

Új módszerek az adattömörítésben

Szakmai beszámoló: OTKA F60787

György András
Gépi Tanulás Kutatócsoport, MTA SZTAKI

2009. március 31.

1. Motiváció

Az utóbbi évek technológiai fejlődése új kihívásokat teremtett az adattömörítés és adattovábbítás terén. Új tömörítési eljárásokra van szükség a hálózati kommunikáció terén, ahol a tömörített adatot több felhasználónak kell eljuttatni, akik különböző kapacitású csatornákon keresztül kapcsolódnak egymáshoz. A veszteséges csomagkapcsolt hálózatokban az előforduló csomagvesztések a hagyományos tömörítési algoritmusok komoly teljesítménycsökkenését eredményezhetik. A kommunikáció során egyre nagyobb teret kapnak a kis késleltetést igénylő alkalmazások, illetve ezzel párhuzamosan - különösen a mobil eszközök elterjedésével - egyre nagyobb az igény a kis számításiigényű kódoló és dekódoló eljárások széleskörű elterjedésére. A kutatás során a modern adattömörítés elméleti és gyakorlati problémáit vizsgáltam, elsősorban a kis komplexitású, kis késleltetésű, adaptív eljárásokra koncentrálnak. Ez utóbbiak egyre fontosabbak lesznek a jövőben, hiszen a kommunikáció mindinkább mobil környezetben történik, és ennek hatására a hálózati környezet a jövőben egyre változókényebbé válik.

Mivel az adattovábbítás minősége a tömörítésen kívül jelentős mértékben függ az átvitel során felmerülő hibák kiküszöbölésétől, azok hatásának minimalizálásától, ez utóbbiak is igen nagy jelentőséggel bírnak, továbbá egyre fontosabbá válhatnak a jövőben az ún. együttes csatorna- és forráskódok, melyek az átviteli csatorna hibázási jellemzőit figyelembe véve optimalizálják a forráskódot (azaz a tömörítő algoritmust). Mindezek alapján a kutatás során mind a hagyományos veszteségmentes adattömörítés egyes kérdéseit, mind a kis késleltetésű univerzális, adaptív veszteséges tömörítés kérdéseit vizsgáltam. Ez utóbbinál a hagyományos kódok mellett a csomagkapcsolt hálózatok csomagvesztéseinek hatását kiküszöbölő többszörös leírású illetve többszörös felbontású kódokat is vizsgáltam. Az átvitel javítása érdekében fontos kérdés, hogy a tömörített adatok milyen útvonalon jutnak el a forrástól a vevőig; erre a korábbiaknál hatékonyabb, robusztus adaptív útvonalválasztási eljárásokat dolgoztunk ki. Az adaptív eljárások háttérében az univerzális predikció területén született eredmények állnak, melyek közül többet e kutatás során dolgoztunk ki. Végezetül egy speciális pre-

dikciós problémára is megoldást adtunk, mely – egyebek mellett – alkalmazható bizonyos szenzorhálózatok adásütemezésére is.

A kidolgozott algoritmusok teljesítményére minden esetben elméleti korlátokat adtunk. Ezek jelentős része az individuális sorozatok predikciójának elméletén alapul, ahol a predikálendő sorozat esetén nem élünk semmilyen sztochasztikus feltevessel, így az analízis általában „legrosszabb eset” típusú. Mindez azt eredményezi, hogy a kidolgozott módszerek igen robusztusak. (Ugyanakkor sztochasztikus feltevések mellett a közölt algoritmusok természetesen javíthatók, ám ez csak akkor igaz, ha a feltevések a gyakorlatban is teljesülnek. A probléma igazi megoldása a két megközelítés ötvözése lenne, mely során mindent modellezünk, amit tudunk, a rosszul modellezhető dolgok hatását pedig a legrosszabb eset feltevés mellett vizsgálánk – ez azonban túlmutat a jelen pályázat keretein.)

Az alábbiak részletesebben kifejtem a különböző területeken elért eredményeket.

