Konvex poliéder tanulás és alkalmazásai

Tézisfüzet

Takács Gábor

Témavezető: Dr. Pataki Béla

Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Méréstechnika és Információs Rendszerek Tanszék

Tartalomjegyzék

1. Témavázlat

Egy lehetséges mérnöki nézőpontból szemlélve a *tanulás* nem más, mint egy jelenség attribútumai közötti kapcsolat feltérképezése. A *gépi tanulás* (adatbányászat) a tanulásnak az a változata, amikor a jelenséggel kapcsolatos megfigyelések adatok formájában állnak rendelkezésre, és az attribútumok közötti összefüggést egy program keresi meg.

A gépi tanulás hatékony eszköz lehet olyan feladatok gépi megoldásánál, ahol a bemenet és az előállítani kívánt kimenet közötti kapcsolat bonyolult, de az összetartozó bemenet–kimenet párok gyűjtése viszonylag könnyű (ilyen pl. a levélszemét szűrés, a hitelkockázat előrejelzés, az arcfelismerés vagy az autóvezetés).

A tanulási probléma két érdekes speciális esete az osztályozás és a kollaboratív szűrés.

- *Osztályozás* [1] esetén a jelenséget egy véletlen $(\mathbf{X}, Y)$ pár modellezi, ahol a d-dimenziós, folytonos $\mathbf{X}$-et bemenetnek, a diszkrét (gyakran bináris) Y-t pedig címkének hívjuk. A cél egy olyan g függvény meghatározása, amely minél kisebbé teszi a $\mathbf{P}\{g(\mathbf{X}) \neq Y\}$ hibavalószínűséget.

- *Kollaboratív szűrés* [2] esetén a jelenséget egy véletlen (U, I, R) hármas modellezi, ahol a diszkrét, U-t felhasználó azonosítónak (user identifier), a diszkrét I-t termékazonosítónak (item identifier), a folytonos R-et pedig értékelésnek (rating value) nevezzük. A cél egy olyan g függvény meghatározása, amely minél kisebbé teszi az $\mathbf{E}\{(g(U, I) - R)^2\}$ várható négyzetes hibát. A kollaboratív szűrés témaköre az 1 millió dollár fődíjú Netflix verseny [3] hatására 2006-ban a tudományos érdeklődés középpontjába került.

A jelenséget modellező véletlen vektor eloszlásáról általában nincsen információnk. Ami a feladat megoldásához rendelkezésre áll, az egy véges *tanító készlet*, amely a szóbanforgó ismeretlen eloszlás szerint generált példákat tartalmaz.

1.1. Konvex poliéder osztályozás

Az osztályozási feladatok fontos részhalmazát alkotják az úgynevezett *kiegyensúlyozatlan feladatok*. Kiegyensúlyozatlanság alatt azt értjük, hogy az egyik osztály előfordulási gyakorisága jelentősen alacsonyabb a másik osztályénál.

Ilyen feladatok többek között az orvosi vagy a műszaki diagnosztika területén szoktak felmerülni. Számítógéppel segített mellrákszűrés esetén például a vizsgálatban részt vevő páciensek túlnyomó többsége egészséges, ezért a szakértői rendszer által megoldandó osztályozási feladat kiegyensúlyozatlan.

A *konvex poliéder osztályozás* egy olyan megközelítés az osztályozás területén belül, amely jól illeszkedik a kiegyensúlyozatlan feladatokhoz. Konvex poliéder osztályozónak azon $g : \mathbb{R}^d \mapsto \{+1, -1\}$ függvényeket nevezzük, amelyek esetén az $\{\mathbf{x} \in \mathbb{R}^d : g(\mathbf{x}) = 1\}$ döntési tartomány egy konvex poliéder (véges sok zárt féltér metszete).

Egy konvex poliéder osztályozót például K darab hipersík segítségével lehet megadni. Az $\mathbf{x} \in \mathbb{R}^d$ bemenet osztályozása során $\mathbf{x}$-et a hipersíkokat definiáló lineáris függvényekbe kell behelyettesíteni. Ha az osztályozandó bemenet bármelyik hipersíknak a negatív oldalára esik, akkor a $g(\mathbf{x})$ függvényérték -1 lesz. Ez azt jelenti, hogy abba is lehet hagyni a számítást, ha valamelyik behelyettesítés eredményeként negatív számot kapunk.

