Support vektor alapú tanulás alkalmazásai

Ormándi Róbert

A témavezetők

Prof. János Csirik és *Dr. Márk Jelasity*

Magyar Tudományos Akadémia és a Szegedi Tudományegyetem Mesterséges Intelligencia Kutatócsoportja

Informatika Doktori Iskola

Szegedi Tudományegyetem

Doktori értekezés összefoglalója

Szeged

2013

Motiváció

Egyre több és több adat halmozódik fel körülöttünk. Emellett olyan szoftvereszközök is könnyedén a rendelkezésünkre állnak, amelyek segítségével rendkívül nagy mennyiségű (large scale) adat is kezelhető, továbbá az adatok tárolása is egyre olcsóbbá válik. Ez a jelenség—habár a gépi tanulás problémája már régóta alapvetőnek számít—tovább növeli az olyan gépi tanuló algoritmusok iránt támasztott igényeket, amelyek *megfelelően* működnek a különféle feladatokon, valamit a megszokottól eltérő számítási környezetekben (pl. elosztott rendszerekben) is *hatékonyan* használhatók; hiszen ezek nélkül az algoritmusok nélkül egyszerűen nem tudunk hasznos információt kinyerni a növekvő mennyiségű adatokból. Azaz a gépi tanulás megfelelő alkalmazása nélkül egyszerűen nem tudjuk kiaknázni a nagy mennyiségű adatokban fellelhető információt. Mindazonáltal a fenti célok elérése továbbra is kihívás, hiszen a tanuló algoritmusok nem megfelelő adaptációja a különféle feladatokra könnyedén olyan modelleket eredményezhet, amelyek messze elmaradnak az optimális modellek teljesítményétől. Másfelől az algoritmusok naiv adaptációja a különféle számítási környezetekhez könnyedén nem várt hatásokat válthat ki a rendszerben, amely az említett algoritmusokat futtatja (mint nagy, kiegyensúlyozatlan terhelés egy elosztott rendszerben).

A tézis az úgynevezett *support vektor alapú tanuló algoritmusokkal* foglalkozik. Az ehhez az algoritmuscsaládhoz tartozó algoritmusok az úgynevezett *maximális margó heurisztika* ötletét használják fel. A heurisztika alapötlete meglehetősen egyszerű: keressünk egy jó elválasztó felületet (a legegyszerűbb esetben hipersíkot), amely mindezek mellett maximalizálja a margót (vagyis a távolságot a legközelebbi eltérő osztályba tartozó tanulópéldák között) is. Ennek az

Table 1: A tézis fejezetei és a hivatkozott saját publikációk közötti kapcsolat (• jelöli az *alap* publikációkat, amíg ○ a kapcsolódóakat).

	3. fejezet	4. fejezet	5. fejezet	6. fejezet	7. fejezet
ICDM 2008 [5]	•				
TSD 2010 [7]		•			
EUROPAR 2010 [8]			•		
WETICE 2010 [6]			•		
EUROPAR 2011 [9]				•	•
CCPE 2012 [10]					•
SASO 2012 [4]				○	○
SISY 2012 [2]				○	○
EUROPAR 2012 [3]				○	○
ICML 2013 [11]				○	○

alapötletnek számos praktikus alkalmazása, valamint formalizmusa ismert.

A tézis célja, hogy különböző megközelítéseket mutasson be a fenti célok (azaz megfelelő adaptáció elérése mind az algoritmikus, mind a rendszermodell-specifikus aspektusok tekintetében) elérésére a support vektor alapú tanuló algoritmusokhoz kapcsolódóan. Azaz a support vektor alapú tanulók adaptálhatóságát vizsgáljuk különböző feladatokban és számítási környezetekben. A tézis fő üzenete a következőképpen foglalható össze. Erőteljes alapötletek (heurisztikák), mint építőkövek, valamint megfelelő tervezés felhasználásával jelentős hatékonyságnövelés érhető el.

