjük a t tanulópont új tanításának. Ha a tanulópont eddig az i osztályhoz tartozott, ezután az $f(i)$ osztályhoz fog. Az új tananyaggal javul az osztályozás pontossága. Néhány alkalmazásnál tapasztalható, hogy különböző osztályok tanulópontjai teljesen összekeveredhetnek, illetve van olyan jelenség is, amikor bizonyos osztályok pontjai markáns klaszterekbe szeparálódnak. Pl. egy digitális műholdfelvételen a kép pixelei a földfelszín olyan osztályaihoz kapcsolhatók, mint a búza, kukorica, napraforgó stb. Azonban a vegetációs időszak kezdeti szakaszában ezek az osztályok teljesen egybeolvadhatnak. Egy későbbi időpontban készített felvételen viszont pl. a búza képpontjai különböző jól szétválasztható alosztályokba tömörülnek a minőség alapján. A tananyag átdefiniálási lehetőségeinek áttekintése ennek a disszertációnak az eredménye.

4. *A $\mathcal{D}_n$ szűrése.* Megfogalmazunk olyan kritériumokat, amelyek az egyes osztályokban a tipikus tanulópontoktól szélsőségesen eltérő tulajdonságú tanulópontokra igazak. Ezek a *deviáns* tanulópontok, melyeket az osztályozásnál máshogy vesszük majd figyelembe. A *deviáns* (vagy outlier) tanulópontok az osztályozások során a hibás besorolások többségét okozhatják. A disszertációban megadjuk a deviáns tanulópontok definícióit. A deviáns pontoknak összesen tíz típusát értelmezzük. Az egyes tanulópontokat deviancia súlyokkal lehet felszerelni, attól függően, melyik deviancia-kritériumnak feleltek meg. Amikor a deviáns pontokat másképpen vesszük figyelembe a döntésben, akkor azt mondjuk, hogy a tananyagot *megszűrtük.*

- *Konjuktív szűrés.* $\mathcal{D}_n$-ből töröljük azokat a tanulópontokat, amelyek az adott deviancia-kritériumok mindegyikének szimultán megfelelnek.

- *Többségi szűrés.* $\mathcal{D}_n$-ből töröljük azokat a tanulópontokat, amelyek az adott r deviancia-kritérium közül legalább $q\,(< r)$-nak megfelelnek.

- *Súlyozott szűrés.* $\mathcal{D}_n$-ből azokat a t tanulópontokat töröljük, melyek deviancia súlya meghalad egy előírt küszöböt: $dev(t) > T_{dev}$.

Szűréskor tehát olyan $T^* \subset \mathcal{T}_n$ részhalmaz előállítása a cél, amelynél T^* pontjait az legközelebbi szomszéd módszerrel pontosabban lehet osztályozni. A

$\mathcal{T}_n \setminus T^*$ halmaz elemei lesznek az adott d metrikával osztályaik rosszul osztályozható (*deviáns* vagy outlier) tanulópontjai. T^*-ot a fenti 1. pont szerint vagy 2. pont szerint további feldolgozásnak vetjük alá, $\mathcal{T}_n \setminus T^*$ pontjait vagy külön kezeljük, vagy pedig mérési hibákból eredőeknek tekintjük, és elhagyjuk a további számolásból őket.

A dolgozat második fejezetének témája **a tanulópontok gyors megkeresése.**

A metrikus térben a háromszög-egyenlőtlenségeken alapuló kizárási feltételeket fogalmazhatunk meg, melyeket beépítve a keresés folyamatába gyorsabban eljuthatunk a legközelebbi társhoz.