2. Blokk-Markov-források szimbólumalapú veszteségmentes tömörítése

Az entrópiakódolás alkalmazásával jelentős sávszélesség takarítható meg mind a hagyományos, mind a hálózati forráskódolás során. Kísérletek alapján megfigyelhető, hogy a mindennapokban hozzáférhető, bájtos alapon működő tömörítési módszerek (pl. zip, bzip) rosszul működnek akkor, ha a forrás nem bájtalapú - ugyanakkor ezen eljárások bitalapú implementációi jól működnek bájtos és nem bájtos források esetén is.

A kutatás során megmutattuk, hogy a bitalapú tömörítés teljesítménye bizonyíthatóan jó: Előbb egy explicit formulát adtunk egy blokk-Markov-forrás (azaz a forrás blokkjai Markov tulajdonságúak) és az azt modellező magasabb rendű, szimbólumalapú Markov-forrás divergenciájátára, és megmutattuk, hogy ez exponenciálisan gyorsan 0-hoz tart, ahogy a modell rendje nő. Ennek segítségével bebizonyítottuk, hogy blokk-Markov-források tömörítése esetén a kódolás során használt szimbólumalapú Markov-modell rendjének növelésével a kódolás redundanciája exponenciális sebességgel eltűnik (az exponens a forrás jellemzőiből meghatározható).

Ennek következtében a bitalapú módszerek - különösen, mivel gyakran kis komplexitással implementálhatók - jó alternatívát jelenthetnek a bájtalapú kódolókkal szemben. A témából egy folyóiratcikket készítettünk, mely az IEEE Transactions on Information Theory-ban jelent meg [J1].

3. Univerzális predikció, időben változó teljesítményű szakértők kombinálása

Az individuális sorozatok előrejelzése (az univerzális predikció) terén ismert alapvető eredmények azt mondják ki, hogy – megfelelő, általános feltételek mellett – léteznek olyan online predikciós algoritmusok, amelyek tetszőleges sorozat esetén majdnem olyan jól tudnak

becsülni, mint az adott sorozatot legjobban előrejelző szakértő egy adott véges osztályból. Ezen algoritmusok teljesítménye (pl. ha a becslés pontosságát mérő veszteségfüggvény korlátos, és minden lépésben minden lehetséges akció teljesítményéről információt kapunk) N szakértő esetén T idő alatt csupán $O(\sqrt{T \ln N})$ -nel marad el a legjobb szakértő teljesítményétől (azaz pl. hosszú távon eléri annak átlagos teljesítményét), pl. [7, 3, 16, 9, 4]. Az a tény, hogy a korlát a szakértők számának logaritmusától függ, lehetővé teszi értelmesen nagy szakértőosztályok alkalmazását. A standard algoritmusok komplexitása azonban többnyire a szakértők számával arányos, azaz ezek nagy szakértőosztályok esetén nem alkalmazhatók.

Ugyanakkor, ha a szakértők osztálya bizonyos struktúrával rendelkezik, sokszor lehetőség nyílik az algoritmusok hatékony implementálására. Az egyik legismertebb ilyen példa az ún. tracking probléma, amikor időben változó szakértőkkel akarunk versenyezni, azaz olyan metaszakértőkkel, akik különböző időintervallumokban más és más (alap-) szakértő tanácsait követhetik [8, 17]. Ez a megközelítés jól modellezi azt a helyzetet, amikor a predikálandó sorozat tulajdonságai időben változnak. Más strukturált szakértőosztályokra példa többek között az, amikor a szakértők döntési fákkal [12, 11] vagy éppen gráfokkal jellemezhetők [15], de ide tartozik az a korábbi eredményünk is, amikor a szakértők osztálya egy adott szintszámú skalárkvantálók halmaza [5].

