Szegedi Tudományegyetem
Informatikai Tanszékcsoport

Various Kernel Methods with Applications

PhD értekezés tézisei

Kovács Kornél

Témavezető:

Dr. Kocsor András

Szeged
2008

Bevezetés

A modellalkotás a tudomány talán legfontosabb komponense, amely absztrakt módon minden tudományágban azonos. Modelleket alkotunk és megvizsgáljuk azok működését, módosításokat végzünk rajta mindaddig, amíg el nem érjük kitűzött célunkat. A mesterséges intelligencia egy különös tudományág, amely rendkívül jó táptalajt ad a számítógépes modellalkotás szerelmeseinek.

A gépi tanulás a mesterséges intelligencia tudományterület részét képezi [25]. Olyan algoritmusok konstrukcióját foglalja magában, amelyek a szamítógépet "tanulási" képességekkel ruházzák fel. A tanulásnak rendszerint két különböző megközelítése van: induktív és deduktív. A tézis az induktív irányzatot követi, amely rendszerint szabályokat vagy deszkriptív mintákat nyer ki masszív adathalmazokból (a tézisben az általánosítást a statisztikai megközelítés módszertanával végezzük). A gépi tanulás fókuszában jelen pillanatban leginkább az automatikus információ kinyerés összetett feladata áll. Ezen kívül fontos alkalmazásokat jelent - a teljesség igénye nélkül - a természetes nyelvfeldolgozás, a szintaktikus mintafelismerés, kereső motorok javítása, orvosi diagnosztika, bioinformatika, beszédfelismerés, tárgyak azonosítása és például számítógépes játékok intelligenciával történő felruházása. Bizonyos gépi tanulási módszerek az embert próbálják kiiktatni az adatanalízis folyamatából, míg más eljárások éppen az ember gép interakciót próbálják emberibbé tenni.

A gépi tanulás, a jelen kor technológiai szintje mellett a mesterséges intelligencia leginkább kutatott és legrelevánsabban fejlődő területévé vált. A dolgozat a gápi tanulás legújszerűbb módszertanához a kernel módszerek világához kapcsolódó eljárások konstrukciójával és vizsgálatával foglalkozik.

A kernel módszerek (KM) a mintafelismerés algoritmusainak egy olyan családja [26], amelynek legjelentősebb tagja a Support Vector Machine (SVM) [31]. A mintafelismerés általános feladata nem más, mint reprezentatív összefüggések keresése és tanulmányozása (például klaszterek, korrelációs összefüggések, klasszifikációs döntések vizsgálata) általános adatokon (vektorok, dokumentumok, szekvenciák, képek, stb.)

A KM megközelítés a nevét a kernel függvényekről kapta, amelyek egy olyan származtatott tulajdonság térben dolgoznak, ahol a minták tényleges koordinátáit soha nem kell kiszámolni. A módszerek csak a mintapontok páronként vett skalárszorzatára támaszkodnak, amelyeket implicit módon a kernel függvények alkalmazásával számítanak ki.

A kernel modszerek csaladjába beletartozik az SVM-en kívül számos más algoritmus: különböző regressziós eljárások, a Fisher-féle lineáris diszkrimináns analízis (LDA) [9], a főkomponens analízis (PCA) [10], a kanonikus korreláció analízis (CCA) [2], a 'ridge' regresszió [22], a spektrális klaszterezés [21], és még sok más eljárás. Általánosságban elmondható, hogy a kernel módszerek többsége hatékonyan megoldható feladatokra, konvex optimalizációra vagy sajátérték-sajátvektor problémára vezetnek.

A 'kernel' ötlet

Legyen adott az (X, y) objektumkettes, mint tanuló adathalmaz. Jelölje az input mintákat az $X = (x_1, \ldots, x_n)$ $(x_j \in \mathbb{R}^d)$, a megfelelő osztálycímkéket pedig y. Klasszifikáció esetén $y \in \{-1, +1\}^n$, regressziós problémáknál pedig $y \in \mathbb{R}$. Tegyük fel továbbá, hogy az (x_i, y_i) minta-osztály párok $(i = 1, \ldots, n)$ független azonos eloszlású véletlen változók.

A hatékony klasszifikáció és regresszió fontos előkészítő lépése az adathalmaz releváns jellemzőinek megtalálása. Ez sok esetben könnyebben elvégezhető, ha az adatokat leképezzük egy megfelelően nagy dimenziós térbe egy alkalmas ϕ nemlineáris leképezés segítségével, majd a transzformált adatokon lineáris algoritmusokat alkalmazunk. Ha valamely algoritmus felépíthető elemi skalárszorzat műveletek alkalmazásával és ha a

$$\phi : \mathbb{R}^d \to \mathcal{H} \tag{1}$$

nemlineáris leképezés olyan, hogy az x_1 és x_2 pontpárok skalárszorzata a ϕ leképezés mellett kiszámítható x_1 és x_2 függvényeként $\mathrm{poly}(d)$ időkomplexitás mellett anélkül, hogy $\phi(x_1)$-et és $\phi(x_2)$-t explicit meghatároznánk, akkor az algoritmus hatékonyan kivitelezhető marad függetlenül $\mathcal{H}$ dimenziójától.

Voltaképpen ez a gondolatmenet teszi lehetővé, hogy nagyon magas vagy akár végtelen dimenziójú képtereket használjunk a klasszifikációs, illetve regressziós problémák megoldásakor. Először választanunk kell egy alkalmas pozitív definit, szimmetrikus függvényt

$$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}. \tag{2}$$

Ezeket a függvényeket – amelyek tulajdonképpen a skalárszorzat művelet implicit elvégzésére alkalmasak a képtérben – kernel függvényeknek nevezzük [4]. Ekkor a

$$\{\, k(x, \cdot) \,|\, x \in \mathbb{R}^d \,\} \tag{3}$$

függvényhalmaz által kifeszített lineáris altér lezártja Hilbert teret alkot a következő belső szorzat definícióval [20]:

$$\langle k(x_1, \cdot), k(x_2, \cdot) \rangle = k(x_1, x_2), \quad x_1, x_2 \in \mathbb{R}^d. \tag{4}$$

Azaz a k kernel függvény megválasztása automatikusan generálja a $\phi : \mathbb{R}^d \to \mathcal{H}$ leképezést a $\phi(x) = k(x, \cdot)$ definícióval.