A dolgozatban öt kizárási feltételt fogalmazunk meg, melyek a következők:

2.1. TÉTEL: A $t \in \mathcal{T}_n \setminus \{t_{i_1}, t_{i_2}, ..., t_{i_k}\}$ tanulópont biztosan nem legközelebbi szomszédja x-nek, ha az alábbi kizárási feltételek valamelyike igaz rá:

$$\mathcal{K}_1: \qquad 2d_{\min}^{(k)} < d\left(t, t_{NN}^{(k)}\right);$$

$$\mathcal{K}_2: \qquad 0 < \max_{j=1,\ldots,k} \left(d\left(t, t_{i_j}\right) - 2d\left(x, t_{i_j}\right)\right);$$

$$\mathcal{K}_3: \qquad d_{\min}^{(k)} < \max_{j=1,\ldots,k} \left|d\left(t, t_{i_j}\right) - d\left(x, t_{i_j}\right)\right|;$$

$$\mathcal{K}_4: \qquad d_{\max}^{(k)} - d_{\min}^{(k)} > d\left(t_{FN}^{(k)}, t\right);$$

$$\mathcal{K}_5: \qquad 2d_{\min}^{(k)} < d\left(t, t_{NN}^{(k)}\right) \text{ vagy } d_{\max}^{(k)} - d_{\min}^{(k)} > d\left(t_{FN}^{(k)}, t\right);$$

ahol $d_{\min}^{(k)} = \min_{j=1,\ldots,k} d(x, t_{i_j}), d_{\max}^{(k)} = \max_{j=1,\ldots,k} d(x, t_{i_j})$,
$\mathrm{t}_{NN}^{(k)} = \arg \min_{j=1,\ldots,k} d(x, t_{i_j}), t_{FN}^{(k)} = \arg \max_{j=1,\ldots,k} d(x, t_{i_j})$.

A felsorolt kizárási feltételek közül csak $\mathcal{K}_3$-nak vannak előzményei az irodalomban, a többi új eredmény.

A 2.2 és 2.3 tételek megadják a kizárási feltételek között meglévő erőkapcsolatot:

2.2. TÉTEL: $\mathcal{K}_i \Longrightarrow \mathcal{K}_j$, ha $(i,j) \in \{(1,2), (2,3), (4,3), (5,3)\}$, ahol a $\Longrightarrow$ arra utal,ha $\mathcal{K}_i$ kizár egy $x \in \mathcal{X}$ pontot, akkor $\mathcal{K}_j$ is kizárja azt.

2.3. TÉTEL: $\mathcal{K}_i \not\Longrightarrow \mathcal{K}_j$, ha
$(i,j) \in \{(3,2), (3,1), (2,1), (3,4), (3,5), (2,4), (4,2), (1,4), (4,1)\}$.

A 2.4 és 2.5 tételekben a $\mathcal{K}_1$ kizárási elv két olyan tulajdonságát tárgyaljuk, ami a kizárásos kereső stratégiák megalkotásánál játszanak szerepet:

2.4. TÉTEL: A $\mathcal{K}_1$ kizárási feltétel antiszimmetrikus és nem tranzitív. (Ha a kizárja b-t, akkor b nem zárhatja ki a-t. Viszont létezhetnek olyan a, b, c tanulópontok, hogy bár a kizárja b-t, b c-t, de a nem zárja ki c-t.)

2.5. TÉTEL: Ha $3d(x,a) < d(a,b)$ és $3d(x,b) < d(b,c)$, akkor $3d(x,a) < d(a,c)$.

A kizárási feltételek egyre élesednek, attól függően mely tanulópontoktól vett távolságokat számoltuk már ki. Fontos tehát, hogy jó sorrendben dolgozzuk fel a tanulópontokat. A feldolgozás sorrendjét meghatározó elvet *keresési stratégiának* nevezzük. A disszertációban megvizsgálunk néhány szóbajöhető keresési stratégiát. A kísérleti futtatások alapján a *minimax keresési stratégia* tűnik a legjobbnak. Tegyük fel, hogy már feldolgoztuk a $t_{\gamma_1}, t_{\gamma_2}, ..., t_{\gamma_k}$ tanulópontokat. A következő, $k+1$-edik tanulópont az lesz, melyre

$$t_{\gamma_{k+1}} = \arg \min_{t \in \mathcal{T} \setminus \{t_{\gamma_1}, t_{\gamma_2}, ..., t_{\gamma_k}\}} \max_{i=1,...,k} \left| d\left(t, t_{\gamma_i}\right) - d\left(t_{\gamma_i}, x\right) \right|$$

