

Szegedi Tudományegyetem
Mesterséges Intelligencia Kutatócsoport

Kernel-Based Feature Extraction and Speech Technology Applications

PhD értekezés tézisei

Kocsor András

Témavezető:

Dr. Csirik János

Szeged
2003

„The problem of learning is arguably at the very core of the problem of intelligence, both biological and artificial.”

T. Poggio and C. R. Shelton

Bevezetés

Az összefoglaló ismerteti a „Kernel-Based Feature Extraction and Speech Technology Applications” című PhD disszertáció eredményeit. A disszertáció témakörét tágabb értelemben a mesterséges intelligencia, szorosabb értelemben pedig a gépi tanulás képezi.

Az emberiségnek már régóta vágya, hogy intelligens gépet alkosson. Már a múlt század közepén felismerték, hogy az intelligens rendszerek létrehozásának az alapja a tanulás és a döntéshozás [9; 22]. A mesterséges intelligenciában a tanulás a modell-módszer alkalmazását jelenti. Megpróbálunk olyan modelleket létrehozni, amelyek jól szimulálják az emberi intelligenciát. Ennek a legjobb lehetséges módja az, hogy minták sokaságát figyelembe véve specifikáljuk a modellünket és a döntéseinket új minták esetére ezen modell alkalmazásával hozzuk meg. A tudomány különböző területei, a filozófia (mint a tudományok tudománya), a fizika, matematika, biológia, kémia, elméleti számítástudomány mind-mind adalékul szolgálnak azokhoz az eszközökhöz, amelyekből a mesterséges intelligencia kutatói modelleiket építik. Ilyen eszközök például a rövidtávú memória, a hosszútávú memória, a hierarchikus modellfelépítés, a modell-hibridizáció, a klaszterezés, az adat-invariáns módszerek, az optimalizáció és az approximáció. Ennek a disszertációnak az eredményeit is egy nagyon dinamikusan fejlődő terület, a kernel módszerek területe implikálta. Maga a kernel ötlet a matematika [11; 23], illetve a matematikai fizika számos területén felbukkant, mielőtt a gépi tanulási kutatások fókuszába került. Az alapötlet mintafelismerési alkalmazását eredetileg majd 40 éve Aizerman javasolta [2]. A módszer azonban igazán ismertté csak sokkal később, a support vektor gépek publikálásakor vált [4; 27]. Az ötlet alkalmazása nem állt meg, sőt igazából manapság éli fénykorát, sorra jelennek meg újabb és újabb kernel eljárások [14].

Mivel a disszertáció két fő részre tagolódik, az eredményeket is ennek megfelelően két csoportra fogjuk felosztani.

Az eredmények *első csoportját* olyan tulajdonságkinyerő eljárások konstrukciója képezi, amelyek hatékonyabbá tehetik a gépi tanulási feladatok megoldását. A disszertációban négy lineáris módszert foglalunk egy egységes szerkezetbe. Az érintett eljárások közül három – a főkomponens analízis (PCA), a független komponens analízis (ICA) és a lineáris diszkrimináns analízis (LDA) – az irodalomból már jól ismertek, míg a negyedik, a rugó alapú diszkrimináns analízis (SDA) [19; 20], egy újszerű ötletre épül. Felhasználva a módszerek sajátos szerkezetét és azt a nemlinearizációs módszertant, amit a kernel ötlet kínál, kidolgoztuk az ICA, LDA és SDA eljárások nemlineáris változatát. Az így kapott eljárások rendre Kernel-ICA [18], Kernel-LDA [17] és Kernel-SDA [19] néven lettek publikálva. A PCA eljárás kernel alapú nemlinearizációját (Kernel-PCA) 1998-ban Schölkopf javasolta [25]. Voltaképpen ennek a munkának a megismerése inspirálta a disszertáció szerzőjét lineáris módszerek egy csoportjának olyan sajátos egységesítésére, amely lehetővé tette a kernel függvények alkalmazását, és ezáltal a módszerek általánosítását [20].

A tézisek *második csoportját* az első csoportban felsorolt eljárások beszédtechnológiai alkalmazása képezi. Különbőféle fonémafelismerési tesztek végeztünk el az említett tulajdonságkinyerő eljárások hasznosságának demonstrálása céljából az OASIS beszédfelismerő rendszer [15; 16; 20], és a „Beszédmester” beszédjavítás-terápiai és olvasásfejlesztő szoftvercsomag [17–19] keretein belül.

I. Kernel alapú tulajdonságkinyerés

A kernel ötlet

Mercer kernelek. A továbbiakra nézve tegyük fel, hogy $\mathcal{X}$ egy kompakt halmaz az n -dimenziós Euklideszi térben.

1. Definíció. A $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ függvény definíció szerint Mercer kernel akkor és csak akkor, ha teljesíti a következő tulajdonságokat: folytonos, szimmetrikus és pozitív definit.

Függvények folytonosságának és szimmetrikusságának definíciója közismert, azonban a pozitív definitéje talán kevésbé.

2. Definíció. Egy $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ függvény pozitív definit, ha tetszőleges $\{\mathbf{x}_1, \dots, \mathbf{x}_k\} \subset \mathcal{X}$ halmazra a $[\kappa(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^k$ mátrix pozitív szemidefinit.

Annak az eldöntése, hogy egy függvény folytonos vagy szimmetrikus viszonylag egyszerű feladat, azonban a pozitív definité ellenőrzése már távol áll a triviálistól.

Kernel függvények indukálta tulajdonságterek. Most megvizsgáljuk, hogy a Mercer kernelek implicit módon milyen tulajdonságteret indukálnak, és ez hogyan használható ki bizonyos típusú algoritmusok nemlinearizálásához. Először a fő tétel következik [6; 21].