A kutatás során olyan módszereket dolgoztunk ki, amelyek a fenti két strukturált probléma előnyeit ötvözve kis komplexitással képesek megoldani a tracking feladatot strukturált alap-szakértőosztály esetén úgy, hogy mind a trackingben, mind az alapstruktúrában rejlő komplexitáscsökkentést kihasználják [J3] (hasonló módszereket dolgoztak ki korábban a szakaszonként stacionárius, memóriamentes források veszteségmentes tömörítése területén [19, 14, 13], azonban a mi eredményünk sok szempontból általánosabb, pl. ezekben a cikkekben mindig a logaritmusos veszteségfüggvényt vizsgálják, míg a mi eredményeink veszteségfüggvények igen általános osztályaira teljesülnek). A kidolgozott módszer előnye, hogy akkor is hatékonyan működik, ha a szakértőosztály mérete időben nő.

Fix méretű szakértőosztály mellett az algoritmus komplexitása négyzetes az optimális konvergenciasebesség elérése esetén, ugyanakkor lehetőség van lineáris komplexitás elérésére a konvergenciasebesség csökkentése mellett. A további kutatás során sikerült az algoritmus komplexitását jelentősen csökkenteni, amely így a lineáris komplexitást csupán egy logaritmusos faktorialal haladja meg, és a konvergenciasebesség is csak logaritmusos faktorokkal romlik [A]. A módszert leíró folyóiratcikk [JS2] elkészítése folyamatban van, és várhatóan április folyamán kerül beküldésre.

4. Univerzális, kis késleltetésű - kis komplexitású veszteséges adattömörítés

A fentiekben ismertetett univerzális predikációs technikák segítségével új módszereket dolgoztunk ki individuális sorozatok veszteséges tömörítésére, melyeket a [J3] cikkben publikáltunk a téma legjelentősebb szakfolyóiratában, az IEEE Transactions on Information Theory-ban.

Az individuális sorozatok témakörében nem tesszük fel, hogy az információforrás egy bizonyos sztochasztikus modell szerint viselkedik - a cél olyan algoritmusok kidolgozása, melyek minden lehetséges forrássorozatra jól működnek algoritmusok egy referenciaosztályához képest (ez jól modellezi azon valós helyzeteket, amikor nem tudjuk előre, hogy milyen típusú adatok érkeznek, azaz tetszőleges adatok tömörítésére fel kell készülni).

Az első kidolgozott módszer aszimptotikusan ugyanolyan jó tömörítést tesz lehetővé, mint a forrássorozathoz illesztett legjobb skalárkvantálási eljárás, mely során az alkalmazott kvantáló többször változhat, lehetővé téve az alkalmazkodást egy időben változó viselkedésű forráshoz. A módszer kis komplexitással implementálható, a konvergenciasebesség enyhe csökkenése árán a számításigény időben lineárisra tehető. (Az irodalomban korábban elérhető kis komplexitású módszerek csupán egy fix skalárkvantáló teljesítményével vették fel a versenyt [5], és fontos problémaként merült fel kis komplexitású, praktikus algoritmusok tervezése értelmese, elegendően nagy szakértőosztályokra [18].)

Hasonló eredményeket sikerült elérni a hálózati kvantálás területén is, ahol aszimptotikusan ugyanolyan jó tömörítést érhetünk el - lineáris komplexitás mellett - mint a legjobb időben változó, intervallum cellákkal rendelkező többszörös felbontású illetve többszörös leírású kvantálási algoritmusok. A módszert kiterjesztettük tetszőleges korlátos késleltetésű, véges memóriájú kódcsaládokra is, ám itt az egyszerű implementáció általában nem oldható meg.

A kidolgozott univerzális kvantálási módszerek hátterében az áll, hogy a tervezési feladatot - megfelelő diszkrétizálási lépések mellett - visszavezetjük a (körmentes) irányított gráfban való legrövidebb út keresésének problémájára, majd az így definiált strukturált szakértőosztályra alkalmazzuk a tracking feladatra kidolgozott predikciós módszerünket. Megjegyzendő, hogy a diszkrétizálás miatt a szakértőosztály mérete időben nő, és ez motiválta a tracking probléma hatékony megoldását nagy, strukturált szakértőosztályokra.