A kutatásom egyik fő célja alacsony számításigényű, és pontos modellt eredményező tanító algoritmusok kidolgozása volt konvex poliéder osztályozókhoz. Ilyen módszerek korábban nem voltak ismertek, és ez komoly akadálya volt a konvex poliéder osztályozók gyakorlatban való alkalmazásának.

1.2. Konvex szeparálhatóság

A konvex poliéder osztályozáshoz kapcsolódó érdekes probléma a konvex szeparálhatóság eldöntésének feladata. A feladat formális megfogalmazása a következő: Adott egy $\mathcal{P}$ és egy $\mathcal{Q}$ véges ponthalmaz az $\mathbb{R}^d$ térben. Kérdés, hogy létezik-e olyan $\mathcal{S}$ konvex poliéder, amely minden $\mathcal{P}$-beli elemet tartalmaz, de egyetlen $\mathcal{Q}$-beli elemet sem tartalmaz.

A gyakorlatban a konvex szeparálhatóság eldöntése például egy gépi tanulási feladat megoldásának első, adatfeltérképező fázisában lehet hasznos. A rendelkezésre álló tanító készlet tulajdonságainak vizsgálata segíthet a megfelelő osztályozó algoritmus kiválasztásában.

1.3. Kollaboratív szűrés konvex poliéderekkel

Konvex poliéder alapú módszereket természetesen nemcsak osztályozás esetén lehet alkalmazni. A megközelítésnek megvan az analóg változata például kollaboratív szűrés esetére is. Ebben az esetben a gyors predikció nem lép fel előnyként, mivel a kimenet előállításához az összes hipersíkon végig kell iterálni. A konvex poliéder módszerek hasznossága ilyenkor az lehet, hogy a hagyományostól eltérő, egyedi megoldását adják a feladatnak, és hasznos tagjai lehetnek egy több módszer eredményét felhasználó, kombinált megoldásnak.

1.4. Modellkomplexitás

A gépi tanulás bizonyos szemszögből nézve modellezésnek tekinthető. A bemenet egy adathalmaz, amelyet egy jelenség megfigyelése során gyűjtöttek. A kimenet egy modell, ami valamilyen értelemben megmagyarázza az eddigi megfigyeléseket, és ami képes előrejelzést adni a jövőre vonatkozólag.

Egy gyakorlati adatbányász projekt esetén többnyire számos kísérletet lefuttatnak és több modellt felállítanak. Nem triviális feladat eldönteni, hogy melyiket érdemes előrejelzésre használni a végleges rendszerben. Ha két modell azonos pontosságot ér el a tanító készleten, akkor nyilván az egyszerűbbet érdemes választani. A kérdés az, hogy hogyan lehet egzakt módon definiálni a modellek bonyolultságát.

A Vapnik–Chervonenkis dimenzió [4] egy széles körben elfogadott modell komplexitási mérték bináris osztályozás esetén. Segítségével valószínűségi korlátot kaphatunk az osztályozók hibavalószínűségére, független tesztkészlet használata nélkül.

2. Az értekezés felépítése

Az értekezés első fejezete (Introduction) röviden bemutatja a gépi tanulás területét, és elhelyezi benne a konvex poliéder tanulás témakörét. Ezután a teljesség igénye nélkül ismertet néhány közismert tanuló algoritmust. A bemutatott módszerek kiválasztásának szempontja az volt, hogy mennyire kapcsolódnak az értekezés további részeihez. A kollaboratív szűrést bemutató rész saját eredményeket is tartalmaz.

Az értekezés második fejezete (Algorithms) olyan algoritmusokat mutat be, amelyek konvex poliédereket használnak valamely gépi tanuláshoz kapcsolódó feladat megoldására. A fejezet első része a konvex és a lineáris szeparálás problémájával foglalkozik. A fejezet második része algoritmusokat ad konvex poliéder osztályozók tanítására. A fejezet harmadik része konvex poliéder alapú algoritmust vezet be kollaboratív szűrésre. Az első két rész ismert és új módszereket is tartalmaz. A harmadik rész kizárólag új eredményeket közöl, mivel kollaboratív szűrés esetén nem állnak rendelkezésre meglévő, konvex poliéder alapú módszerek.

Az értekezés harmadik fejezete (Model complexity) a konvex poliéder osztályozók Vapnik–Chervonenkis dimenziójával kapcsolatos eddigi ismereteket foglalja össze, valamint új eredményeket bizonyít. Az értekezés negyedik fejezete (Applications) az algoritmusokkal végzett számítógépes kísérleteket ismerteti. A futtatások kisebb része mesterséges, nagyobb része valódi adatokon történt.