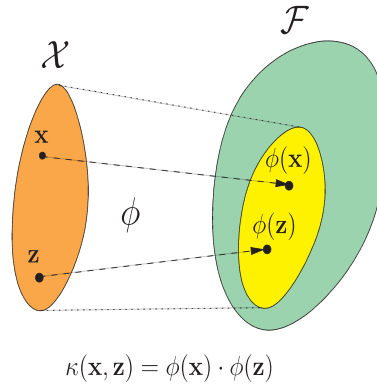
1. Tétel. Tetszőleges $\mathcal{X} \times \mathcal{X}$ feletti κ Mercer kernel esetén létezik egy $\mathcal{F}$ skalárszorzat tér és egy $\phi : \mathcal{X} \rightarrow \mathcal{F}$ leképezés úgy, hogy minden $\mathbf{x}, \mathbf{z} \in \mathcal{X}$ -re $\kappa(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$.

Rendszerint az $\mathcal{F}$ teret kernel-tulajdonságtérnek, a ϕ -t pedig tulajdonságtér-leképezésnek nevezzük. Két következmény azonnal adódik. Amikor ϕ az identikus leképezés, akkor a $\kappa(\mathbf{x}, \mathbf{z}) = \mathbf{x} \cdot \mathbf{z}$ függvény (az $\mathcal{X}$ feletti skalárszorzás) szimmetrikus, folytonos és pozitív definit, amiből következően egy valódi Mercer kernelt alkot. A kernel módszerek szempontjából azonban a másik irány sokkal fontosabb. Ha definiálunk egy Mercer kernelt, akkor feltehetjük egy $\mathcal{F}$ tér létezését, amely felett skalárszorzat számításokat végezhetünk. Ez a tér és a skaláris szorzat számítás ezen tér felett implicit módon definiált a kernel függvényen keresztül (ld. 1-es ábra).

Az 1-es tétel alapján a kernel ötlet lényege a következő módon foglalható össze: *Ha valamely algoritmus outputja kifejezhető Mercer kernelek függvényeként, akkor alternatív algoritmusok konstrukciójára nyílik lehetőség, ha a formulákban az eredetitől eltérő Mercer kerneleket alkalmazunk.*

Lineáris tulajdonságkinyerés

A legtöbb klasszifikációs probléma megoldása során a klasszifikálandó komplex objektumokat célszerű egy dimenzióját tekintve megfelelően nagy tulajdonságtér pontjaival ábrázolni. Ennek a térnek elég szabadsági fokának kell lennie ahhoz, hogy a különböző osztályokhoz tartozó objektumok elégségesen elszeparálhatóak legyenek, azonban a felesleges komponensek megzavarhatják a klasszifikációs algoritmusok működését. Általános gyakorlati tapasztalat, hogy érdemes a tulajdonságtér dimenzióját csökkenteni amennyiben az adatok struktúrája nem sérül. Ennek egy egyszerű lehetősége a lineáris leképezések használata, amikor is az eredeti tulajdonságteret lineárisan leképezzük egy új, rendszerint kisebb dimenziós tulajdonságtérbe. Ennek a transzformációnak nemcsak a dimenziócsökkentés lehet a célja, hanem az adatok struktúrájának kiemelése, világosabbá, egyszerűbbé tétele.



1. ábra. A „kernel ötlet”. A skaláris szorzat az $\mathcal{F}$ kernel-tulajdonságtérben implicit módon definiált.

Bevezetés. Az általánosság megszorítása nélkül feltehetjük, hogy egy véletlen vektorváltozó realizációjaként olyan n -dimenziós valós vektorok állnak elő egy $\mathbb{R}^n$ feletti kompakt $\mathcal{X}$ halmazból, amelyek valamilyen objektumhalmazt jellemeznek. Legyen adott továbbá egy $n \times k$ méretű $X = (\mathbf{x}_1, \dots, \mathbf{x}_k)$ mintamátrix, amely a vektorváltozónk k véletlen megfigyelését foglalja magába.

Az $\mathcal{X}$ tér alkotja az iniciális tulajdonságtérét, az X mátrix pedig az input adatokat jelenti a lineáris tulajdonságkinyerő eljárások számára, amelyek meghatároznak egy $h : \mathcal{X} \rightarrow \mathbb{R}^m$ lineáris függvényt az új tulajdonságok kinyeréséhez. A lineáris leképezés $m \times n$ -es ($m \leq n$) mátrixát, amely dimenzióredukciót is magába foglalhat, V -vel jelöljük. Egy-egy lineáris tulajdonságkinyerő eljárás egy-egy optimális V mátrixot határoz meg, ahol az optimalitás definíciója módszerről-módszerre változik. Habár lehetséges olyan mértékek definiálása is, amelyek a lineáris transzformáció jóságát m irányra (a V mátrix sorvektoraira) egyszerre mérik, mi minden egyes irányt külön-külön egy $\tau : \mathbb{R}^n \rightarrow \mathbb{R}$ függvény segítségével határozzunk meg. Intuitíve, ha a τ függvény nagyobb értékei jobb irányokat jelölnek, és ha az m kiválasztott iránynak valamilyen módon függetlennek kell lennie, akkor az m legnagyobb függvényértékű stacionárius pont kiválasztása egy indokolható eljárás lehet. A stacionárius pontok meghatározása egy általános függvény esetén globális optimalizálással történik. Azonban, ha τ a Rayleigh-hányadossal definiált, azaz

$$\tau(\mathbf{v}) = \frac{\mathbf{v}^\top B_1 \mathbf{v}}{\mathbf{v}^\top B_2 \mathbf{v}}, \quad (1)$$

ahol B_1 és B_2 $n \times n$ -es szimmetrikus mátrixok, B_2 pozitív definit akkor, a megoldás relatíve gyors és egyszerű, mivel sajátérték-sajátvektor problémára vezet.

1. Propozíció. A $\tau(\mathbf{v})$ függvény stacionárius pontjai éppen a $B_2^{-1}B_1$ mátrix sajátvektorai, ahol a megfelelő sajátértékek képezik a stacionárius pontokban felvett függvényértékeket.