A tézisben tárgyalt fő rendszermodell az úgynevezett teljesen elosztott környezet. Ez a rendszermodell nagy számú számítási egységből, csomópontokból áll. Ezek a csomópontok lokális számítások elvégzésére, valamint hálózati kommunikációra képesek. Feltételezzük, hogy a rendszer nagyon nagy számú csomópontból, tipikusan személyi számítógépekből, úgymint PC-ből, mobiltelefonokból és tabletekből áll. Nem tételezünk fel homogenitást a csomópontok között. Azaz különbözők lehetnek számítási kapacitásuk, operációs rendszerük, valamint bármilyen más jellemzőjük tekintetében. Mindazonáltal feltesszük, hogy a kommunikáció üzenetküldéssel történik bármiféle központi irányítás felhasználása nélkül (teljesen elosztott aspektus). Azaz minden csomópont küldhet üzenetet bármelyik csomópontnak, feltéve hogy a célcsomópont címe lokálisan ismert. Ezek az üzenetek tetszőlegesen késhetnek, elveszhetnek valamint a felhasználók bármikor ki-, illetve beléphetnek a rendszerbe. Azaz a kilépő és a meghibásodott csomópontok ugyanúgy kezelhetők. Feltesszük, hogy a kilépett csomópontok újracsatlakozhatnak és amíg nem elérhetők, megtartják az állapotinformációjukat.

A tézis egy áttekintő fejezettel (2. fejezet) indul, amely összegzi a megértéshez szükséges

háttérinformációkat a support vektor alapú tanulást és a teljesen elosztott rendszereket illetően. A tézis fő része (3-7. fejezetek) két részre osztható. Az első rész (3-5. fejezet) a support vektor alapú tanulás *algoritmikus adaptivitását* vizsgálja. Itt arra fókuszálunk, hogy a support vektor alapú tanulás alapötletét, a maximális margó heurisztikát, hogyan tudjuk különféle alkalmazásokhoz adaptálni, mint a idősorelemzés (3. fejezet), doménadaptáció (4. fejezet) és ajánlórendszerek (5. fejezet).

A tézis második felében (5-7. fejezetek) folyamatosan elkezdjük vizsgálni az *adaptáció rendszermodell aspektusát*. Ebben a részben elsősorban arra a kérdésre keressük a választ, hogy teljesen elosztott környezetben hogyan implementálható egy support vektor alapú tanuló. A 6. fejezetben egy pletykaalapú support vektor implementációt, a P2PEGASOS algoritmust vezetjük be. Ezután az utolsó fejezetben (7. fejezet) tovább javítjuk a P2PEGASOS konvergenciasebességét. Az itt létrehozott algoritmusok konvergenciájának bizonyítását is közöljük.

A tézis minden egyes fejezete legalább egy elfogadott nemzetközi publikációra épül. A fejezetek és a publikációk közötti kapcsolatot az 1. táblázat foglalja össze[1].

[1] A publikációk teljes listája a weblapom megfelelő oldalán érhető el: `http://www.inf.u-szeged.hu/~ormandi/papers`

A tézis eredményeinek összefoglalása

Ebben a fejezetben áttekintjük a tézis fő céljait és eredményeit. Ennek során egy rövid összegzést adunk minden fejezetről (3-7. fejezetek). Az összegzés során felsorolás jelleggel kiemeljük az adott fejezet főbb hozzájárulásait és azok hatását.

VMLS-SVM idősorelemzésre

A 3. fejezetben az idősorelemzés problémáját vizsgáltuk, vagyis azt, hogy hogyan tudjuk kiterjeszteni a Least Squares SVM-eket, hogy az idősorok előrejelzésére megfelelőbb algoritmust kapjunk. Az általunk ajánlott algoritmus (VMLS-SVM) egy súlyozott variancia taggal bővíti az LS-SVM eredeti célfüggvényét. A módosítás alapötlete azon az előzetes megfigyelésen alapul, hogy két, azonos predikciós teljesítménnyel rendelkező, idősorelemzésre alkalmas modell közül a kisebb varianciával illesztő modell nyújt összegzett teljesítményt a túlillesztéstől (overfitting) eltekintve. Az ajánlott módszer az LS-SVM egy kiterjesztésének tekinthető, amely megtartja az eredeti algoritmus összes előnyét, mint pl. a kernel-trükk alkalmazhatósága és lineáris megoldás. Mindamellett egy új hiperparamétert hoz be az algoritmus, ami megnehezíti a módszer finomhangolását.