3. Új tudományos eredmények összefoglalása

1. **Téziscsoport**: Új algoritmusokat dolgoztam ki ponthalmazok lineáris és konvex szeparálhatóságának az eldöntésére. Kísérletekkel demonstráltam, hogy a javasolt algoritmusok számos tesztfeladaton gyorsabbak a közismert módszereknél. A téziscsoport eredményei a következő publikációkban jelentek meg: [P10], [P11], [P24].

   [T1.1] Megvizsgáltam a lineáris szeparálási feladat inkrementális módszerrel való megoldásának lehetőségét. Kidolgoztam egy lineáris programozásra épülő direkt módszert (LSEP_2), amely kedvező tulajdonságokkal rendelkezik az inkrementális módszer részeként történő alkalmazáshoz. Heurisztikákat javasoltam a következő lépés aktív kényszereinek ill. aktív változóinak megválasztására (LSEPX, LSEPY, LSEPZX, LSEPZY).

   [T1.2] Új, alacsony időigényű közelítő algoritmust adtam két ponthalmaz konvex szeparálhatóságának eldöntésére (CSEPC).

   [T1.3] Új, alacsony várható időigényű pontos algoritmust adtam két ponthalmaz konvex szeparálhatóságának eldöntésére (CSEPX). A módszer az előző algoritmust használja előfeldolgozóként.

2. **Téziscsoport**: Új, derivált alapú algoritmusokat dolgoztam ki konvex poliéder osztályozók tanítására valamint kollaboratív szűrésre. A téziscsoport eredményei a következő publikációkban jelentek meg: [P22], [P23].

   [T2.1] Megvizsgáltam a maximumképzés sima függvénnyel való közelítésének lehetőségét, és bevezettem hat, paraméterezhető sima maximum függvénycsaládot.

   [T2.2] Új, sima maximum függvényen alapuló algoritmuscsaládot dolgoztam ki konvex poliéder osztályozók tanítására (SMAX). Az algoritmusok hatékonyságát (osztályozási pontosság és futási idő tekintetében) mesterséges és valódi adathalmazokon végzett kísérletekkel demonstráltam.

   [T2.3] Új, sima maximum függvényen alapuló algoritmust adtam konvex poliéder alapú modellek tanítására kollaboratív szűrés esetén ($\mathrm{SMAX}_{\mathrm{CF}}$). Az algoritmus hatékonyságát (előrejelzési pontosság és futási idő tekintetében) a jelenleg elérhető legnagyobb benchmark adatbázison (Netflix) végzett kísérletekkel demonstráltam.

3. **Téziscsoport**: Új eredményeket értem el a konvex poliéder osztályozók Vapnik–Chervonenkis dimenziójának meghatározásában. A téziscsoport eredményei a következő publikációkban jelentek meg: [P8], [P9].

   [T3.1] Meghatároztam a síkbeli konvex K-szög osztályozók Vapnik–Chervonenkis dimenzióját. Fontos megjegyezni, hogy nem rögzítem előre a belső (konvex) tartomány címkéjét. (Az egy könnyebb, megoldott probléma.)

[T3.2] A triviálisnál jobb alsó korlátot bizonyítottam a d dimenziós konvex K-poliéder osztályozók Vapnik–Chervonenkis dimenziójára. A $d = 3$ és $d = 4$ speciális esetekben még tovább javítottam az alsó korlátot.

4. **Téziscsoport**: Új, jól skálázható, és nagy pontosságú modellek építésére képes algoritmusokat vezettem be a kollaboratív szűrés témakörében. A téziscsoport eredményei a következő publikációkban jelentek meg: [P1], [P2], [P3], [P4], [P5], [P6].

[T4.1] Új mátrix faktorizációs módszert vezettem be BRISMF (biased regularized incremental simultaneous matrix factorization) néven. A módszer a hatékonyságát a Netflix versenyben bizonyította.

[T4.2] Új tanító algoritmust adtam a Paterek-féle NSVD1 modellhez [5], amely jelenleg az egyik legfontosabb szomszédság alapú megközelítés a kollaboratív szűrés területén belül. Az algoritmus jelentősen kisebb számításigénnyel rendelkezik, mint a gradiens módszer naiv megvalósítása, ugyanakkor a végeredményként kapott modell teljesen azonos a két esetben.