Ezt a sémát kernel ötletnek szokás nevezni a gépi tanulásban [1; 23; 30] és ez az a módszertan, amely köré a tézisben vázolt újszerű eljárások kötődnek.

Eredmények

A disszertáció tézisei lényegében két különböző módon két-két csoportra oszthatók. Az egyik osztályozás szerint a szerző eredményei a gépi tanulás témakörének kernel-alapú tulajdonságkinyerő és klasszifikációs módszereinek tárgykörébe esik. A másik felosztás szerint pedig algoritmikus konstrukciókról és gyakorlati alkalmazásokról beszélhetünk. A következőkben követve a disszertáció felépítését az első felosztás szerint vesszük számba az elért eredményeket. Fontosnak tartjuk megjegyezni, hogy az alább részletezett eredménylista a kapcsolódó cikkek nóvumainak csak azon részét katalogizálja, ahol a disszertáció szerzőjének hozzajárulása volt domináns.

Az eredmények első csoportját a szerző Kernel alapú tulajdonságkinyerő módszerei képezik. Ezen eredményeket a 2-es és 3-as fejezetek írják le a tézisben.

I/1. A szerző kidolgozta az MMDA algoritmus direkt változatát [11; 16]. A módszer egy olyan tulajdonságkinyerő eljárás, amely megnövelheti a klasszifikációs módszerek hatékonyságát. Az UCI gépi tanulási adatbázis számos példáján sikerült bizonyítani az eljárás használatának létjogosultságát [3].

I/2. Az MMDA algoritmus a motivációjából következően alkalmas nagydimenziós tulajdonságterek redukálására a nagyobb klasszifikációs hatékonyság növelése érdekében. A szerző megkonstruálta a módszer arcfelismerésre kidolgozott módosított változatát. A FERET 'gold standardot' képező arcfelismerési adatbázis [6] segítségével bizonyította a bevezetett eljárás eredményességét. Az irodalomban fellelhető eredményeket sikerült több ponton jelentősen meghaladni [16].

I/3. A tulajdonságkinyerés fókuszában a klasszifikáción túl a regresszió is állhat. A szerző kidolgozta az MMDA algoritmus regressziós problémák megoldására adaptált változatát, korrelációmentes tulajdonságok kinyerésére. A módszer neve Kernel Decorrelated Learning Regression (KDLR) [28]. Standard regressziós feladatokon történő teszt alapján kimondható, hogy az eljárás a gyakorlatban hatékony regresszióhoz vezethez.

I/4. A szerző javaslatot tett a statisztikából ismert átlagos deriváltbecslő eljárás és a kernel függvények alkalmazásának kombinációjára. A Kernel Average Derivative Estimation-nek (KADE) elnevezett eljárás célja a regresszió szempontjából releváns alterek megtalálása [28]. Mesterséges adatokon történő teszteléssel és a kapcsolódó algoritmusokkal való összehasonlítással a szerző kimutatta, hogy a megtalált alterek számos esetben hatékonyabb regressziót tesznek lehetővé.

A tézis eredményeinek második csoportjába újszerű Kernel alapú klasszifikációs algoritmusok tartoznak. Az eredményeket a 4-es és 5-ös fejezetek részletezik.

II/1. A szerző definiálta hipersík alapú klasszifikációs módszerek egy családját [13]. Három geometriai megfontolásokat követő módosítást/konstrukciót javasolt. i) különféle veszteségfüggvényeket használt a hipersik alapú klasszifikációban. ii) a lineáris regressziót sajátos módon alkalmazta klasszifikációhoz. iii) az output teret beágyazta az input térbe és kidolgozta a Minor Component Classifier (MCC) módszert, amely egy a mintapontokbol számított mátrix legkisebb sajátértékéhez tartozó sajátvektora segítségével definiál klasszifikációs hipersíkot. A szerző kialakította a módszerek tesztelési környezetét, majd végrehajtotta a működést demonstráló teszteket. Az eredmények azt igazolják, hogy az SVM-el [26] kompetitív eljárások kialakítására került sor.

II/2. A szerző megkonstruált egy klasszifikációs sémát az ún. 'Convex Machine' technikát, amely bázisfügvények ritka kombinációját alkalmazza. A kidolgozott módszertan magában foglal számos gépi tanulási technikát. A teljesség igénye nélkül ezek közé tartozik a Support Vector Machine (SVM) [26], a Smooth Support Vector Machine [18], a Least Square Support Vector Machine (LSVM) [27] es a Kernel Logistic Regression (KLR) [8]. Kialakított továbbá három alapvető numerikus matematikai módszerek által inspirált bázisfüggvény-kiválasztási módszert (RANDOM, MGRAMM, CORR) [14], amelyeket tesztelt az UCI adatbazis [3] egyes elemein.

Tézis eredmények	[11]	[13]	[14]	[15]	[16]	[28]	Fejezet	Eredmény kategóriája
MMDA	•						2	Tulajdonságkinyerés
MMDA arcfelismerő	•				•		2	Tulajdonságkinyerés
KDLR	•					•	3	Tulajdonságkinyerés
KADE						•	3	Tulajdonságkinyerés
Hipersík alapú klasszifikáció		•					4	Klasszifikáció
Konvex gépek			•	•			5	Klasszifikáció
Alapvető bázisszelekciós eljárások			•				5	Bázisfüggvény-szelekció
Komplex bázisszelekciós eljárások				•			5	Bázisfüggvény-szelekció

1. táblázat. A tézisek és a kapcsolódó publikációk összefüggése.

II/3. A szerző kidolgozott három komplexebb bázisfüggvényszelekciós módszert (SFS, SFFS és PTA) a klasszifikáció hatékonyságának javítására és a klasszifikációs modell méret-komplexitásának csökkentésére. A definiált eljárások az irodalomból ismert hatékony tulajdonságtér-szelekciós módszerek analógiájára épülnek. A kialakított tesztek eredménye alapján elmondható, hogy ezek az eljárások segítik a hatékony klasszifikációt [28].

Az összefoglaló végén az 1-es táblázat mutatja a tézispontok és azok publikáltságának összefüggését.