A fejezet első felében áttekintjük a kapcsolódó módszereket, valamint az eredeti LS-SVM módszer néhány tulajdonságát. Azután az ajánlott módszer részletes leírása következik, beleértve annak matematikai hátterét. Majd egy módszert ajánlunk a bevezetett hiperparaméter miatt bekövetkezett paraméterhangolás kezelésére. Ezután egy alapos empírikus kiértékelés

zárja a fejezetet. A kiértékelés alapján azt mondhatjuk, hogy az ajánlott módszer—megfelelő paraméterhangolást feltételezve—szignifikánsan jobb teljesítményt képes elérni számos baseline algoritmus ellen. A kiértékeléseket három, széles körben használt benchmark adatbázison lettek végeztük el.

Főbb hozzájárulások és azok hatása:

- A VMLS-SVM elméleti és algoritmikus levezetése
- Egy, a hiperparaméterek finomhangolására alkalmas módszer
- Az algoritmus szignifikánsan felülmúl számos baseline módszert három elterjedt benchmark adatbázison

A fejezet eredményei a korábbi [5] publikáción alapulnak.

Transzformációalapú doménadaptáció

A 4. fejezet a doménadaptáció problémáját vizsgálta véleménydetekciós feladaton. A fejezetben egy általános keretrendszert ajánlunk (DOMÉN MAPPING LEARNER (DML)) és annak két példányosítását: az első SVM-alapú modelleket használ forrásmodellként, míg a második logisztikus regresszió (LR) alapú tanulókat alkalmaz mint forrásmodell. Az általános módszer alapötlete a következőképpen foglalható össze: a forrás- és céldomének között fennálló relációt egy olyan modelltranszformációs mechanizmussal modellezi, amely egy, a céldoménből származó, nagyon kis méretű tanuló adatbázis alapján tanulható.

A fejezetben röviden áttekintjük a doménadaptáció feladatához kapcsolódó algoritmusokat. Azután megadjuk a probléma transzformációalapú formális definícióját, valamint bevezettünk egy újszerű, absztrakt, transzformációalapú megközelítést, a DML algoritmust. Ezután részletesen tárgyaljuk egy SVM-alapú, valamint egy LR-alapú példányosítását az alap algoritmusnak. Az ezt követő empirikus kiértékelések validálták azt, hogy az általunk javasolt módszerek képesek jó modellek előállítására nagyon kis számú címkézett tanulóadat felhasználásával is a céldoménből. Továbbá ez a jelenség akkor is fennáll, amikor van elegendő tanulópélda a céldoménből, de a baseline módszerek nem képesek jó általánosítási tulajdonságot felmutatni. Ezekből a kiértékelésekből arra a következtetésre jutottunk, hogy az SVM-alapú példányosítás az előnyösebb választás az LR-alapú megközelítéssel szemben, mivel az előbbi robosztusabb a transzformáció tanulására.

Főbb hozzájárulások és azok hatása:

- A doménadaptációs feladat transzformációalapú formalizmusa.
- Az általános DML megközelítés.
- Az általános megközelítés két példányosítása (SVM és LR alapú megközelítések)
- Az ajánlott algoritmusok jobb teljesítményt érnek el, mint a direkt megközelítés (baseline) és két state-of-the-art algoritmus (SCL és SCL-MI).

A fejezetben bemutatott eredmények egy korábbi munkánkon alapulnak [7].