A Rayleigh-hányados alapú megközelítés képezi a disszertációban tárgyalt lineáris tulajdonságkinyerő eljárások egységes keretét. Ezek a módszerek a főkomponens analízis (PCA), a független komponens analízis (ICA), a lineáris diszkrimináns analízis (LDA) és a rugó alapú diszkrimináns analízis (SDA). Mivel a 4 módszer közül 2 az ún. felügyelt módszerek családjába tartozik, szükségünk lesz a különböző osztályok címkéire is¹. Legyen r osztályunk, és legyen adott egy $\mathcal{L} : \{1, \dots, k\} \rightarrow \{1, \dots, r\}$ indikátor függvény, ahol $\mathcal{L}(i)$ megadja az $\mathbf{x}_i$ minta osztálycímkejét. Továbbá k_j jelöli a j címkéhez társított mintavektorok elemszámát az X mintamátrixban.

¹A felügyelt és nem felügyelt eljárásokat éppen az különbözteti meg, hogy figyelembe veszik-e az osztályinformációt vagy sem.

A következőkben összefoglaljuk a négy eljárás általános koncepcióját:

- a) A PCA algoritmus olyan független irányokat választ ki a tulajdonságkinyeréshez, amelyek mentén a levetített adatok varianciája nagy [7; 13]. A PCA eljárás esetében az irányok kiválasztására szolgáló τ célfüggvény a következő alakban definiált:

$$\tau(\mathbf{v}) = \frac{\mathbf{v}^T \mathbf{C} \mathbf{v}}{\mathbf{v}^T \mathbf{v}}, \quad (2)$$

ahol

$$\mathbf{C} = E\{(\mathbf{x} - E\{\mathbf{x}\})(\mathbf{x} - E\{\mathbf{x}\})^T\} \quad (3)$$

a minta kovarianciamátrixa. A (3)-as egyenlőségben E a várható értéket jelöli, a (2)-es egyenlőség pedig a $\tau(\mathbf{v})$ függvényt úgy definiálja, mint a $\mathbf{v}/\|\mathbf{v}\|$ vektorra levetített centralizált $\mathbf{x}_1 - E\{\mathbf{x}\}, \dots, \mathbf{x}_k - E\{\mathbf{x}\}$ mintavektorok varianciája.

- b) Az ICA eljárás olyan független irányokat keres, amelyek mentén a minta vetületének eloszlása jelentősen eltér a Gauss-eloszlástól [5]. A módszer azért ilyen irányok kiválasztását tűzi ki céljául, mert ezen irányok mentén az egyes osztályokat esetleg könnyebb elkülöníteni egymástól. Az ilyen típusú irányok kiválasztására számos célfüggvényt definiáltak az irodalomban, amelyek, mint később kiderült, közelítőleg ekvivalens eredményre vezettek. A tézisben mi Hyvärinen megközelítését követtük [12]. Az általa javasolt FastICA algoritmusban az új, független $\mathbf{v}$ irányok kiválasztására a következő τ célfüggvény került bevezetésre:

$$\tau_G(\mathbf{v}) = \left(E\{G(\mathbf{v}^T \mathbf{x})\} - E\{G(\nu)\} \right)^2, \quad (4)$$

ahol $G: \mathbb{R} \rightarrow \mathbb{R}$ egy nem kvadratikus függvény, E az előzőekhez hasonlóan a várható értéket jelöli, ν egy standard Gauss-eloszlású változó, végül $\mathbf{v}^T \mathbf{x}$ a $\mathbf{v}$ irányvektor és az $\mathbf{x}$ minta belső szorzata. A FastICA módszer a $\tau_G(\mathbf{v})$ függvényt egy közelítő Newton iterációs eljárással optimalizálja. Mielőtt azonban az optimalizációra sor kerülne, előfeldolgozást kell végrehajtani az X mintaadatakon. Ez a lépés a centralizációt és az ún. kifehérítést jelenti, ahol az utóbbi a minta kovarianciamátrixának egységmátrixra történő transzformációját foglalja magába. Közismert, hogy a főkomponens analízis a kovarianciamátrixot diagonális alakra hozza [13], aminek egyszerű módosításával az egységmátrix alak könnyen elérhető. Előfeldolgozás után, ha levetítjük a mintát egy tetszőleges egységnormájú $\mathbf{v}$ vektorra, akkor zérus várható értékű és 1 varianciájú mintát kapunk. Sőt az is teljesül, hogy tetszőleges W mátrixszal balról megszorozva az előfeldolgozott mintánkat akkor és csak akkor lesz az új minta kovarianciamátrixa egységmátrix, ha a W mátrix ortogonális. Az előfeldolgozást követő Newton-iteráció meghatároz egy olyan optimális W forogatómátrixot, amely bázisvektorai mentén a levetített mintára kiszámolt τ_G függvényértékek megfelelően nagyok. Mivel az ICA algoritmus lényegében a PCA transzformáció egy módosított változatából és egy azt követő optimális ortogonális transzformációból áll, a PCA algoritmus sajátos továbbfejlesztésének is tekinthető.