teljesül. A minimax keresés tehát arra a tanulópontra pozícionál rá, amely leginkább úgy helyezkedik el a metrikus térben mint az x a $t_{\gamma_1}, t_{\gamma_2}, ..., t_{\gamma_k}$ tanulópontokhoz képest. A disszertációban foglalkozunk a kizárásos kereső stratégiák hatékonyságának jellemzésével. Megadunk egy olyan algoritmust, amely a tanulópontokat úgy rendezi át adott x osztályozandó pont ismeretében, hogy a legközelebbi társ megkeresésének lépésszáma minimális legyen. Ezzel a minimális lépésszámmal lehet összehasonlítani egy tesztelendő NN-kereső stratégia lépésszámát. A minimális lépésszámnak a meghatározásához elő kell állítani az x ponttól függő ún. *optimális janicsárhalmazt.* Az optimális janicsárhalmazt előállító algoritmus a disszertáció eredménye. Két kizárásos NN-kereső stratégia esetében meg is tesszük a hatékonyság ellenőrzését, amikoris 30 számítógéppel szimulált esetben elvégezzük különböző beállítások mellett a keresést.

A dolgozatban megmutatjuk, hogyan lehet a hagyományos szekvenciális és a legjobbnak tartott $\mathcal{K}_3$ kizárásos minimax kereső stratégiával működő algoritmusokat összehasonlítani. Megállapítható, hogy bonyolult metrikák esetén még kis mintaelemszámnál is a kizárásos keresés a jobb.

Nemcsak pontonként, hanem *csoportonkénti kizárási elv* is érvényesíthető, ha előzőleg megfelően struktúráljuk a $\mathcal{T}_n$ tanulópont halmazt. A *lefedő gömb* fogalma segítségével definiáljuk a *szeparálhatóság* fogalmát.

2.1. DEFINÍCIÓ Legyen $A \subset \mathcal{T}_n$. A $G(c,R) = \{x;\ x \in \mathcal{X},\ \mathrm{d}(c,x) \leq R\}$ c centrumú, R sugarú zárt gömb az A-t lefedő gömb, ha $c \in A$ és $A \subset G(c,R)$. Az A-t lefedő zárt gömbök között azt a $G_A(c_A, R_A)$-val jelöltet fogjuk *legkisebb lefedő zárt gömb*nek nevezni, amelynél sugár minimális. A c_A az A halmaz *centruma*, R_A pedig a *takaró-sugár.*

2.2. DEFINÍCIÓ: Legyenek $A_1, A_2 \subset \mathcal{T}_n$. Jelölje rendre c_1, c_2 a centrumaikat, és R_1, R_2 a takaró-sugaraikat. Az A_1, A_2 halmazok *szeparálhatóak*, ha $\mathrm{d}(c_1, c_2) > R_1 + R_2 + \max\{R_1, R_2\}$.

Szeparált halmazok esetén lehetőség nyílik a kereséskor csoportos kizárásra.

2.6. TÉTEL: Legyenek $A_1, A_2 \subset \mathcal{T}_n$ szeparálhatóak. Ha $x \in \mathcal{X}$-re teljesül, hogy $\mathrm{d}(x, c_1) < R_1$, akkor $\min_{z \in A_1} \mathrm{d}(x, z) < \min_{y \in A_2} \mathrm{d}(x, y)$.

2.7. TÉTEL: Legyenek $A_1, A_2 \subset \mathcal{T}_n$ szeparálhatóak. Ha $x \in \mathcal{X}$-re teljesül, hogy $\mathrm{d}(x, c_1) + R_2 < \mathrm{d}(x, c_2)$, akkor $t_{NN}(x) \notin A_2$.

2.8. TÉTEL: Tegyük fel, hogy már kiszámoltuk a $d(x, t_1), d(x, t_2), ..., d(x, t_k)$ távolságokat. Legyen $d_{\min}^{(k)} = \min_{i=1,...,k} d(x, t_i)$, és $A \subset \mathcal{T}_n \setminus \{t_1, ..., t_k\}$.
Ha $d_{\min}^{(k)} + R_A < d(x, c_A)$, akkor $t_{NN}(x) \notin A$.