Konklúzió

Ebben a rövid fejezetben összefoglaljuk a jelen tézis konzekvenciáit. A tudomány fejlődésének elsődleges mozgatórugója a technikai újdonságok megjelenése. Míg a hatvanas években a mesterséges intelligencia kutatások leginkább a lineáris és kvadratikus modellek vizsgálatára korlátozódtak, addig a mai szamítógépes kapacitások mellett jelentős fejlődés tapasztalható a modellépítés területén. A másik momentum, amely hozzajárult a tudományterület további fejlődéséhez az az, hogy jelentős mértékben megnövekedett a "tanulásra" alkalmas adathalmazok száma és mérete.

A dolgozat a kernel ötlet köré szövődik, amely a lineáris modellek nemlineáris transzformációjára alkalmas a modell komplexitásának kismértékű növekedése mellett. A kernel ötlet azon algoritmusok transzformációjára képes, amelyek inputként csak a mintavektorok egy halmazának páronként vett skalárszorzatát használják fel. Ekkor a skalárszorzat művelet nemlineáris módon történő újradefiniálásával alternatív modelleket kaphatunk. A cél tehát nem más, mint a gépi tanulás alapfeladatainak, mint például a tulajdonságkinyerés, klasszifikáció, regresszió az átdefiniálása skalárszorzatok formájában. A tézisben szereplő eredmények ehhez a temakörhoz járulnak hozzá egy-egy újszerű algoritmussal.

A tézis első részében olyan tulajdonságkinyerő eljárásokat definiálunk, amelyek hatékonyan növelik a klasszifikációs és regressziós módszerek pontosságát. A gépi tanulás standard adathalmazain és az arcfelismerés feladatán sikerült a kidolgozott eljárások létjogosultságának bizonyítása. Összegzéskent elmondhatjuk, hogy a kidolgozásra került négy eljárás irányt mutat a nagydimenziós tulajdonságterek hatékony redukciójára.

A második részben a klasszifikáció témakörével foglalkoztunk. Elsőként hipersík alapú klasszifikációs módszerek egy családját mutattuk be.

A kialakított módszercsalád vezérlő motívuma a módszerek hátterében nyugvó geometriai szemlélet volt. Szintén a második részben megadtunk egy konvex függvények optimalizációjára vezető általános klasszifikációs sémát, amely számtalan az utóbbi években kidolgozott eljárást foglal magában. Az egységes szemlélet, amely itt a gondolatok vezetője volt szintén egy hagyományos numerikus matematikai gondolkodást feltételezett. A második részben bevezetett eljárások valós és mesterséges adatokon is jól viselkedtek. Így a konstrukciós eredményeken túl az előző rész eredményeihez hasonlóan a gyakorlati használhatóság szempontja is előtérbe kerül.

Melléklet

A.1 Az MMDA eljárás

Rögzítsünk egy nemnegatív C számot, amelyet a téves klasszifikáció súlyozására fogunk használni. Először definiáljuk a maximális margó szeparációs problémát (MMSO), amely egy a klasszifikációt segítő tulajdonságkinyerő algoritmus. Legyen u egy d-dimenziós vektor: $u \in \mathbb{R}^d$. Az (X, y, C, u) objektumnégyessel paraméterezett MMSO probléma a következő feltételes szélsőértékfeladatot jelenti:

$$\begin{aligned} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i \to \min \text{ s.t.} \\ y_i(w^T x_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, \quad i = 1, \ldots, n, \\ u^T w = 0, \end{aligned} \tag{5}$$

ahol a $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ és $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n$ változók optimális értékeit kell meghatározni. A formulák alakja hasonló marad, ha a merőlegességi feltételt egyről, mondjuk r-re bővítjük. Ekkor az (X, y, C, U) paraméterezésű MMSO problémát kapjuk, ahol az $U = (u_1, \ldots, u_r)$ mátrix oszlopvektorai az $U^T w = 0$ egyenlettel definiálják a merőlegességi feltételeket.

Egyszerűen látható , hogy az (X, y, C, U) objektumnégyessel paraméterezett MMSO probléma a következő duális kvadratikus programozási probléma megoldására vezet:

$$\begin{aligned} -\frac{1}{2}\left(\alpha^\top Y K Y \alpha + \gamma^\top U^T U \gamma\right) + \alpha^\top 1 + \gamma^\top U^T X Y \alpha \to \max_{\alpha,\gamma} \\ y^T \alpha = 0, \quad 0 \leq \alpha \leq C1, \end{aligned} \tag{6}$$

ahol $K = X^T X$ és $1 = (1, \ldots, 1)^T$. Mivel az U mátrix oszlopvektorainak száma r, a γ változóvektor is r dimenziós és így a fenti feladatban a változók száma $n + r$-nek adódik.

A direkt MMDA módszer a következő módon működik: legyen adott az (X, y, C) objektumhármas, továbbá az $(X, y, C, 0)$-val paraméterezett MMSO probléma megoldását jelölje (w_1, b_1). ξ-t elhagyjuk a megoldásvektorból, mivel a klasszifikáció döntési felületében nem vesz részt. Feltéve, hogy az MMSO probléma optimalizálásával már generáltunk r tulajdonságvektort, nevezetesen a $(w_1, b_1), \ldots, (w_{r-1}, b_{r-1})$ vektorokat, akkor az r-edik (w_r, b_r)-el jelölt tulajdonságvektort, az (X, y, C, W_{r-1})-el paraméterezett MMSO probléma optimalizálásával kapjuk. A paraméterezésben a $W_{r-1} = (w_1, \ldots, w_{r-1})$.

Algorithm 1 MMDA tulajdonságkinyerő eljárás arcfelismeréshez

input: $(m, (x_1, y_1) \ldots, (x_N, y_N))$ // különböző arcok száma, arckép-személyazonosító párok listája

$F := ()$; $X^i := \{x_j | y_j = i\}$, $i = 1, \ldots, m$; // arcképek az i. személytől

$(w_1, \ldots, w_n) := \mathrm{FE}((x_1, y_1) \ldots, (x_N, y_N))$; // n tulajdonság kinyerése az FE módszerrel

for $i \in \{1, \ldots, n\}$ **do**

$z_j := w_i^T x_j$, $j = 1, \ldots, N$; // a képek vetülete

$Z^i := \{z_j | y_j = i\}$, $i = 1, \ldots, m$; // az i. személyhez tartozó képek vetületei

Keressük $(v_1, \ldots, v_m) \in \{-1, 1\}^m$-et feltéve, hogy

$$\sum_{\substack{v_i=-1, v_j=-1 \\ i \neq j}} \sum_{\substack{z \in Z^i \\ z' \in Z^j}} (z - z')^2 + \sum_{\substack{v_i=1, v_j=1 \\ i \neq j}} \sum_{\substack{z \in Z^i \\ z' \in Z^j}} (z - z')^2$$

minimális.