- c) Az LDA módszer azokat az irányokat részesíti előnyben, amelyek mentén a különböző osztályok távol helyezkednek el egymástól, miközben az osztályok átlagos varianciája kicsi [9; 10]. Az

LDA eljárás esetében a módszer felügyelt mivoltának köszönhetően a $\tau : \mathbb{R}^n \rightarrow \mathbb{R}$ célfüggvény nemcsak az X mintamátrixtól függ, hanem az $\mathcal{L}$ indikátor függvénytől is. Legyen

$$\tau(\mathbf{v}) = \frac{\mathbf{v}^\top B \mathbf{v}}{\mathbf{v}^\top W \mathbf{v}}, \quad (5)$$

ahol B az osztályok közötti, W pedig az osztályokon belüli szórásmatrix. A B osztályok közötti szórásmatrix azt mutatja, hogy milyen a minta $\mathbf{m}_j$ osztályátlagainak a szórása a totális $\mathbf{m}$ átlagvektorhoz képest:

$$\begin{aligned} B &= \sum_{j=1}^r \frac{k_j}{k} (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^\top \\ \mathbf{m} &= \frac{1}{k} \sum_{i=1}^k \mathbf{x}_i \\ \mathbf{m}_j &= \frac{1}{k_j} \sum_{\mathcal{L}(i)=j} \mathbf{x}_i \end{aligned} \quad (6)$$

A W osztályokon belüli szórásmatrix pedig a j osztályú minták C_j kovarianciamatrixának súlyozott összege:

$$\begin{aligned} W &= \sum_{j=1}^r \frac{k_j}{k} C_j \\ C_j &= \frac{1}{k_j} \sum_{\mathcal{L}(i)=j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^\top. \end{aligned} \quad (7)$$

- d) Az SDA eljárás az egy osztályba eső minták között vonzó, míg a különbözőek között taszító erőket feltételez rugók és antirugók felhasználásával [20]. A modell definiálása után a módszer azokat az irányokat választja ki, amelyek mentén a rendszer potenciális energiája maximális. Legyen most a $\tau(\mathbf{v})$, a rugómodell potenciális energiája a $\mathbf{v}$ irányban, a következő módon definiálva:

$$\tau(\mathbf{v}) = \frac{\mathbf{v}^\top D \mathbf{v}}{\mathbf{v}^\top \mathbf{v}}, \quad (8)$$

ahol

$$D = \sum_{i,j=1}^k (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top [M]_{ij} \quad (9)$$

és

$$[M]_{ij} = \begin{cases} -1, & \text{ha } \mathcal{L}(i) = \mathcal{L}(j) \\ 1, & \text{különben} \end{cases} \quad i, j = 1, \dots, k. \quad (10)$$

Természetesen az M mátrix elemeit ± 1 -től eltérően is inicializálhatjuk. Ez a mátrix tekinthető a rugóegyütthatók mátrixának, amely akár mintapáronként is különböző lehet.

Nemlineáris tulajdonságkinyerés kernel függvényekkel

A tulajdonságkinyerés vagy lineáris, vagy nemlineáris, de a kernel ötlet bizonyos értelemben eltörli a határt közöttük. Ahogy már láttuk, ha valamely lineáris eljárás csak az input vektorok páronként vett skaláris szorzatát használja fel a számítások során, akkor a skalárszorzat alkalmas megváltoztatásával a módszer egy nemlineáris változatát definiálhatjuk. A műveletcsere hatása az, hogy az eredeti lineáris algoritmus egy másik, talán nagyobb szabadsági fokot megengedő belső szorzat térben fog végrehajtódni.

A következőkben ezt az ötletet használjuk fel a PCA, ICA, LDA és SDA módszerek nemlineáris változatainak megadásához. Ehhez legyen a belső szorzat egy κ Mercer kernel által definiált, ami

nemlineáris módon indukál egy $\mathcal{F}$ teret a ϕ leképezéssel (ld. 1-es Tétel). Először vizsgáljuk meg a Rayleigh-hányados (1)-ben definiált alakját az $\mathcal{F}$ térben. Formálisan,

$$\tau(\mathbf{v}) = \frac{\mathbf{v}^\top \mathcal{B}_1 \mathbf{v}}{\mathbf{v}^\top \mathcal{B}_2 \mathbf{v}}, \quad \mathbf{v} \in \mathcal{F} \quad (11)$$

ahol $\mathcal{B}_1$ és $\mathcal{B}_2$ szimmetrikus $\dim(\mathcal{F}) \times \dim(\mathcal{F})$ méretű mátrixok, $\mathcal{B}_2$ pozitív definit. Sajnos az előbbi formula még nem elég specifikus ahhoz, hogy $\tau(\mathbf{v})$ a κ Mercer kernellel kifejezhető legyen.

Vegyük észre, hogy a korábban definiált lineáris tulajdonságkinyerő módszereknél a számlálóban elhelyezkedő mátrix az X mintamátrix sajátos függvénye. Minden esetben a következő általános alakot ölti:

$$X \Theta X^\top = \sum_{i=1}^j [\Theta]_{ij} \mathbf{x}_i \mathbf{x}_j^\top, \quad [\Theta]_{ij} \in \mathbb{R} \quad (12)$$

(ld. (2)-es, (5)-ös és (8)-as formulák), ahol a Θ egy módszerspecifikus valós szimmetrikus mátrix. A nevezőben megtalálható mátrixok pedig a PCA, ICA, SDA esetében az egységmátrix, míg az LDA-nál szintén a (12)-es egyenlőségben definiált formában áll elő (ld. (7)). Ezen megfigyelések alapján a lineáris módszerek Rayleigh-hányadosa a következő alakú:

$$\frac{\mathbf{v}^\top X \Theta_1 X^\top \mathbf{v}}{\mathbf{v}^\top (X \Theta_2 X^\top + \delta I) \mathbf{v}}, \quad (13)$$

ahol $\mathbf{v} \in \mathcal{X}$, X a mintamátrix, Θ_1, Θ_2 módszerfüggő $n \times n$ -es valós szimmetrikus mátrixok és $\delta \in \mathbb{R}_+$. Az LDA esetében $\delta = 0$, a PCA, ICA és SDA módszereknél $\delta = 1$ és Θ_2 a zérómátrix.