A harmadik fejezet tárgya **a metrikák optimális átskálázása**. Egy adott metrikát alkalmas súlyok segítségével hogyan lehet úgy átkariblálni, hogy az osztályozás pontossága növekedjen? Az átkalibrált metrikával hogyan változik meg az NN-keresés sebessége? Ezeket a kérdéseket tisztázzuk ebben a fejezetben. Hogyan lehet adott $\mathcal{X}$ alaphalmazhoz metrikafüggvények egy Δ halmazából az optimálisat kiválasztani, amivel egy teszthalmazon az osztályozási hiba relatív gyakoriságát minimalizáljuk. A leírás során az $\mathcal{X}$ alaphalmaz a p-dimenziós $\mathbb{R}^p$ vektortér lesz. Egy $d(\underline{x}, \underline{u}) = \sum_{i=1}^{p} \rho(x_i, u_i)$ szerkezetű bázismetrikából indulunk ki, ahol ρ egy metrika $\mathbb{R}$-en. Ilyen pl. az *Euklideszi-*, *Minkowsky-*, *City-blokk* vagy a *Canberra*-metrika.

A bázismetrikával generáljuk a

$$\Delta = \left\{ d_{\underline{A}}; d_{\underline{A}}(\underline{x}, \underline{u}) = \sum_{i=1}^{p} A_i \rho(x_i, u_i), \underline{A} \geq \underline{0} \right\}$$

halmazt. Δ minden eleme metrika $\mathbb{R}^p$-n. A célkitűzésünk az, hogy ebből válasszuk ki az optimálisat.

Az n elemszámú tananyagot egy m elemszámú $\mathcal{D}_m^*$ és egy l elemszámú $\mathcal{D}_l^{**}$ részhalmazra bontjuk fel. ($n = m + l$). A továbbiakban $\mathcal{D}_m^*$ játsza a tananyag szerepét az osztályozáskor. $\mathcal{D}_l^{**}$-et *tesztanyagnak* fogjuk nevezni. Jelölje a tananyagot

$$\mathcal{D}_m = \{(\underline{x}_1, y_1), (\underline{x}_2, y_2), ..., (\underline{x}_m, y_m)\}, \text{ ahol}$$

$\underline{x}_i$ az i-edik *tanulópont*, y_i az i-edik *tanítás*.

A tesztanyag

$$\mathcal{D}_l^{**} = \left\{ \left(\underline{x}_{m+1}, y_{m+1}\right), \left(\underline{x}_{m+2}, y_{m+2}\right), ..., \left(\underline{x}_{m+l}, y_{m+l}\right) \right\}$$

lesz, ahol $\underline{x}_{m+j}$ az j-edik *tesztpont.*

Jelölje a Δ metrika-családhoz tartozó súlyozott metrikás legközelebbi társ döntésfüggvények halmazát C_m! A döntésfüggvényeket a kapcsolatos $\underline{A}$ skálázó-vektorral fogjuk indexelni.

A $\phi_{\underline{A}}$ döntésfüggvényhez tartozó, a $\mathcal{D}_m$ feltételre vett feltételes döntési hiba valószínűségét $L\left(\phi_{\underline{A}}\right) = \mathbf{P}\left(\phi_{\underline{A}}\left(\underline{X}\right) \neq Y \mid \mathcal{D}_m\right)$ jelöli. A $\phi_{\underline{A}}$ döntésfüggvény az $L\left(\phi_{\underline{A}}\right) - \inf_{\phi \in C_m} L\left(\phi\right)$ különbséggel jellemezhető.