4. Az eredmények alkalmazhatósága

A lineáris és konvex szeparálhatóság eldöntésére adott algoritmusok az adatbányászat kezdeti, adatfeltérképező fázisában lehetnek hasznosak. Tegyük fel, hogy van egy bináris osztályozási feladatunk, és egy „elegendően nagy" méretű tanító készletünk. Mivel a feladat megoldására fordítható erőforrásaink korlátozottak, fontos az elvégzendő kísérletek jó megválasztása.

Ha a tanító készletben az osztályok lineárisan szeparálhatók, akkor lineáris osztályozókkal érdemes elkezdeni a kísérletezést. Ha az osztályok csak konvex módon szeparálhatók, akkor konvex poliéder osztályozókkal érdemes először próbálkozni, egyébként pedig általános nemlineáris osztályozókkal. Ha azt tapasztaljuk, hogy nem a tanító készlet szeparálhatósági szintjének megfelelő osztályozó éri el a legjobb eredményt, akkor ebből az adatok zajosságára vagy a tanító készlet elégtelen méretére lehet következtetni.

A konvex poliéder osztályozók tanítására adott hatékony algoritmusok új eszközt adnak az adatbányász kezébe. A megközelítés bármilyen osztályozási feladat esetén kipróbálható, de különösen kiegyensúlyozatlan feladatok esetén lehet hasznos.

A kollaboratív szűrésre adott algoritmusok ajánlórendszerek készítéséhez használhatók. Ha van egy prediktorunk, amely képes megbecsülni, hogy egy felhasználó egy terméket hogyan értékelne, akkor az adott felhasználónak való ajánlás egyszerűen megvalósítható. Egyszerűen végigiterálunk az ajánlható termékek listáján, és mindegyikre kiszámolniuk a prediktor válaszát (az adott felhasználó esetén). A felhasználónak azokat a termékeket ajánljuk, amelyeknek a becsült értékelése a legmagasabb.

Ha a prediktor pontosabb az értékelések előrejelzésében, akkor a felhasználónak jobban tetsző termékek előbbre kerülnek a listában. Érdemes megemlíteni, hogy a predikciók értékének kis változása nagy változást tud okozni a termékek sorrendjében, ezért gyakran megéri energiát fektetni a prediktor pontosságának növelésébe.

A konvex poliéder osztályozók Vapnik–Chervonenkis dimenziójával kapcsolatos eredményeknek kétféle haszna lehet. Egyrészt egy kicsit pontosabbá teszik az osztályozó hibavalószínűségére adott becslést, másrészt ötletet adhatnak másoknak az alsó és felső korlátok további javításához.

5. Publikációk listája

[P1] G. Takács, I. Pilászy, B. Németh, and D. Tikk. Scalable collaborative filtering approaches for large recommender systems. *Journal of Machine Learning Research* (Special Topic on Mining and Learning with Graphs and Relations), 10: 623–656, 2009.

[P2] G. Takács, I. Pilászy, B. Németh, and D. Tikk. Matrix factorization and neighbor based algorithms for the Netflix Prize problem. *Proc. of the 2008 ACM Conference on Recommender Systems (RECSYS'08)*, pages 267–274, Lausanne, Switzerland, 2008.

[P3] G. Takács, I. Pilászy, B. Németh, and D. Tikk. Investigation of various matrix factorization methods for large recommender systems. *Proc. of the 2nd KDD Workshop on Large Scale Recommender Systems and the Netflix Prize Competition*, Las Vegas, Nevada, USA, 2008.

[P4] G. Takács, I. Pilászy, B. Németh, and D. Tikk. A unified approach of factor models and neighbor based methods for large recommender systems. *Proc. of the 1th IEEE ICADIWT Workshop on Recommender Systems and Personalized Retrieval*, pages 186–191, Ostrava, Czech Republic, 2008.

[P5] G. Takács, I. Pilászy, B. Németh, and D. Tikk. Major components of the Gravity Recommendation System. *ACM SIGKDD Explorations Newsletter*, 9(2): 80–83, 2007.

[P6] G. Takács, I. Pilászy, B. Németh, and D. Tikk. On the Gravity Recommendation System. *Proc. of the KDD Cup and Workshop 2007*, pages. 22–30, San Jose, California, USA, 2007.

[P7] G. Takács and B. Pataki. Case-level detection of mammographic masses. *International Journal of Applied Electromagnetics and Mechanics*, 25(1–4): 395–400, 2007.

[P8] G. Takács. The Vapnik–Chervonenkis dimension of convex n-gon classifiers. *Hungarian Electronic Journal of Sciences*, 2007.