Most már formalizálhatjuk a (13)-as kifejezést a kernel tulajdonságtérben azáltal, hogy az $X = (\mathbf{x}_1, \dots, \mathbf{x}_k)$ mátrixot $F = (\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_k))$ -val és a $\mathbf{v} \in \mathcal{X}$ vektort $\mathbf{v} \in \mathcal{F}$ -el helyettesítjük:

$$\tau(\mathbf{v}) = \frac{\mathbf{v}^\top F \Theta_1 F^\top \mathbf{v}}{\mathbf{v}^\top (F \Theta_2 F^\top + \delta I) \mathbf{v}}. \quad (14)$$

Most tekintsük $\tau(\mathbf{v})$ stacionárius pontjait.

2. Propozíció. $\tau(\mathbf{v})$ stacionárius pontjaira $\mathbf{v} \in \text{SPAN}(\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_n))$ teljesül.

Azaz feltehetjük, hogy $\mathbf{v} = \alpha_1 \phi(\mathbf{x}_1) + \dots + \alpha_n \phi(\mathbf{x}_n) = F \boldsymbol{\alpha}$, amivel (14) alapján a következő $\boldsymbol{\alpha}$ -tól függő Rayleigh-hányadost kapjuk:

$$\tau(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^\top F^\top F \Theta_1 F^\top F \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top F^\top (F \Theta_2 F^\top + \delta I) F \boldsymbol{\alpha}} \quad (15)$$

És mivel $F^\top F$ megegyezik a $K = [\kappa(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^k$ kernelmátrixszal, kapjuk, hogy

$$\tau(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^\top K \Theta_1 K \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top (K \Theta_2 K + \delta K) \boldsymbol{\alpha}}, \quad (16)$$

ahol az 1-es propozíciónak megfelelően a

$$(K \Theta_2 K + \delta K)^{-1} K \Theta_1 K \quad (17)$$

mátrix sajátvektorai adják a keresett stacionárius pontokat. Ha az $\mathcal{A}$ mátrix sorvektorait a domináns sajátértékekhez tartozó sajátvektorokkal definiáljuk, akkor a kernel alapú tulajdonságkinyerés tetszőleges $\mathbf{z} \in \mathcal{X}$ mintára $\mathcal{A} F^\top \phi(\mathbf{z})$ -vel számítható. Mivel azonban ϕ explicit módon nem ismert, ez a kifejezés ebben a formában nem kiértékelhető, de ha észrevesszük, hogy $F^\top \phi(\mathbf{z}) = (\kappa(\mathbf{x}_1, \mathbf{z}), \dots, \kappa(\mathbf{x}_k, \mathbf{z}))^\top$, akkor már kernel függvények kifejezését kapjuk.

A módszerek egységes matematikai keretrendszerének tárgyalása után rátérünk az egyes módszerek formuláira, illetve összefoglaljuk a négy eljárás általános koncepcióját:

- a) A Kernel-PCA [25] eljárás azokra a nemlineáris irányokra koncentrál, amelyek mentén a minta varianciája nagy (ld. [16; 20]). A Kernel-PCA esetében a τ célfüggvény a következő alakban áll elő:

$$\tau(\alpha) = \frac{\alpha^\top \frac{1}{k} K(I - \hat{I}) K \alpha}{\alpha^\top K \alpha}, \quad (18)$$

ahol $[K]_{ij} = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ a kernelmátrix és

$$\hat{I} = \frac{1}{k} \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}. \quad (19)$$

- b) A Kernel-ICA módszer nemlineáris módon olyan független irányokat keres, amelyek mentén a minta eloszlása jelentősen eltér a Gauss-eloszlástól [18]. Ahogy korábban láttuk, a Hyvärinen-féle ICA algoritmus két fő részből áll: centralizáció és kifehérítés, illetve az azt követő közelítő Newton eljárás. A tézisben mi csak az első, az előfeldolgozó fázist nemlinearizáltuk, mivel ezután már az adataink nemlineáris módon korrelálatlanokká válnak az $\mathcal{F}$ kernel tulajdonságtérben. Habár megtehető,² de a második iteratív rész nemlineáris kiterjesztését nem végeztük el.
- c) A Kernel-LDA nemlineáris módon próbálja elérni az osztályok jó szeparációját és ezáltal nyer ki új tulajdonságokat [17]. A különböző osztályok elemeit eltávolítja, amíg az ugyanazon osztályba tartozókat közelíti egymáshoz. A Kernel-LDA $\tau(\alpha)$ célfüggvénye:

$$\frac{\alpha^\top K(R - \hat{I}) K \alpha}{\alpha^\top K(I - R) K \alpha}, \quad (20)$$