Jelölje $g_m^* \in C_m$ azt a döntésfüggvényt, amelynél

$$\hat{L}_{m,l}\left(g_m^*\right) = \min_{\phi_{\underline{A}} \in C_m} \hat{L}_{m,l}\left(\phi_{\underline{A}}\right), \text{ ahol}$$

$\hat{L}_{m,l}\left(\phi_{\underline{A}}\right) = \frac{1}{l} \sum_{i=1}^{l} I_{\left\{\phi_{\underline{A}}\left(\underline{x}_{m+i}\right) \neq y_{m+i}\right\}}$ az $L\left(\phi_{\underline{A}}\right)$ döntési hibavalószínűség empirikus becslése.

Az optimális $\underline{A}^* \geq \underline{0}$ skálázó vektort úgy fogjuk előállítani, hogy felírjuk azt egy $\underline{\underline{G}} \cdot \underline{A} \leq \underline{0},\ \underline{A} \geq \underline{0},\ \underline{A} \neq \underline{0}$ homogén lineáris egyenlőtlenség-rendszer megoldásaként, ahol a $\underline{\underline{G}}$ mátrix minden sorában van legalább egy negatív komponens. A $\mathcal{D}_l^{**}$ teszthalmaz minden eleméhez megkísérlünk előállítani egy vagy több sorvektort a $\underline{\underline{G}}$ mátrixban. Amelyik tesztpontnál ez nem sikerül, azokat rosszul fogja osztályozni még az optimális g_m^* döntésfüggvény is, de ezeket minden más $\phi_{\underline{A}} \in C_m$ döntésfüggvény is rosszul osztályozná. Az ismertetendő algoritmus két fázisban oldja meg a feladatot. Az első fázisban megadjuk a $\underline{\underline{G}}$ mátrixot. A feladat felállításához szükséges $\underline{\underline{G}}$ együtthatómátrix komponenseinek előállítása ennek a dolgozatnak az eredménye. A második fázisban meghatározzuk a g_m^*-hoz tartozó $\underline{A}^* \geq \underline{0}$ skálázó vektort. A felállított egyenlőtlenségrendszer megoldása FARKAS GYULA (1902) klasszikus módszerével történik.

A legfontosabb eredményt a 3.1. tételben mondjuk ki.

3.1. TÉTEL: Az $\hat{L}_{m,l}\left(\phi\right) = \frac{1}{l} \sum_{i=1}^{l} I_{\left\{\phi\left(\underline{x}_{m+i}\right) \neq y_{m+i}\right\}}$ empirikus hibavalószínűséget minimalizáló $g_m^* = \arg \min_{\phi_{\underline{A}} \in C_m} \hat{L}_{m,l}\left(\phi_{\underline{A}}\right)$ döntésfüggvényre minden $\varepsilon > 0$ esetén fennáll, hogy

$$\mathbf{P}\left(L\left(g_m^*\right) - \inf_{\phi \in C_m} L\left(\phi\right) > \varepsilon \mid \mathcal{D}_m \right) \leq 4\,\mathrm{e}^8 \left(l^2 \binom{m}{2} \right)^p \mathrm{e}^{-\frac{l\varepsilon^2}{2}}, \text{ ahol}$$

$$L\left(\phi\right) = \mathbf{P}\left(\phi\left(\underline{x}\right) \neq y \mid \mathcal{D}_m\right).$$

A fenti tétel segítségével meg lehet adni azt, hogy a súlyvektor előállításakor a tananyag-tesztanyag elemszám $m : l$ aránya hogyan választandó meg. Az $l \approx \frac{n}{\ln n}$, $m \approx n - \frac{n}{\ln n}$ választás esetén a becslő kifejezés minimalizálható az $n \to \infty$ átmenetkor, azaz viszonylag kis elemszámú tesztanyag leválasztásával pontossá tehető a metrikasúlyozásos osztályozás.