$F_0 := MMDA(\cup_{v_i=-1} X^i, \cup_{v_j=+1} X^j)$; // MMDA-val kinyert tulajdonságok

append$(F, F0)$; // az eddig kinyert tulajdonságok bővítése

end for

return F

A.2 Az MMDA tulajdonságkinyerő eljárás arcfelismeréshez

Az emberi arcok felismerése egy speciális klasszifikációs feladatot képez, hiszen az osztályok száma rendkívül nagy, az egy-egy osztályba eső elemek száma viszont csekély. A feladatnak éppen ezen tulajdonsága teszi a hatékony klasszifikációt nehézzé.

Mivel az MMDA algoritmus bináris klasszifikációs problémák megoldásához készült, a többosztályos tanulás megoldásánál osztálycsoport-párok kialakítására van szükség. Az MMDA algoritmus kialakításakor [11] az ilyen feladatokra eredendően az "egy a több ellen" megközelítést javasoltuk. Nevezetesen, ha adott egy m-osztályos feladat, akkor az MMDA algoritmust m-szer kell végrehajtani minden egyes osztályra a kimaradó osztályok ellenében. Annak ellenére, hogy ez egy rendkívül egyszerű megközelítés, alkalmazásával a jóval összetettebb algoritmusok - mint például az output-kód - eredményeivel hasonlóakat lehet kapni.

Sajnos az "egy a több ellen" megközelítés azonban az arcfelismerés feladatán nem eredményez reprezentatív tulajdonságokat, mivel az így kapott kétosztályos feladat nagyon kiegyensúlyozatlan lesz az osztályokba eső minták elemszámát tekintve. A minták ilyen jellegű eloszlása a különböző egy a több ellen részfeladatok megoldása esetén rendkívűl korreláló tulajdonságokat fog eredményezni. A szokásos megoldás ezekben az esetekben az osztályok klaszterezése és a hasonló jellegű osztályok összevonása. Követve ez utóbbi ötletet alakítottuk ki az MMDA algoritmus arcfelismeréshez szánt változatát.

A feladat tehát alproblémák definiálása. Ennek egy elegendően jó megközelítése, ha megfelelő független 1-dimenziós altereket választunk a tulajdonságtérben, majd ezen 1-dimenzió mentén klaszterezzük az arcokat két csoportba. Ha független altereket valasztottunk ki, akkor a definiált részproblémák is igen különbözőek lesznek. Az 1-dimenzios alterek kiválasztásához (ld. FE az 1-es algoritmusban) például használható a főkomponens analízis [10], vagy akár a lineáris diszkrimináns analízis [9] módszere is. Ezek után az alproblémákon az MMDA eljárás alkalmazását javasoljuk végrehajtani, rendre az egyes tulajdonságok definiálásához. Az így kinyert tulajdonságok összessége alkotja a végső tulajdonságteret (ld. 1-es algoritmus).

A.3 KDLR: az MMDA regressziós változata

Tekintsük a regressziós problémát, ahol az (X_i, Y_i) minták független azonos eloszlású változók. Legyel L egy veszteségfüggvény, például a kvadratikus veszteségfüggvény $L(y, z) = (y - z)^2$ és keressük az $f(x) = \operatorname{argmin}_y E[L(Y, y)|X = x]$ optimális regressziós függvényt. Először vegyük szemügyre a következő modellt:

$$Y = \sum_i \beta_i g_i(X) + \epsilon, \tag{7}$$

ahol $g_i : X \to \mathbb{R}$ ismeretlen függvényeket jelöl, továbbá ϵ jelöli a zajt, ami feltételezésünk alapján független az Y, X változóktól.

A g_i függvények becslését iteratív módon szeretnénk végezni. A fenti modell egy lehetséges egyszerűsítése, hogy feltételezzük, hogy Y a következő lineáris regressziós alakban áll elő: $Y = \beta^T\gamma+\epsilon$, ahol $\gamma = (g_1, \ldots, g_m)$. Feltesszük továbbá azt is, hogy $0 \leq \beta \leq 1$, $\beta^T e = 1$, ahol $e = (1, 1, \ldots, 1)^T$, azaz a kimenet a $g_1(X), \ldots, g_m(X)$ tulajdonságok "zajos" konvex kombinációja. A veszteségfüggvény a továbbiakban legyen a kvadratikus veszteségfüggvény.

Legyen $g = \sum_i \beta_i g_i$, f tetszőleges. Ekkor könnyen látható, hogy

$$\operatorname{Loss}(g) = \sum_i \beta_i \operatorname{Loss}(g_i) - \sum_i \beta_i E[(g_i(X) - g(X))^2] \tag{8}$$

és

$$\operatorname{Loss}(g) = E[(g(X) - f(X))^2]. \tag{9}$$

A fenti formalizmust először [17] -ben definiálták és a az eljárást „ambiguity decomposition" (AD)-nek nevezték. Folytatva a gondolatmenetet kapjuk, hogy

$$\sum_i \beta_i E[(g_i(X) - g(X))^2] = \sum_i (\beta_i^2 - \beta_i)\Big(E[g_i(X)]^2 + \operatorname{Var}[g_i(X)]\Big) - \sum_{i \neq j} \beta_i \beta_j \operatorname{Cov}(g_i(X), g_j(X)).$$

Így, ha valamely (g_i) és $(\hat{g}_i)$ függvényhalmazokra teljesül, hogy $E[g_i(X)] = E[\hat{g}_i(X)]$ és $\operatorname{Var}[g_i(X)] = \operatorname{Var}[\hat{g}_i(X)]$, akkor

$$\sum_{i \neq j} \beta_i \beta_j E[g_i(X)\, g_j(X)] < \sum_{i \neq j} \beta_i \beta_j E[\hat{g}_i(X)\, \hat{g}_j(X)] \Rightarrow \operatorname{Loss}(g) < \operatorname{Loss}(\hat{g}).$$

A levezetés tanulsága, hogy az $E[g_i(X)g_j(X)] = 0, \quad i \neq j$ feltétel csökkenti a veszteségfüggvény értékét.