5. Adaptív útvonalválasztás

A hálózati kommunikáció során igen fontos, hogy a csomagokat kis késleltetéssel illetve veszteséggel továbbítsuk, melyet jelentős mértékben támogat a megfelelő útvonalválasztás. A gyakorlatban nehézséget jelent, hogy egy csomópont általában nem ismeri az egész hálózatot, csupán az általa küldött csomagok késleltetéséről/veszteségéről van információja.

A tömörítéshez hasonlóan az útvonalválasztás is megfogalmazható egy szekvenciális döntési feladatként (minden csomaghoz új utat kell választani). Ennek megfelelően, a korábbiakhoz hasonlóan az individuális sorozatos keretet alkalmazzuk, azaz olyan útvonalválasztási algoritmusokat dolgoztunk ki, melyek a hálózat tetszőleges viselkedése esetén (aszimptotikusan) optimálisan viselkednek (egy szakértőosztályhoz képest). Bár ez a megközelítés túl konzervatívnak tűnhet a standard hálózatok esetén, amikor az egyes linkeken a forgalom csupán lassan változik, de komoly létjogosultsága van például ad-hoc hálózatokban (ahol az összeköttetések minősége akár pillanatról pillanatra változhat), vagy akár a lehető legjobb működés biztosításában szolgáltatásmegtagadás-típusú (denial of service) támadások esetén.

Új helyzet azonban a 3. részben ismertetett problémához képest, hogy egy-egy döntés

után csupán a választott út (akció) teljesítményéről kapunk információt, a többitől nem. Ez a szokásos algoritmusok teljesítményét jelentősen csökkenti a korábban ismertetett ún. teljes információs esethez képest: a standard (ún. bandita-) algoritmusok a legjobb szakértő teljesítményénél N szakértő esetén átlagosan akár $O(\sqrt{N \ln N/T})$ -vel rosszabb teljesítményt nyújthatnak T idő alatt [1]. Mivel a szakértők száma a legrövidebb út probléma esetén tipikusan exponenciálisan nagy a feladat (a gráf) méretében, ebben az esetben a standard algoritmusok alkalmazását nem csak a komplexitás, hanem a gyenge teljesítmény is gátolja.

Ennek a problémának a megoldására készítettünk egy olyan algoritmust, amelynek extra vesztesége a legjobb útéhoz képest nem az utak számában, hanem a gráf méretében polinomiális, az átlagos extra veszteség az időben továbbra is optimális $O(T^{-1/2})$ nagyságrendű, továbbá az algoritmus hatékonyan megvalósítható (a korábbi kis komplexitású algoritmusoknál a konvergenciasebesség rosszabb, $O(T^{-1/3})$ [2] vagy $O(T^{-1/4})$ [10] volt). A megoldás alapja az az észrevétel, hogy mivel az egyes utak közösen használnak bizonyos linkeket, ezért egy út teljesítményéről nyert információ információval szolgál más utak teljesítményéről is. A fenti algoritmus feltételezi, hogy a linkenkénti késleltetések ismertek, azonban bizonyos körülmények között (például csomagvesztés vizsgálata esetén) realisztikusabb azt feltételezni, hogy csupán a csomag teljes útjára vonatkozó minőségi paramétereket ismerjük. Erre az esetre kidolgoztunk egy olyan módszert, melynek teljesítménye eléri a korábban publikált legjobb hasonló algoritmusét (mely a hálózat tetszőleges viselkedése esetén bizonyos értelemben optimális teljesítményt nyújt [2]), ugyanakkor a mi algoritmusunk koncepcionálisan és implementációs szempontból is lényegesen egyszerűbb. Kidolgoztuk továbbá a fenti algoritmusok hatékony trackinges változatát is, amikor nem a legjobb fix útvonal, hanem a legjobb időben (lassan) változó útvonal teljesítményét akarjuk elérni.