SVM-mel támogatott elosztott ajánlás

Ebben a fejezetben (5. fejezet) az elosztott ajánlás problémájával foglalkoztunk. A fejezet célja kettős. Először bemutatja és motiválja az implicit felhasználói tevékenységekből történő értékelések előrejelzését. Ezután két újszerű heurisztikus eljárást (direkt és időcsúsztatásos módszer) vezet be a probléma megoldására. Ennek során egy érdekes, mindazonáltal rendhagyó alkalmazását mutattuk be az SVM-nek az SMO algoritmus használatával. Arra használjuk az SVM-et, hogy validáljuk a heurisztikáinkat és empirikusan bizonyítsuk azt, hogy az előrejelzett adatbázisnak "érdekes" tulajdonságai vannak, továbbá a szofisztikáltabb heurisztika olyan adatbázist eredményez, amely még inkább érdekes belső struktúrával rendelkezik. Másodsorban újszerű overlay kezelő protokollokat vezettünk be, amelyek a teljesen elosztott collaborative filtering (CF) alapú ajánló algoritmusok teljesen elosztott környezetben történő hatékony megvalósítását teszik lehetővé.

A fejezetben először áttekintettük mind a központosított, mind az elosztott CF eljárásokat. Ezután bevezettünk a következtető heurisztikákat, a tanulhatóságon alapuló validációs módszert és elvégeztük azt az SMO (egy SVM-megvalósítás) algoritmus használatával. Később bemutattuk azt, hogy a legtöbb ajánló adatbázis bemeneti éleinek eloszlása közel hatvány eloszlást mutat, ami komoly problémákat okozhat az elosztott környezetben történő CF implementációk működése során. A problémák leküzdése érdekében bevezettünk néhány overlay kezelő protokollt (véletlen példa alapú kNN és annak T-MAN alapú megközelítései (GLOBAL, VIEW, BEST és PROPORTIONAL), valamint ezek randomizált változatai) és alapos empirikus kiértékelést végeztünk. A kiértékelésben az általunk ajánlott módszerek egymás ellen, valamint a BUDDYCAST nevű baseline protokoll ellen vettek részt. A fejezet végén többszörös konklúziót tudtunk megállapítani: az agresszív peer választás kezelhetetlen hálózati terhelést okozhat, továbbá az ajánlott, T-MAN alapú megközelítés a GLOBAL peer választási stratégiával kom-

binálva jó választás, hiszen teljesen uniform hálózati terhelés mellett, elfogadható konvergenciasebességet biztosít.

Főbb hozzájárulások és azok hatása:

- Előrejelzési heurisztikák (direkt és időcsúsztatásos módszer).
- Tanulhatóságalapú indirekt validációs technika.
- A FileList.org-ból kinyert ajánló adatbázis [1].
- Overlay kezelő megközelítések: véletlen minta alapú kNN megközelítés, T-MAN alapú variánsok (GLOBAL, VIEW, BEST és PROPORTIONAL) és azok randomizált változatai.
- A GLOBAL peer választáson alapuló technika megfelelő kompromisszumot eredményez a konvergenciasebesség és a hálózati terhelés között.

A fejezet eredményei a korábbi [6, 8] publikációkon alapulnak.

P2PEGASOS—Egy teljesen elosztott SVM

A 6. fejezetben a teljesen elosztott adatbányászat problémájára fókuszáltunk. A célunk egy általános SVM-alapú algoritmus kidolgozása volt, amely egy teljesen elosztott hálózatban jó minőségű modellek tanulására képes elfogadható kommunikációs komplexitás mellett, mialatt a lehető legkevesebb elvárást támasztjuk a kommunikációs modellel szemben. Az ajánlott algoritmus—a P2PEGASOS algoritmus—egy koncepcionálisan egyszerű, mégis erőteljes SVM-implementáció. Az alapötlet szerint sok modell végez véletlen sétát a hálózatban, mialatt egy grádiens alapú, online tanuló szabály alkalmazásával javítják magukat.

A fejezet elején definiáltuk a teljesen elosztott rendszermodellt. Aztuán áttekintettük az Pegasos SVM solver általános tulajdonságait és a kapcsolódó teljesen elosztott tanuló algoritmusokat. Az általunk javasolt algoritmus részletes tárgyalása után annak alapos kiértékelése következik. Ennek során számos, főként centralizált SVM algoritmus ellen vizsgáltuk az ajánlott módszer konvergenciasebességét több benchmark adatbázison. A kiértékelések során több szimulációs forgatókönyvet vizsgáltunk, köztük számos olyannal, amely extrém hálózati hibákat modellezett. Ezek a kiértékelések kimutatták, hogy az ajánlott algoritmus rendkívül robosztus a különféle hálózati hibák ellen, mialatt jó modelleket biztosít elfogadható hálózati terhelés mellett.