Most legyen a $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ egy pozitív definit kernel, a $\mathcal{H}$ pedig a hozzá tartozó reprodukálható kernel-Hilbert tér (RKHS). $\{(x_i, y_i)\}_{i=1}^n$ jelölje az adatokat (itt is független, azonos eloszlású véletlen változókat kell feltételezni), $L(y, z) = (y - z)^2$ és $f \in \mathcal{H}$. Definiáljuk a következő feladatot

$$R(f) = \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i) + \lambda \|f\|_2, \tag{10}$$

ahol $\| \ \|_2$ a $\mathcal{H}$ Hilbert tér normája. Wahba 'reprezentációs tétele' [33] alapján ekkor $f \in \operatorname{span}(\Phi)$, $\Phi = (\phi_1, \ldots, \phi_n)$ és a $\phi_i : \mathbb{R}^d \to \mathbb{R}$ függvényt a $\phi_i(x) = k(x_i, x)$ definiálja. Tegyük fel, hogy $f = \Phi\alpha$ valamely $\alpha \in \mathbb{R}^n$-re. Ekkor (10) megoldása a következő

$$R(\alpha; X; k) = \frac{1}{n} \sum_{i=1}^{n} L((\Phi\alpha)(x_i), y_i) + \lambda \alpha^T K \alpha, \tag{11}$$

ahol $K_{ij} = k(x_i, x_j)$ és $X = (x_1, \ldots, x_n)$. Az X adathalmazra és rögzített k kernel függvényre egyszerűen csak $R(\alpha)$-at írunk $R(\alpha; X; k)$ helyett.

Most tegyük fel, hogy $g_i = \Phi\alpha_i$ és $g_j = \Phi\alpha_j$. Ha a várható érték operátort felcseréljük az empirikus várható értékkel, akkor kapjuk, hogy

$$0 = \sum_{k=1}^{n} g_i(x_k) g_j(x_k) = \alpha_k^T K^2 \alpha_j. \tag{12}$$

Így az 'ambiguity' dekompozíció alapján nyert ortogonalitási feltétel esetén $R(\alpha)$ iteratív optimalizációjára a következőt kapjuk: ha már adott $\alpha_1, \ldots, \alpha_i$, akkor

$$\alpha_{i+1} = \underset{\alpha}{\operatorname{argmin}} \{ \, R(\alpha) \, | \, \alpha_j^T K^2 \alpha = 0, \, 1 \leq j \leq i \, \}. \tag{13}$$

Ha valamely $k > 0$-ra $\alpha_1, \ldots, \alpha_k$ már előállt, akkor az optimális β_i-k is előállíthatók például a legkisebb négyzetek módszerével. A fenti (13) egyenlet és a β_i meghatározása képezi a 'Decorrelated Learning Regression (DLR)' eljárást.

A (13)-as probléma megoldása a Lagrange duálisának előállításával történik. Ehhez tegyük fel, hogy a megoldás i. lépésében ΦA_i formában adott a megoldás, ahol $A_i = [\alpha_1, \ldots, \alpha_i]$. Most vegyük V. Vapnik [30] ϵ-veszteség függvényét, amelyet az $L(y, z) = \max(0, |y - z| - \epsilon)$ kifejezés definiál. Ekkor algebrai átalakítások után a következő dualis kvadratikus programozási feladathoz jutunk:

$$\begin{aligned} L(\alpha, \alpha^*, \beta) = &-\frac{1}{2}(\alpha - \alpha^*)^T K (\alpha - \alpha^*) - (\alpha - \alpha^*)^T K^2 A_i \beta \\ &- \frac{1}{2} \beta^T A_i K^3 A_i \beta + (\alpha - \alpha^*)^T y - \epsilon(\alpha + \alpha^*)^T e \to \max \\ \text{s.t. } &0 \leq \alpha_i, \alpha_i^* \leq C \ \forall i. \end{aligned}$$

A.4 KADE: átlagos deriváltbecslő eljárás kernelekkel

Tegyük fel, hogy az ismeretlen f regresszionálandó függvény a következő alakban írható

$$f(x) = f_0(Bx), \tag{14}$$

ahol $B \in \mathbb{R}^{m \times d}$, $m \ll d$ és $BB^T = I_m$. A fenti definícióban f_0 egy ún. közvetítő függvény. A célunk egy releváns m-dimenziós $\mathcal{S} = \Im B^T$ altér megtalálása, ahol a regresszió elvégzése hatékonyabb [19]. Az atlagos deriváltbecslő eljárás alapötlete a következő: tekintsük f deriváltját minden $x \in \mathbb{R}^d$ pontban ekkor az

$$F(x) \stackrel{\text{def}}{=} B^T f_0'(Bx) \tag{15}$$

definícióval kapjuk, hogy $F(x) \in \mathcal{S}$.

Az alapötlet szerint most becsüljük az f függvényt valamely nem parametrikus módszer alkalmazásával. Jelölje ezt $\hat{f}$, továbba a becslésben felhasznált mintapontokat definiálja $x_1, \ldots, x_n$.

Most legyen

$$\hat{F}(x) = d/dx\hat{f} \tag{16}$$

és számítsuk ki a sajátérték-sajátvektor felbontását az

$$M = \sum_i \hat{F}(x_i)\hat{F}(x_i)^T. \tag{17}$$

mátrixnak. Ha $\hat{F} = F$, akkor könnyen láthatjuk, hogy az M mátrixnak csak az első m sajátértéke különbözik nullától. Mivel $\hat{F}$ csak egy közelítése F-nek azt várhatjuk, hogy M-nek m-nél több nem zérus sajátvektora lesz. A módszer gyakorlati alkalmazása során a spektrumban egy jelentősebb esés segít a releváns altér dimenziójának meghatározását.

A kernel gépek alkalmazását javasoljuk az f becslésére, azaz $\hat{f}$ előállítására. Az így kialakult módszertant kernel átlagos deriváltbecslő eljárásnak (KADE) nevezzük. A kernel módszerek implantálását az ADE eljárásba az a gépi tanulásban széleskörben elfogadott nézet indukálja, hogy ezek az eljárások kevésbé érzékenyek az input tér dimenziójára. A módszerek ezen tulajdonsága a KADE eljárás iteratív alkalmazásánál például egy fontos szempont lehet.