Az adaptív útvonalválasztási algoritmusok területén elért eredményeinket a [C1, JC] konferenciákon, illetve a téma egyik vezető folyóiratában, a Journal of Machine Learning Research-ben publikáltuk [J2].

6. Szekvenciális on-line ládapakolási probléma

A kutatás utolsó fázisában a klasszikus ládapakolási probléma egy szekvenciális változatát vizsgáltuk. Ebben az esetben a tárgyak egyenként érkeznek, és mielőtt egy tárgy méretét megtudjuk, el kell döntenünk, hogy a tárgyat az éppen nyitott ládába tesszük, vagy az aktuális ládát bezárjuk, és nyitunk egy új, üres ládát. Ha egy tárgy nem fér be, akkor elvész, továbbá ha egy ládát bezárunk, akkor az ott megmaradó üres helyet is veszteségként könyveljük el. A feladat az, hogy adott idő alatt minimális veszteséget érzünk el.

Ez a feladat több különböző ütemezési problémánál is felmerülhet. Egy érdekes példa bizonyos speciális szenzorhálózatok adásütemezése. A szenzorok méréseket végeznek, a mérési eredményeket pedig változó szóhosszúsággal tömörítik, és a memóriájukban tárolják (ha van elég szabad hely). Az eredményeket fix méretű frame-ek elküldésével kommunikálják, ám – mivel memóriájuk korlátozott – csupán egy frame-nyi adatot képesek tárolni. Mivel a szenzor energiaellátása is korlátos, minimalizálnia kell a fölöslegesen elküldött adatok mennyiségét (ezek a kitöltetlen adatokat a frame-ekben), ugyanakkor a lehető legtöbb mérési eredményt is

továbbítaniuk kell. Könnyen látható, hogy annak eldöntése, hogy egy szenzor mikor küldjön el egy frame-et, analóg azzal, hogy mikor nyissunk új ládát a szekvenciális ládapakolási feladatban.

A feladat nehézségét az adja, hogy a standard predikciós feladatoktól eltérően itt a veszteségfüggvénynek memóriája van, és az állapottér végtelen. Kidolgoztunk egy algoritmust, melynek átlagos teljesítménye lényegében ugyanolyan jó, mint a legjobb algoritmusé egy véges referenciaosztályból. Egy természetesen adódó referenciaosztály a küszöb-stratégiák osztálya, melyek akkor nyitnak új ládát, ha a rendelkezésre álló üres hely egy korlátnál kisebb. Ezen referencia-algoritmusosztály esetén bebizonyítottuk, hogy algoritmusok egy igen nagy osztályában nem létezik univerzálisan jó algoritmus minden tárgysorozatra. Ugyanakkor mutattunk egy algoritmust, amely univerzálisan jó teljesítményt nyújt akkor, ha a tárgyak mérete legalább kis mértékben véletlen. Az eredményeket leíró folyóiratcikket márciusban küldtük be a *Journal of Machine Learning Research*-hez [JS1], illetve egy részüket korábban publikáltuk a [C2] konferencián. Jelentőségüket növeli, hogy tudomásunk szerint ez az első publikált nem-triviális probléma memóriával rendelkező, végtelen állapotterű veszteségfüggvények esetére.