Főbb hozzájárulások és azok hatása:

- A P2PEGASOS algoritmus.

- Szavazásalapú mechanizmus a predikciós teljesítmény javítására.
- Az algoritmus meglepően jó konvergenciatulajdonságokat mutat még az extrém hibákkal terhelt szimulációkban is.

A fejezetben közölt eredmények főként egy korábbi munkánkra [9] támaszkodnak.

A P2PEGASOS konvergenciájának további gyorsítása

A 7. fejezetben folytattuk a teljesen elosztott környezetben történő tanulás vizsgálatát, és egy olyan mechanizmust dolgoztunk ki, amely közel egy nagyságrenddel gyorsítja a P2PEGASOS algoritmus konvergenciasebességét. A létrejött algoritmusok a P2PEGASOSMU és a P2PEGASOSUM. Az alapötlet az, hogy vezessünk be egy modellkombinációs eljárást, amely átlagolja a találkozó modelleket. A fejezetben bemutattuk, hogy az ajánlott eljárás az Adaline modell esetében éppen úgy viselkedik, mintha exponenciálisan növekvő számú modellt gyűjtenénk és szavaztatnánk a predikció során, de konstans tárral. Továbbá megmutattuk, hogy habár a P2PEGASOS algoritmus esetén az egzakt ekvivalencia nem valósul meg, a viselkedése nagyon hasonló. A P2PEGASOSMU algoritmus konvergenciájának bizonyítását is közöljük.

A fejezet a teljesen elosztott adatmodell motivációjával indul, azután a modellkombinációs eljárások áttekintése következik. Majd az ajánlott módszer részletes algoritmikus leírása következik, amely tartalmazza az Adaline és a Pegasos modellekkel történő példányosítás bemutatását és a P2PEGASOSMU algoritmus konvergenciájának bizonyítását. Az empirikus kiértékelés kimutatta azt, hogy az ajánlott módszer közel egy nagyságrenddel megnöveli a konvergenciasebességet az eredeti P2PEGASOS algoritmushoz képest, mialatt annak összes előnyét megtartja. Kimutattuk továbbá, hogy a P2PEGASOSMU algoritmus jobban fenntartja a modellek közötti függetlenséget, ami miatt ez a kívánatosabb választás a két algoritmus közül.

Főbb hozzájárulások és azok hatása:

- Az egyesítő (merging) mechanizmus a P2PEGASOS algoritmushoz, ami a P2PEGASOSMU és P2PEGASOSUM algoritmusokat eredményezi.
- A P2PEGASOSMU algoritmus konvergenciájának bizonyítása.
- Az algoritmusok jelentősen gyorsabb konvergenciasebességet mutatnak, mint az eredeti P2PEGASOS, mialatt megőrzik annak összes előnyös tulajdonságát.

A fejezet eredményei egy korábbi munkánkon [10] alapulnak.

A téziseredmények összesítése

A tézis főbb eredményei az alábbiakban foglalhatók össze:

- A VMLS-SVM algoritmus bevezetése a 3. fejezetben (eredetileg a [5] cikkben publikálva).
- Paraméteroptimalizációs eljárás a VMLS-SVM algoritmus részére a 3. fejezetben (eredetileg a [5] cikkben publikálva).
- A doménadaptációs eljárás transzformációalapú formalizmusa a 4. fejezetben (eredetileg a [7] cikkben publikálva).
- Az általános DML algoritmus bevezetése a 4. fejezetben (eredetileg a [7] publikációban közölve).
- A DML algoritmus SVM- és LR-alapú példányosítása a 4. fejezetben bevezetve (eredetileg a [7] cikkben közzétéve).
- A következtető heurisztikák (direkt és időcsúsztatás alapú megközelítések) az 5. fejezetben (eredetileg a [6] cikkben publikálva).
- Tanulhatóságalapú validációs technika az SMO algoritmus használatával az 5. fejezetben (eredetileg a [6] cikkben közölve).
- Overlay kezelő megközelítések: véletlen példa alapú kNN megközelítés, T-MAN-alapú variánsok (GLOBAL, VIEW, BEST és PROPORTIONAL) és ezek véletlenített változatai az 5. fejezetben (eredetileg a [8] publikációban közzétéve).
- A P2PEGASOS teljesen elosztott SVM tanuló a 6. fejezetben (eredetileg a [9] cikkben publikálva).
- Szavaztatásalapú megközelítés a P2PEGASOS SVM teljesítményének javítására a 6. fejezetben (eredetileg a [9] cikkben bevezetve).
- A modellkombinációs mechanizmus a P2PEGASOS SVM számára, amely a P2PEGASOSMU és P2PEGASOSUM algoritmusokat eredményezi, valamint maguk az algoritmusok a 7. fejezetben (eredetileg a [10] cikkben közölve).
- A P2PEGASOSMU algoritmus konvergenciájának bizonyítása a 7. fejezetben (eredetileg a [10] publikációban közzétéve).

Hivatkozások

[1] The FileList.org-based inferred recommendation dataset. `http://www.inf.u-szeged.hu/rgai/recommendation/`.

[2] István Hegedűs, Nyers Lehel, and Ormándi Róbert. Detecting concept drift in fully distributed environments. In *2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics*, SISY'12, pages 183–188. IEEE, 2012.

[3] István Hegedűs, Busa-Fekete Róbert, Ormándi Róbert, Jelasity Márk, and Kégl Balázs. Peer-to-peer multi-class boosting. In *Euro-Par 2012 Parallel Processing*, volume 7484 of *Lecture Notes in Computer Science*, pages 389–400. Springer Berlin / Heidelberg, 2012.

[4] István Hegedűs, Ormándi Róbert, and Jelasity Márk. Gossip-based learning under drifting concepts in fully distributed networks. In *2012 IEEE Sixth International Conference on Self-Adaptive and Self-Organizing Systems*, SASO'12, pages 79–88. IEEE, 2012.

[5] Róbert Ormándi. Variance minimization least squares support vector machines for time series analysis. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 965–970, Washington, DC, USA, 2008. IEEE Computer Society.

[6] Róbert Ormándi, István Hegedűs, Kornél Csernai, and Márk Jelasity. Towards inferring ratings from user behavior in bittorrent communities. In *Proceedings of the 2010 19th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises*, WETICE '10, pages 217–222, Washington, DC, USA, 2010. IEEE Computer Society.

[7] Róbert Ormándi, István Hegedűs, and Richárd Farkas. Opinion mining by transformation-based domain adaptation. In *Proceedings of the 13th international conference on Text, speech and dialogue*, TSD'10, pages 157–164, Berlin, Heidelberg, 2010. Springer-Verlag.

[8] Róbert Ormándi, István Hegedűs, and Márk Jelasity. Overlay management for fully distributed user-based collaborative filtering. In *Proceedings of the 16th international Euro-Par conference on Parallel processing: Part I*, EuroPar'10, pages 446–457, Berlin, Heidelberg, 2010. Springer-Verlag.

[9] Róbert Ormándi, István Hegedűs, and Márk Jelasity. Asynchronous peer-to-peer data mining with stochastic gradient descent. In *17th International European Conference on Parallel and Distributed Computing (Euro-Par 2011)*, volume 6852 of *Lecture Notes in Computer Science*, pages 528–540. Springer, 2011.

[10] Róbert Ormándi, István Hegedűs, and Márk Jelasity. Gossip learning with linear models on fully distributed data. *Concurrency and Computation: Practice and Experience*, pages n/a–n/a, 2012.

[11] Balázs Szörényi, Róbert Busa-Fekete, István Hegedűs, Ormándi Róbert, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *Proceedings of The 30th International Conference on Machine Learning (ICML), 3rd Cycle*, volume 28 of *JMLR: Workshop and Conference Proceedings*, pages 19–27. JMLR: W&CP, 2013.