A.5 Hipersík alapú klasszifikáció

A hipersík alapú klasszifikációs módszerek látszólag gyenge aproximatornak tünnek a pozitív es negatív osztályok elválasztása szempontjából. Hiszen kis dimenziós térben könnyen definiálhatók olyan mintahalmazok, ahol a lineáris elválasztás nem müködik. A kernel ötlet azonban éppen ezt a limitet oldja fel es ezért újra fontossá vált a hipersík alapú módszerek vizsgálata.

Nagyon röviden három módszert mutatunk be. Az első a hipersíkos klasszifikációt veszteségfüggvényekkel bővíti ki, a második módszer a regresszió formalizmusát használja fel klasszifikációra az irodalomban ismerttől különböző módon. A harmadik eljárás a módszercsalád talán legfontosabb eredménye, a 'Minor Component Classifier (MCC)', amely az input es output tér egy közös térben történő kezelésével végez klasszifikációt.

A kernel ötlet alkalmazása után a következő formulákat kapjuk:
a) Lineáris klasszifikáció veszteségfüggvényekkel a kernel tulajdonság térben:

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} g\left(y_i \sum_{j=1}^{n} \alpha_i k\left(\begin{pmatrix} \mathbf{x}_i \\ 1 \end{pmatrix}, \begin{pmatrix} \mathbf{x}_j \\ 1 \end{pmatrix} \right) \right), \tag{18}$$

ahol g egy veszteségfüggvény, k pedig egy kernel. A megoldás hatékonyan Newton vagy kvázi-Newton módszerekkel határozható meg.
b) Lineáris regresszió klasszifikációhoz: a döntési felületet a következő kifejezés definiálja egy tetszőleges z mintára:

$$sign\left((z^T 1) X_1 (K^T K)^+ K^T Y \right), \tag{19}$$

ahol

$$X_1 = \begin{pmatrix} \mathbf{x}_1^T & 1 \\ \vdots & \vdots \\ \mathbf{x}_n^T & 1 \end{pmatrix} \text{ és } K_{ij} = k\left(\begin{pmatrix} \mathbf{x}_i \\ 1 \end{pmatrix}, \begin{pmatrix} \mathbf{x}_j \\ 1 \end{pmatrix} \right). \tag{20}$$

c) A 'Minor Component Classifier' alakja a következő:

$$\min_{\boldsymbol{\beta}} \frac{\boldsymbol{\beta}^T \bar{K} \bar{K} \boldsymbol{\beta}}{\boldsymbol{\beta}^T \bar{K} \boldsymbol{\beta}}, \tag{21}$$

ahol a $\bar{K}$ mátrix a kiterjesztett mintavektorok páronkénti skalárszorzatával definiált:

$$\bar{K}_{ij} = k\left(\begin{pmatrix} \mathbf{x}_i \\ y_i \\ 1 \end{pmatrix}, \begin{pmatrix} \mathbf{x}_j \\ y_j \\ 1 \end{pmatrix} \right). \tag{22}$$

A (21)-es feladat megoldása a

$$\bar{K} \bar{K} \boldsymbol{\beta} = \lambda \bar{K} \boldsymbol{\beta} \tag{23}$$

sajátérték-sajátvektor probléma legkisebb nem triviális sajátértékéhet tartozó sajátvektorának meghatározásával történik.

A.5 Konvex gépek

Tekintsük ismét a klasszifikáció feladatát. Legyen adott n minta az $\mathbb{R}^m$ feletti $\mathcal{X}$ kompakt halmazon. A mintákat jelöljék az $\mathbf{x}_1, \ldots, \mathbf{x}_n$ vektorok, minden egyes $\mathbf{x}_i$ mintára legyen definiált annak osztálya. Az osztálycímkék általában c osztályos feladat esetén az $\{1, \ldots, c\}$ halmazból kerülnek ki és szokás szerint $y_1, \ldots, y_n$ jelöli. Mivel a többosztályos problémák visszavezethetők kétosztályos feladatokra csak ez utóbbival foglalkozunk [12; 34]. Az osztálycímkéket – szintén az általános jelölésrendszert követve – $c = 2$ esetén praktikusan -1, $+1$-re módosítjuk az egyszerűbb írásmód kedvéért.

Legyen V a valós függvények vektortere. $S \subset V$ pedig a függvények egy kiválasztott halmaza:

$$S = \{f_1(\mathbf{x}), \ldots, f_k(\mathbf{x})\}, \qquad f_i : \mathcal{X} \to \mathbb{R}. \tag{24}$$

$Span(S)$ jelölje az S elemei által kifeszített lineáris teret:

$$Span(S) = \left\{ f_{\boldsymbol{\alpha}} : \mathcal{X} \to \mathbb{R} \mid f_{\boldsymbol{\alpha}}(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i f_i(\mathbf{x}),\ \mathbf{x} \in \mathcal{X},\ \boldsymbol{\alpha} \in \mathbb{R}^k \right\}. \tag{25}$$

A klasszifikáció feladatát a következő optimalizációs problémával definiáljuk:

$$\min_{f_{\boldsymbol{\alpha}}(x) \in Span(S)} E_{\mathbf{x},y} L(f_{\boldsymbol{\alpha}}(x), y), \tag{26}$$

ahol L egy veszteségfüggvény, az $f_{\boldsymbol{\alpha}}(x)$ prediktor minőségét méri, E pedig a várható érték operátor (x, y) felett.