ahol K a kernelmátrix, $\hat{I}$ -t a (19)-es formula definiálja, továbbá

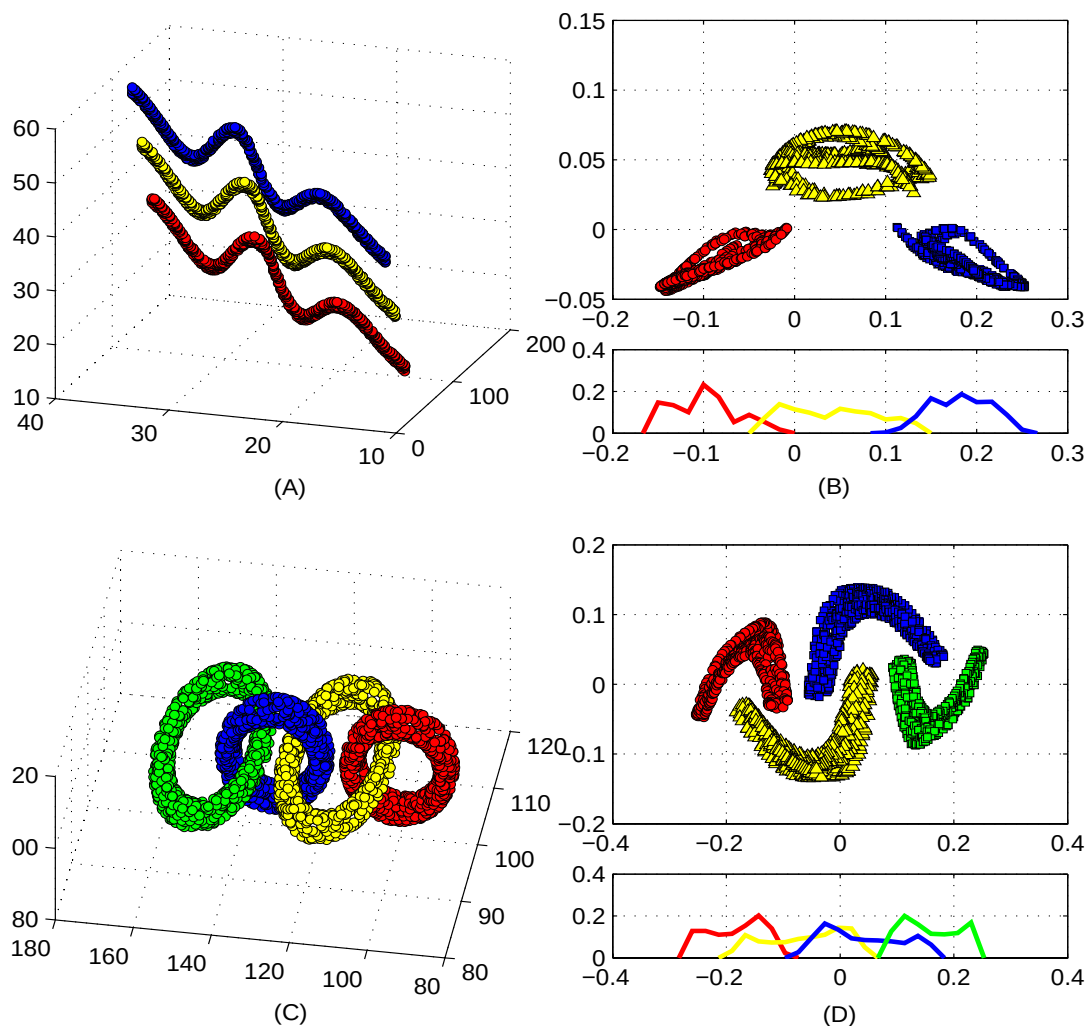
$$[R]_{ij} = \begin{cases} \frac{1}{k_t} & \text{ha } t = \mathcal{L}(i) = \mathcal{L}(j) \\ 0 & \text{különben.} \end{cases} \quad (21)$$

- d) A Kernel-SDA nemlineárisan képezi le a tulajdonságtérre, és az új térben a Kernel-LDA-hoz hasonlóan osztályszeparációt végez vonzó és taszító erők definiálásával (ld. 2-es ábra) [19]. A Kernel-SDA módszer Rayleigh-hányadosa:

$$\tau(\alpha) = 2 \frac{\alpha^\top K(\tilde{M} - M) K^\top \alpha}{\alpha^\top K \alpha}, \quad (22)$$

ahol K a kernelmátrix és $\tilde{M}$ egy olyan diagonális mátrix, amelynek az elemeit a (10)-es egyenlőségben definiált M rugóegyüttható mátrix sorvektorainak összege adja.

²Nyilvánvalóan a (4)-es kifejezés könnyen nemlineárisra tehető kernelekkel, hiszen a formulában egy $\mathbf{v}^\top \mathbf{x}$ skalárszorzat szerepel. Azonban ezt a lépést elvetjük, mert ez a Rayleigh-hányados alapú megközelítéstől nagyon eltérne.



2. ábra. A Kernel-SDA módszer hatása 3-dimenziós adatokon. Az (A) és a (C) ábrán megjelenített mintákra alkalmazott Kernel-SDA eljárás eredménye a (B) és (D) ábrákon látható.

I/1. Tézis

A szerző egy olyan egységes matematikai keretet definiált lineáris tulajdonságkinyerő algoritmusok egy halmazához, amely egy, a gépi tanulási kutatások fókuszában álló ötlet, a kernel ötlet alkalmazásával lehetővé tette ezen eljárások nemlineáris változatának kidolgozását. A disszertációban 8 tulajdonságkinyerő eljárást, 4 lineáris (PCA, ICA, LDA, SDA) és 4 nemlineáris módszert (Kernel-PCA, Kernel-ICA, Kernel-LDA, Kernel-SDA) mutat be egy egységes megközelítésben a Rayleigh-hányados optimalizálásával [15–20].

I/2. Tézis

Az irodalomból korábban ismert 3 lineáris módszer (PCA, ICA, LDA) kiegészítéseként a szerző megkonstruált egy újszerű lineáris eljárást az SDA-t [20].

I/3. Tézis

A szerző megadta az ICA, LDA és SDA lineáris eljárások nemlineáris változatát, amelyek eredményeként 3 további új algoritmus, a Kernel-ICA [18], Kernel-LDA [17] és Kernel-SDA [19] jött létre.

II. Beszédtechnológiai alkalmazások

Beszédfelismerés

Az automatikus beszédfelismerés egy olyan mintafelismerési probléma, amelynek a célja az ember beszédfeldolgozási képességének modellezése. Ezért nyilvánvalóan a gépi tanulással és mesterséges intelligenciával foglalkozó tudomány részét képezi. Tradicionális okokból azonban az elektromérnöki tudományok részterületeként szokás megjelölni. Az utóbbi néhány évtizedben a beszédfelismerés domináns technológiája a rejtett Markov modell (HMM) alapú megközelítés volt. Eközben viszont a gépi tanulás elmélete sokat fejlődött és számos új tanuló és klasszifikációs eljárás vált elérhetővé [3; 8; 10]. A disszertációban a beszédfelismerési fejezetek célja a bevezetett tulajdonságkinyerő eljárások alkalmazhatóságának vizsgálata a fonémafelismerés feladatán.