Publikációk

- [J1] D. Nagy, A. György, and T. Linder. Symbol-based modeling and coding of block markov sources. *IEEE Trans. Inform. Theory*, 52(12):5570–5578, Dec. 2006. IF=2.18
- [J2] A. György, T. Linder, G. Lugosi, and Gy. Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 58:2369–2403, Oct. 2007. IF=2.68
- [J3] A. György, T. Linder, and G. Lugosi. Tracking the best quantizer. *IEEE Transactions on Information Theory*, 54:1604–1625, Apr. 2008. IF=2.32
- [JC] A. György, T. Linder, and Gy. Ottucsák. The shortest path problem under partial monitoring. In *Proceedings of the 19th Annual Conference on Learning Theory, COLT 2006, LNAI 4005*, pages 468–482, Pittsburgh, Pennsylvania, USA, June 2006. Springer. IF=0.300
- [C1] A. György, T. Linder, and G. Lugosi. The shortest path problem in the bandit setting. In *Proc. 2006 IEEE Information Theory Workshop*, pages 87–91, Punta del Este, Uruguay, Mar. 2006.
- [C2] A. György, G. Lugosi, and Gy. Ottucsák. On-line sequential bin packing. In *Proceedings of the 21st Annual Conference on Learning Theory, COLT 2008*, pages 447–454, Helsinki, Finland, July 2008.
- [A] A. György, T. Linder, and G. Lugosi. Efficient tracking of the best of many experts. *Information and Communication Conference*, pages 3–4, Budapest, Aug. 25–28 2008.

- [JS1] A. György, G. Lugosi, and Gy. Ottucsák. On-line sequential bin packing. *Journal of Machine Learning Research*, March 2009, submitted.
- [JS2] A. György, T. Linder, and G. Lugosi. Efficient tracking of the best of many experts. *IEEE Transactions on Information Theory*, in preparation.

Hivatkozások

- [1] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.
- [2] B. Awerbuch and R. D. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In *Proceedings of the 36th Annual ACM Symposium on the Theory of Computing, STOC 2004*, pages 45–53, Chicago, IL, USA, Jun. 2004. ACM Press.
- [3] D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.
- [4] N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, D. Haussler, R. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.
- [5] A. György, T. Linder, and G. Lugosi. Efficient algorithms and minimax bounds for zero-delay lossy source coding. *IEEE Transactions on Signal Processing*, 52:2337–2347, Aug. 2004.
- [6] A. György, T. Linder, and G. Lugosi. Tracking the best of many experts. In *Proceedings of the 18th Annual Conference on Learning Theory, COLT 2005*, pages 204–216, Bertinoro, Italy, Jun. 2005. Springer.
- [7] J. Hannan. Approximation to bayes risk in repeated plays. In M. Dresher, A. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games*, volume 3, pages 97–139. Princeton University Press, 1957.
- [8] M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.
- [9] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.
- [10] H. B. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *Proceedings of the 17th Annual Conference on Learning Theory, COLT 2004*, pages 109–123, Banff, Canada, Jul. 2004. Springer.
- [11] F. Pereira and Y. Singer. An efficient extension to mixture techniques for prediction and decision trees. *Machine Learning*, 36:183–199, 1999.

- [12] R. E. Schapire and D. P. Helmbold. Predicting nearly as well as the best pruning of a decision tree. *Machine Learning*, 27:51–68, 1997.
- [13] G. I. Shamir and D. J. Costello, Jr. Asymptotically optimal low-complexity sequential lossless coding for piecewise-stationary memoryless sources – part i: The regular case. *IEEE Trans. Inform. Theory*, IT-46:2444–2467, Nov. 2000.
- [14] G. I. Shamir and N. Merhav. Low-complexity sequential lossless coding for piecewise-stationary memoryless sources. *IEEE Trans. Inform. Theory*, IT-45:1498–1519, July 1999.
- [15] E. Takimoto and M. K. Warmuth. Predicting nearly as well as the best pruning of a planar decision graph. *Theoretical Computer Science*, 288:217–235, 2002.
- [16] V. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 372–383, Rochester, NY, Aug. 1990. Morgan Kaufmann.
- [17] V. Vovk. Derandomizing stochastic prediction strategies. *Machine Learning*, 35(3):247–282, Jun. 1999.
- [18] T. Weissman and N. Merhav. On limited-delay lossy coding and filtering of individual sequences. *IEEE Trans. Inform. Theory*, 48:721–733, Mar. 2002.
- [19] F. M. J. Willems. Coding for a binary independent piecewise-identically-distributed source. *IEEE Trans. Inform. Theory*, IT-42:2210–2217, Nov. 1996.