Az (26)-os egyenlet egy lehetséges megszorítása a következő:

$$\min_{f_{\boldsymbol{\alpha}}(x) \in Box(S,\mathcal{B})} E_{\mathbf{x},y} H(y_i f_{\boldsymbol{\alpha}}(x_i)). \tag{27}$$

Itt az $L(f_{\boldsymbol{\alpha}}(x), y)$ -ről felteszzük, hogy $H(y_i f_{\boldsymbol{\alpha}}(x_i))$ alakban áll elő, ahol $H : \mathbb{R} \to \mathbb{R}$ egy kétszer folytonosan differenciálható, nemnegatív, csökkenő, konvex függvény. $Box(S, \mathcal{B})$ továbbá a $Span(S)$ megszorítása egy 'téglatest' típusú tartományra. Legyen $\mathcal{B} = \mathcal{B}_1 \times \cdots \times \mathcal{B}_n \subseteq \mathbb{R}^n$ nem üres intervallumok szorzata. Ekkor

$$Box(S, \mathcal{B}) = \left\{ f_{\boldsymbol{\alpha}} : \mathcal{X} \to \mathbb{R} \mid f_{\boldsymbol{\alpha}}(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i f_i(\mathbf{x}),\ \mathbf{x} \in \mathcal{X},\ \boldsymbol{\alpha} \in \mathcal{B} \right\}. \tag{28}$$

A függvényapproximáció feladata ritka minták esetén egy aluldefiniált feladat [32], már eddig is két különböző megszorítást kellett tennünk a prediktor függvény alakjára: i) $f_{\boldsymbol{\alpha}}(x)$-t bázisfüggvények véges lineáris kombinációjának alakjában keressük, ii) a lineáris kombináció komponensei egy-egy rögzített intervallumba esnek. A regularizációs elmélet [7; 29] egy további praktikus megszorítást indukál. Nevezetesen hozzáadunk a célfüggványhez egy regularizációs tagot:

$$\min_{f_{\boldsymbol{\alpha}}(x) \in Box(S,\mathcal{B})} E_{\mathbf{x},y} H(y_i f_{\boldsymbol{\alpha}}(x_i)) + \lambda \|\boldsymbol{\alpha}\|_A^2, \tag{29}$$

ahol $\lambda > 0$ és $A \in \mathbb{R}^{k \times k}$ egy tetszőleges szimmetrikus pozitív definit mátrix. Ezt a formalizmust 'Konvex gépeknek' (Convex Machines - CM) nevezzük.

A.5 Alapvető heurisztikák bázisfüggvény szelekcióhoz

Egy CM modell által előállított diszkriminatív hiperfelület

$$\{\mathbf{z} \mid \sum_{j=1}^{k} \alpha_j f_j(\mathbf{z}) = \gamma,\ \mathbf{z} \in \mathcal{X}\}, \qquad f : \mathcal{X} \to \mathbb{R}. \tag{30}$$

alakban írható fel rögzített $\gamma \in \mathbb{R}$ küszöbszámra. A modell kiértékelésekor az összegből elhagyhatóak az $\alpha_j = 0$ együtthatóval szereplő elemi elválasztó függvények. A megmaradt tagok definiálják a CM tárkomplexitását, ritkaságát. Természetesen minél több az elhagyható elem, annál gyorsabb lesz a CM

modell kiértékelése, lehetőséget teremtve a gyors válaszidejű alkalmazásokban történő felhasználásra. Az együtthatók értékét azonban a feladat optimális megoldása határozza meg, amelyre ritkaság szempontjából közvetlen hatással nem lehet élni a paramétereken keresztül, a teljesítmény romlása nélkül.

A megoldás ritkaságának kontrollálása végett korlátozzuk az elemi elválasztó függvények számát, azaz a CM modell értelmezési tartományát

$$\sum_{i=1}^{k} |sign(\alpha_i)| \leq q \tag{31}$$

feltétellel szűkítjük le. Egy ilyen tipusú megszorítás sérti a tartomány zárt és konvex tulajdonságát, így a feladat kombinatorikus úton történő megoldására késztet. Célunk a lehetséges elemi elválasztó függvényeknek azt a q számosságú részhalmazát kiválasztani, amellyel a klasszifikációs feladat a legjobban oldható meg. Ez a probléma NP nehéz [5], így heurisztikák alkalmazása kerül előtérbe, amelyek épülhetnek a saját célfüggvényükre, vagy a CM modell eltérő paraméterezésű működtetésére is. A továbbiakban ezeket tekintjük át.

A következőkben olyan eljárásokkal foglalkozunk, amelyek egy q elemű részhalmaz kiválasztása során nem a CM modell célfüggvényét használják fel.

RANDOM A legegyszerűbb stratégia a véletlen mintaválasztás, amikor a k bázisfüggvényből q elemű véletlen mintát veszünk. Az eljárás nem rendelkezik célfüggvénnyel, így többszöri végrehajtás után az adódó legjobb mintát tartjuk meg.

MGRAMM A CM modell a bázisfüggvények egy lineáris kombinációjával közelíti az optimális elválasztófelületet. Ezért a közelítést elegendő elvégezni a bázisfüggvények terének egy tetszőleges bázisán, amit például Gramm-Schmidt ortogonalizációval kaphatunk meg. A bázis számossága a függvényhalmaz rangja, ami túllépheti a kiválasztani kívánt q elemet. Másrészről az eljárás egy ortogonális függvényrendszert állít elő az eredeti bázisfüggvények lineáris kombinációival, az egyes függvények kiválasztása helyett.

Ezért egy olyan iteratív kiválasztási stratégiát definiálunk, amely a Gramm-Schmidt ortogonalizáció egy módosított változatára épül. Minden lépésében a lehetséges függvények közül azt választjuk, amely maradéktagjának normája maximális. Az eljárás eredménye nem maga az ortogonális függvényrendszer lesz, hanem azon bázisfüggvények, amelyek lineáris kombinációjaként felépülnek az ortogonalizáció során kialakított függvények.

Tegyük fel, hogy az elemi elválasztófüggvények az L_2 tér elemei, így a skalárszorzás a szorzatfüggvény integráljával számítható. Ha az integrál kiszámítására nincs lehetőség, akkor a következő közelítéssel élünk az algoritmus során:

$$\langle f, g \rangle = \sum_{i=1}^{n} f(\mathbf{x}_i) g(\mathbf{x}_i) \qquad f, g : \mathcal{X} \to \mathbb{R}. \tag{32}$$

CORR Az MGRAMM módszerhez hasonló eljárást kaphatunk, ha az iteráció során az új bázisfüggvényt úgy vesszük be a kiválasztott báziselemek listájába, hogy az a már korábban kiválasztott elemekre leginkább merőleges legyen. A *CORR* módszer ennek a mértékét a kiválasztott elemekkel történő merőlegesség négyzetének összegével definiálja.