Amikor beszédfelismerési kutatásaink irányát megválasztottuk, elhatároztuk, hogy olyan flexibilis rendszert fejlesztünk, amely lehetővé teszi a kísérletezést különféle előfeldolgozó, tulajdonságkinyerő és gépi tanuló algoritmusokkal. Ezek az elvárások elvezettek egy sztochasztikus szegmentális beszédfelismerő rendszer, az OASIS [26] kifejlesztéséhez, amelynek a moduljai egy sajátos script nyelv segítségével vezérelhetők. A moduláris felépítés és a magas szintű vezérlési lehetőség intenzív kutatómunkát tesz lehetővé a beszédfelismerés területén.

A disszertációban az OASIS beszédfelismerő rendszer keretein belül szegmentális fonémafelismerési tesztek hajtottunk végre [15; 16; 20]. A tesztekben azt vizsgáltuk, hogy a tulajdonságkinyerő algoritmusok kombinálva különféle klasszifikációs algoritmusokkal (Timbl, OC1, C4.5, GMM, ANN) hogyan befolyásolják a felismerési pontosságot.

A nagyszámú teszt eredményének ismeretében kijelenthetjük, hogy a hatékonyabb klasszifikáció reményében tanulás előtt érdemes tulajdonságkinyerő algoritmusokat alkalmazni.

A fonológiai tudatosság tanítása

Az alfabetikus nyelvek esetében az olvasástanulás folyamatában nagyon fontos a szavakat alkotó egymást követő hangok szeparálásának és azonosításának képessége, valamint a beszédhangok és írásjelek helyes asszociációja [1; 24]. Ahhoz, hogy az olvasástanulás eredményes legyen, a tanulóknak valamiféle fonológiai tudatosságnak kell kialakulni, sőt a gyerekeknek képesnek kell lenniük manipulálni is ezeket. A legtöbb tanulási nehézségekkel küzdő gyermeknek problémája van a fonológiai információk feldolgozásával. Mindezek mellett a fonológiai tudatosság kialakításának és tanításának szintén nagy szerepe van a siket gyermekek beszédjavítás-terápiájában is.

Az általunk kifejlesztett „Beszédmester” szoftvercsomag (ld. 3-as ábra) beszédfelismerési technológiákat alkalmaz olvasásfejlesztésre és beszédjavítás-terápiára (ld. IKTA-055/2001 számú projekt). A beszédfelismerés feladata egy vizuális fonetikai visszacsatolás megvalósítása egy megbízható valós idejű fonémafelismerő segítségével [17]. A képernyőn a kiejtés pillanatában a felismerő kimenete alapján megjelenik egy betű, amelynek a fényessége éppen a felismerésének valószínűségével arányos. Míg a beszédjavítás-terápiában a siket gyermekek hiányzó auditív visszacsatolásának helyettesítése a cél (3B. ábra), addig az olvasástanításban a fonéma-graféma párok asszociációjának megerősítése (3A. ábra). A számítógéppel segített beszédterápia lehetősége jelentős, hiszen a tanár állandó jelenléte nem szükséges hozzá. Rendszerint a siket gyermekek esetében nagyon hosszú terápiás folyamat



(A)



(B)

3. ábra. A „Beszédmester” olvasásfejlesztő és beszédjavítás-terápiai szoftvercsomag.

eredményez előrehaladást, amelyet a „Beszédmester” akár jelentősen is meggyorsíthat. Továbbá a tapasztalatok azt mutatják, hogy a gyermekek előnyben részesítik a számítógépes munkát a hagyományos feladatokkal szemben.

A disszertációban, a „Beszédmester” programcsomag keretein belül elvégzett tesztek során szintén azt vizsgáltuk, hogy a javasolt tulajdonságkinyerő eljárások milyen hatást fejtenek ki klasszifikációs algoritmusok egy halmazára (ANN, PPL, GMM, SVM), a felismerési pontosság tekintetében [17–19]. Az eredmények azt mutatták, hogy a transzformációk közül a nemlineárisak rendszerint kisebb klasszifikációs hibát eredményeztek. A felügyelt és nem felügyelt módszerek viszonylatában pedig a felügyelt eljárások voltak sikeresebbek. Mindezek az eredmények nagyban hozzájárultak, hogy a „Beszédmester” szoftverünk hatékonyabb és megbízhatóbb valós idejű fonémaklasszifikációt, és ezáltal eredményesebb terápiát végezhesen.

II/1. Tézis

A szerző megtervezett, és munkatársaival együtt végrehajtott az OASIS beszédfelismerő rendszer keretében olyan szegmens-alapú fonémafelismerési teszteket, amelyek demonstrálják a kifejlesztett tulajdonságkinyerő eljárások hatását a felismerési pontosság tekintetében [15; 16; 20]. A szerző saját munkáját képezi a tervezésen kívül a tulajdonságkinyerő eljárások implementálása és futtatása is.

II/2. Tézis

A Beszédmester beszédjavítás-terápiai, olvasásfejlesztő és olvasásterápiai rendszer valós idejű fonémafelismerési hatékonyságának megnövelése érdekében a szerző további felismerési teszteket tervezett meg és végzett el munkatársaival. A munkavégzés itt is az előbbi tézispontban leírt munkavégzési kondícióknak megfelelően történt³.

³Noha az elmúlt évek során mind az OASIS, mind a „Beszédmester” rendszerek esetében a szerző projekt- és témavezetőként vett részt a munkálatokban, a két rendszert magát nem sorolja a disszertáció eredményei közé, hiszen ezekben az esetekben a munkatársakkal közös és oszthatatlan eredményekről van szó.

sorszám	PCA	ICA	LDA	SDA	Kernel-PCA	Kernel-ICA	Kernel-LDA	Kernel-SDA	keretrendszer
[15]	•	•	•						OASIS
[16]	•				•				OASIS
[17]			•				•		Beszédmester
[18]		•				•			Beszédmester
[19]								•	Beszédmester
[20]	•	•	•	•	•	•	•	•	OASIS

1. táblázat. A táblázatból leolvasható a disszertáció téziseinek, valamint a publikált eljárásoknak és azok alkalmazásának a viszonya.