[P9] G. Takács and B. Pataki. Lower bounds on the Vapnik–Chervonenkis dimension of convex polytope classifiers. *Proc. of the 11th International Conference on Intelligent Engineering Systems (INES 2007)*, Budapest, Hungary, 2007.

[P10] G. Takács and B. Pataki. Deciding the convex separability of pattern sets. *Proc. of the 4th IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS'2007)*, Dortmund, Germany, 2007.

[P11] G. Takács and B. Pataki. An efficient algorithm for deciding the convex separability of point sets. *Proc. of the 14th PhD Mini-Symposium, Budapest University of Technology and Economics, Department of Measurement and Information Systems*, pages 54–57, Budapest, Hungary, 2007.

[P12] G. Takács and B. Pataki. Nearest local hyperplane rules for pattern classification. *AI*IA 2007: Artificial Intelligence and Human-Oriented Computing*, pages 302–313, Rome, Italy, 2007.

[P13] G. Takács and B. Pataki. A lépcsőzetes döntéshozás elvének műszaki alkalmazásai, (in Hungarian). *Elektronet*, 16(8): 76–78, 2007.

[P14] M. Altrichter, G. Horváth, B. Pataki, Gy. Strausz, G. Takács and J. Valyon. *Neurális hálózatok*, (in Hungarian). Panem, 2006.

[P15] G. Takács and B. Pataki. Local hyperplane classifiers. *Proc. of the 13th PhD Mini-Symposium, Budapest University of Technology and Economics, Department of Measurement and Information Systems*, pages 44–45, Budapest, Hungary, 2006.

[P16] G. Takács and B. Pataki. Fast detection of masses in mammograms with difficult case exclusion. *International Scientific Journal of Computing*, 4(3): 70–75, 2005.

[P17] G. Takács and B. Pataki. Case-level detection of mammographic masses. *Proc. of the 12th International Symposium on Interdisciplinary Electromagnetic, Mechanic and Biomedical Problems (ISEM 2005)*, pages 214–215, Bad Gastein, Austria, 2005.

[P18] G. Takács and B. Pataki. Fast detection of mammographic masses with difficult case exclusion. *Proc. of the 3rd IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS'2005)*, pages 424–428, Sofia, Bulgaria, 2005.

[P19] G. Takács and B. Pataki. Computer-aided detection of mammographic masses. *Proc. of the 12th PhD Mini-Symposium, Budapest University of Technology and Economics, Department of Measurement and Information Systems*, pages 24–25, Budapest, Hungary, 2005.

[P20] N. Tóth, G. Takács, and B. Pataki. Mass detection in mammograms combining two methods. *Proc. of the 3rd European Medical & Biological Engineering Conference (EMBEC'05)*, Prague, Czech Republic, 2005.

[P21] G. Horváth, B. Pataki, Á. Horváth, G. Takács, and G. Balogh. Detection of microcalcification clusters in screening mammography. *Proc. of the 3rd European Medical & Biological Engineering Conference (EMBEC'05)*, Prague, Czech Republic, 2005.

[P22] G. Takács. The smooth maximum classifier. Accepted at: Second Győr Symposium on Computational Intelligence, 2009.

[P23] G. Takács. Smooth maximum based algorithms for classification, regression, and collaborative filtering. Accepted at: *Acta Technica Jaurinensis, Series Intelligentia Computatorica*, 2009.

[P24] G. Takács. Efficient algorithms for determining the linear and convex separability of point sets. Accepted at: *Acta Technica Jaurinensis, Series Intelligentia Computatorica*, 2009.

[P25] G. Takács, I. Pilászy, B. Németh, and D. Tikk. Unifying collaborative filtering approaches. Veszprém Optimization Conference: Advanced Algorithms (VOCAL 2008), Veszprém, Hungary, 2008.

[P26] R. Horváth-Bokor, Z. Horváth, and G. Takács. Kockázatelemzés logisztikus regresszióval nagy adathalmazokon, (in Hungarian). 28. Magyar Operációkutatási Konferencia, Balatonőszöd, Hungary, 2009.

Irodalomjegyzék

[1] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer, New York, 1996.

[2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17:734–749, 2005.

[3] J. Bennett and S. Lanning. The Netflix Prize. In *Proc. of the KDD Cup and Workshop 2007*, pages 3–6, 2007. (URL: `http://www.netflixprize.com/`)

[4] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. *Lecture Notes in Artificial Intelligence*, 3176:169–207, 2004.

[5] A. Paterek. Improving regularized singular value decomposition for collaborative filtering. In *Proc. of the KDD Cup and Workshop 2007*, pages 39–42, 2007.