A fejezet végén megmutatjuk, hogy a súlyvektort előállító algoritmus alkalmazható metrikák keveréséhez felhasználható együtthatórendszer előállítására is. Legyenek $d_1, d_2, ..., d_p$ különböző metrikák $\mathcal{X}$-hez. Ezekkel, mint bázismetrikákkal képezzük a

$$\Delta = \left\{ d_{\underline{A}};\ d_{\underline{A}}(x,u) = \sum_{i=1}^{p} A_i d_i(x,u),\ \underline{A} \geq \underline{0} \right\}$$

metrika-halmazt. Feladatunk az, hogy $\mathcal{D}_m^*$ és $\mathcal{D}_l^{**}$ segítségével olyan $\underline{0} \leq \underline{A}^* = \underline{A}^*(\mathcal{D}_m^*, \mathcal{D}_l^{**})$ skálázó vektort megadjunk, amivel $d_{\underline{A}^*}$ minimalizálja az

$$\hat{L}_{m,l}(\phi_{\underline{A}}) = \frac{1}{l} \sum_{i=1}^{l} I_{\left\{\phi_{\underline{A}}(x_{m+i}) \neq y_{m+i}\right\}}$$

hibabecslést. Ezekkel minden megismételhető, amit az optimális súlyvektorok előállításáról elmondtunk. Ezzel a technikával jó metrikákból keverhetünk ki még jobb metrikát. Ugyanis, ha $W_i \subset T_l$ jelöli a teszthalmaz azon részét, melyet a d_i metrikával a legközelebbi társ módszerrel jól osztályozzuk, akkor a $d_{\underline{A}^*}$ metrikával a jól osztályozható tesztpontok halmaza $\bigcup_{i=1}^{p} W_i$ lesz.

A szerző publikációi a témakörben

KETSKEMÉTY L.: A legközelebbi szomszéd gyors megkeresése, Alkalmazott Matematikai Lapok, **21**, (2003)

KETSKEMÉTY L.: A legközelebbi szomszéd módszer hatékonysága automatikus skálázás esetén, Alkalmazott Matematikai Lapok, **21**, (2003)

KETSKEMÉTY L.: Nearest Neighbor Classifying with Excluding Assumption, XXI. Seminar of Stability Problems of stochastic Models, Eger, 2001. 28 januar-3 februar.

KETSKEMÉTY L.: Digitális műholdképek számítógépes klaszterezése, egyetemi oktori disszertáció, 1984, KLTE.

GULYÁS O., KETSKEMÉTY L., KORÁNDI M.: Többsávos műholdképek számítógépes klaszterezése, Idõjárás,88.vf.,No.3.,(1984), 161.-173.

ASZMUSZ,V.V., VADÁSZ V., KARASZEV, A.B., KETSKEMÉTY L.: Adaptivnaja avtomatizirovannaja szisztema inventarizacii szelszkohozjajsztvennüh kultur po koszmicseszkim szmjimkam, Iszledovanija Zemli iz Koszmosza, No.6, (1987), 79-88

ASZMUSZ,V.V., VADÁSZ V., KARASZEV, A.B., KETSKEMÉTY L., SZPIRIMIDONOV,J.G.: Programnüj komplex klaszterizacii mnogozonalnüh dannüh Iszledovanija Zemli iz Koszmosza, No 3, (1988), 86-93

TÄNCZER T., KETSKEMÉTY L., LÉVAI G.: Kísérlet a borultsági viszonyok műholdas meghatározására, Időjárás, 92.évf.,4.sz.,(1988), 242-250

TÄNCZER T., KETSKEMÉTY L., LÉVAI G. Attempt for Estimation of Cloud Amounts Whithin Satellite Pixels in Visible Spectrum, Adv.Space Res.,Vol.9.,No. 7., (1989), 181-184.

KETSKEMÉTY L.: Statistical methods for Preparing the Training Set ZAMM, 70 (1990) 6, 733-735

GULYÁS, O., BARTA GY., SZÁSZ, G., KETSKEMÉTY, L., PRÖHLE, K.: Alakfelismerési módszerek alkalmazása a minõségbiztosításban, Tanulmány (az OMFB támogatásával), (1995)

GULYÁS, O., BARTA GY., SZÁSZ, G., KETSKEMÉTY, L., PRÖHLE, K.: Alakfelismerési programrendszer minõségbiztosítási alkalmazásokhoz, Programdokumentáció (az OMFB támogatásával), (1996)