Konklúzió

A disszertációban szereplő lineáris tulajdonságkinyerő eljárások, illetve ezek kernel függvényeket használó nemlinearizált változata jól demonstrálja, hogy talán matematikai értelemben nem is olyan nagy a különbség a lineáris és a nemlineáris modellek között. Pontosabban a nemlineáris modellek egy halmaza lineáris, de egy másik térben.

A beszédtechnológiai alkalmazások eredményeiből kiderült, hogy a hatékonyság növelése érdekében érdemes a felismerés problémáját két részre bontani: először tulajdonságkinyerésre, majd tanulásra. Noha mind a tulajdonságkinyerést végző, mind a tanulást megvalósító matematikai modellek célja a szeparáció, a végcél több lépésben történő elérése sokszor célravezetőnek bizonyul.

Végül az 1-es táblázat összefoglalja, hogy a disszertáció eredményeihez kapcsolódó publikációk milyen tulajdonságkinyerő eljárásokat ismertetnek, és hogy az alkalmazások melyik keretrendszer felhasználásával készültek.

Hivatkozások

- [1] M. J. Adams, Beginning to read: Thinking and learning about print, Cambridge, MA: MIT Press, 1990.
- [2] M. A. Aizerman, E. M. Braverman, L. I. Rozonoer, "Theoretical foundation of the potential function method in pattern recognition learning," Automat. Remote Cont., Vol. 25, pp. 821-837, 1964.
- [3] C. M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press Inc., New York, 1996.
- [4] B. E. Boser, I. M. Guyon, V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in Proc. of the Fifth Annual ACM Conference on Computational Learning Theory, D. Haussler (eds.), ACM Press, Pittsburg, pp. 144-152, 1992.
- [5] P. Comon, "Independent component analysis, A new concept?" Signal Processing, Vol. 36, pp. 287-314, 1994.
- [6] F. Cucker, S. Smale, "On the mathematical foundations of learning," Bull. Am. Math. Soc., Vol. 39, pp. 1-49, 2002.
- [7] K. I. Diamantaras, S. Y. Kung, Principal Component Neural Networks: Theory and Applications, John Wiley, New York, 1996.
- [8] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, John Wiley & Sons, New York, 2001.

- [9] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, Vol. 7, pp. 179-188, 1936.
- [10] K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, New York, 1989.
- [11] D. J. Hand, *Kernel discriminant analysis*, Research Studies Press, New York, 1982.
- [12] A. Hyvärinen, E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, Vol. 9, No. 7, pp. 1483-1492, 1997.
- [13] I. J. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, 1986.
- [14] Kernel Machines Web site, <http://kernel-machines.org>.
- [15] A. Kocsor, L. Tóth, A. Kuba Jr., K. Kovács, M. Jelasity, T. Gyimóthy, J. Csirik, "A Comparative Study of Several Feature Space Transformation and Learning Methods for Phoneme Classification", *International Journal of Speech Technology*, Vol. 3, No. 3/4, pp. 263-276, 2000.
- [16] A. Kocsor, A. Kuba Jr., L. Tóth, "Phoneme Classification Using Kernel Principal Component Analysis", *Periodica Polytechnica*, Vol. 44, No. 1, pp. 77-90, 2000.
- [17] A. Kocsor, L. Tóth, D. Paczolay, "A Nonlinearized Discriminant Analysis and its Application to Speech Impediment Therapy", in: V. Matousek, P. Mautner, R. Moucek, K. Tauser (Eds.): *Proceedings of Text, Speech and Dialogue: 4th International Conference, TSD 2001, LNAI 2166*, pp. 249-257, Springer Verlag, 2001.
- [18] A. Kocsor, J. Csirik, "Fast Independent Component Analysis in Kernel Feature Spaces", in: L. Pacholski and P. Ruzicka (Eds.): *Proceedings of SOFSEM 2001: Theory and Practice of Informatics: 28th Conference on Current Trends in Theory and Practice of Informatics, LNCS 2234*, pp. 271-281, Springer Verlag, 2001.
- [19] A. Kocsor, K. Kovács, "Kernel Springy Discriminant Analysis and Its Application to a Phonological Awareness Teaching System", in: P. Sojka, I. Kopecek, K. Pala (Eds.): *Proceedings of Text, Speech and Dialogue: 5th International Conference, TSD 2002, LNAI 2448*, pp. 325-328, Springer Verlag, 2002.
- [20] A. Kocsor, L. Tóth, "Application of Kernel-Based Feature Space Transformations and Learning Methods to Phoneme Classification", accepted for *Applied Intelligence*.
- [21] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Philos. Trans. Roy. Soc. London, A*, Vol. 209, pp. 415-446, 1909.
- [22] J. V. Neumann, O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, 1947.
- [23] E. Parzen, "On estimation of probability density function and mode", *Annals of Mathematical Statistics*, Vol. 33, pp. 1065-1076, 1962.
- [24] D. J. Sawyer, B. J. Fox, *Phonological Awareness in Reading: The Evolution of Current Perspectives* (Springer Series in Language and Communication, Vol 28), Springer-Verlag, New York, 1991.
- [25] B. Schölkopf, A. J. Smola, K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, Vol. 10, pp. 1299-1319, 1998.
- [26] L. Tóth, A. Kocsor, K. Kovács, "A Discriminative Segmental Speech Model and its Application to Hungarian Number Recognition," in: P. Sojka, I kopecek, K. Pala (Eds.), *Proceedings of Text, Speech and Dialogue: 3th International Conference, TSD 2000, LNAI 1902*, pp. 307-313, Springer Verlag, 2000.
- [27] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons Inc., 1998.