A.6 Komplex heurisztikák bázisfüggvény szelekcióhoz

A mérték alapú részhalmaz-kiválasztási probléma a mesterséges intelligencia más területein is előfordul, ilyen terület például a tulajdonság szelekció. Itt a mintapontok m dimenziójából kell azt az r komponenst kiválasztani, amellyel a klasszifikáció a legjobban megoldható az adott gépi tanulási algoritmussal [24]. A területen kidolgozott algoritmusok felhasználhatóak az elemi szeparátorfüggvények q elemű részhalmazának kiválasztására is, ha a mértéket a CM modell célfüggvényével helyettesítjük.

SFS A *Sequential Forward Selection* (SFS) algoritmus a mérték minimalizálásának egy mohó megközelítése. Az üres halmazból kiindulva minden lépésben a lokálisan optimális elemmel bővíti az indexhalmazt, visszalépések nélkül.

PTA Az SFS algoritmus szekvenciális működésű, azaz a megelőző lépések lokális optimum alapú választásának kihatásait a későbbi lépésekben nem lehet korrigálni. A probléma feloldásának egyik lehetősége a *Plus* l *Take Away* r (PTA(l,r)) megközelítés. Minden l SFS tipusú bővítési lépés után r elemet eltávolítunk szekvenciálisan úgy, hogy a mérték értéke minden lépésben a lokális optimumba kerüljön.

SFFS A PTA eljárás működése során l bővítési lépést mindig r szűkítési követi. Emiatt előfordulhat, hogy akkor is végrehajtunk szűkítési lépést, amikor az előálló megoldás rosszabb mértékkel rendelkezik, mint az eddigi vele megegyező számosságú. Fordított helyzet is adódhat, azaz nem szűkítünk, miközben jobb mértékkel rendelkező megoldást kapnánk adott szinten. Ezt küszöböli ki a Sequential Forward Floating Selection (SFFS) algoritmus, amely minden SFS tipusú bővítési lépés után addig távolít el elemeket szekvenciálisan, amíg a mérték adott szinten meghaladja az eddigi legjobbat.

Hivatkozások

[1] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[2] S. Akaho. A kernel method for canonical correlation analysis. In *International Meeting of Psychometric Society (IMPS), Osaka*, 2001.

[3] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/mlrepository.html, 1998.

[4] B.E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learing Theory*, pages 144–152, 1992.

[5] Thomas M. Cover and Jan Van Campenhout. On the possible orderings in the measurement selection problem. *IEEE Trans. Systems, Man, and Cybernetics*, 7:657–661, 1977.

[6] M. Teixeira D. Bolme, R. Beveridge and B. Draper. The csu face identification evaluation system: Its purpose, features and structure. In *International Conference on Vision Systems*, pages 304–311. Springer-Verlag, 2003.

[7] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

[9] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, California, 1990.

[10] I. T Jolliffe. *Principal component analysis*. New York : Springer, 2002.

[11] A. Kocsor, K. Kovács, and Cs. Szepesvári. Margin maximizing discriminant analysis. In Fosca Giannotti et al. Jean-François Boulicaut, Floriana Esposito, editor, *Machine Learning: ECML 2004: 15th European Conference on Machine Learning, vol. 3201*, pages 227–238. Springer-Verlag GmbH, September 2004.

[12] E. B. Kong and T. G. Dietterich. Error-correcting output coding corrects bias and variance. In *International Conference on Machine Learning (ICML)*, pages 313–321, 1995.

[13] K. Kovács and A. Kocsor. Various hyperplane classifiers using kernel feature spaces. *Acta Cybernetica*, 16(2):271–278, 2003.

[14] K. Kovács and A. Kocsor. Classification using sparse combination of basis functions. *Acta Cybernetica*, 17(2):311–323, 2004.

[15] K. Kovács and A. Kocsor. Improving a basis function based classification method using feature selection algorithms, accepted for ieee international workshop on soft computing applications. In *IEEE International Workshop on Soft Computing Applications, (IEEE-SOFA)*, pages 208–211, 2005.

[16] K. Kovács, A. Kocsor, and Cs. Szepesvári. Maximum margin discriminant analysis based face recognition. In M. Vincze D. Chetverikov, L. Czuni, editor, *Proceedings of the Joint Hungarian-Austrian Conference on Image Processing and Pattern Recognition, HACIPPR*, pages 71–78. Oesterreichische Computer Gesellschaft, 2005.

[17] A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems NIPS 11*, pages 231–238, 1995.

[18] Y. Lee and O. L. Mangasarian. SSVM: A smooth support vector machine. *Computational Optimization and Applications*, 20:5–22, 2001.

[19] K.-C. Li. Sliced inverse regression for dimension reduction. (With discussion). *J. Amer. Statist. Ass.*, 86(414):316–342, 1991.

[20] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc. London, A*, 209:415–446, 1909.

[21] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14 (NIPS)*, pages 849–856. MIT Press, Cambridge, 2002.

[22] I. Nouretdinov, T. Melluish, and V. Vovk. Ridge regressioon confidence machine. In *Proc. 18th International Conf. on Machine Learning*, pages 385–392. Morgan Kaufmann, San Francisco, CA, 2001.

[23] T. Poggio. On optimal nonlinear associative recall. *Biological Cybernetics*, 19:201–209, 1975.

[24] P. Pudil, J. Novovicova, and J. Kittler. Feature selection based on the approximation of class densities by finite mixtures of the special type. *Pattern Recognition*, 28(9):1389–1397, 1995.

[25] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1995.

[26] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[27] J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

[28] Cs. Szepesvári, A. Kocsor, and K. Kovács. Kernel machine based feature extraction algorithm for regression problems. In Lorenza Saitta Ramon López de Mántaras, editor, *Proceedings of the 16th Eureopean Conference on Artificial Intelligence, ECAI 2004*, pages 1091–1091, Valencia, Spain, August 2004. IOS Press.

[29] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed Problems*. W. H. Winston, Washington, D.C., 1977.

[30] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, NY, USA, 1995.

[31] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons Inc., 1998.

[32] G. Wahba. *Splines models for Observational Data*. Vol. 59, SIAM, Philadelphia, 1990.

[33] Grace Wahba. Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In *Advances in kernel methods: support vector learning*, pages 69–88. MIT Press, 1999.

[34] J. Weston and C. Watkins. Support vector machines for multiclass pattern recognition. In *Proceedings of the Seventh European Symposium On Artificial Neural Networks (ESANN)*, pages 219–224, 1999.