

Szekvenciajelölés gráfalapú, részben felügyelt tanulási módszerrel

Molnár Gábor József¹, Farkas Richárd²

¹Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
6720, Szeged, Árpád tér 2.
gjmolnar@inf.u-szeged.hu

²MTA-SZTE, Mesterséges Intelligencia Kutatócsoport,
6720 Szeged, Tisza Lajos krt. 103. III. lépcsőház
rfarkas@inf.u-szeged.hu

Kivonat: A felügyelt tanulás fő problémája, hogy az egyedek kézi jelölése költséges és időigényes. Ez különösen igaz a szekvenciajelölés esetében, ahol egy tanítóhalmaz elkészítése több ezer token átvizsgálását igényli. Természetesen adódik az az igény, hogy olyan módszereket dolgozzunk ki, amelyek kevesebb tanítópélda ellenére is megfelelő modellt képesek építeni. Továbbá a klasszikus, szekvenciajelölésre használt algoritmusok kis méretű tanítóhalmazokon legtöbbször rosszul teljesítenek. Ezzel szemben a részben felügyelt tanulás éppen az előző igénynek próbál eleget tenni. Kísérleteinkben arra igyekeztünk rámutatni, hogy kis számú tanítópéldán alkalmazva a gráfalapú, részben felügyelt tanulási módszereket, azok jobb eredményt érnek el, mint a manapság gyakran alkalmazott szekvenciajelölők.

1 Bevezetés

Számos valós életbeli osztályozási probléma létezik, amelyekhez nem áll rendelkezésre megfelelő egyedszámú tanítóhalmaz. Az egyedek manuális jelölése gyakran költséges és időigényes. Ez különösen igaz a természetes nyelvi feldolgozás problémáinál, pl. a szekvenciajelölésnél, ahol gyakran több százezer tokenes tanítóadatbázisra van szükség. A probléma megoldására, a részben felügyelt tanulás módszere kínálhat megoldást.

Részben felügyelt esetben jelölt és jelöletlen példáink is vannak. Célunk a jelöletlen példák közötti mintázatok felismerésének segítségével, és a jelölt adatokból származó információ felhasználásával jelöléseket hozzárendelni a jelöletlen példákhoz. Azt várjuk, hogy ilyen módon kevesebb jelölt példa mellett is tanulható megfelelő pontosságú modell. A részben felügyelt tanulási technikákról egy kitűnő áttekintést ad [4].

A részben felügyelt tanulás egyik legfiatalabb részterületei a gráf alapú módszerek [1]. Ebben az esetben az egyedek alkotják a gráf pontjait, a gráf élei pedig a köztük lévő hasonlóságot reprezentálják. Ezeknél a módszereknél a kiértékelő adatbázist is felhasználjuk annak jelölései nélkül, hiszen a célunk nem az ismeretlen példákat jól klasszifikáló modell építése (induktív megközelítés), hanem a kiértékelő adatbázis felcímkézése (transzduktív megközelítés).

2 Szekvenciajelölés gráfok felhasználásával

Szekvenciajelölésen egy olyan osztályozási problémát értünk, ahol egyedek (tokenek) sorozatához (szekvenciához) rendelünk jelöléssorozatot. Tipikus szekvenciajelölési probléma a tulajdonnév-felismerés, ahol a mondatok szavait jelöljük be aszerint, hogy azok mely tulajdonnévosztályba tartoznak. Ebben a cikkben egy adott tulajdonnévosztályt jelöltünk szekvenciákban (bináris szekvenciajelölés), azaz mondatokban. A problémát a gráfalapú részben felügyelt tanulási paradigmába illesztve a gráf pontjainak a szekvenciák tokenjei felelnek meg. A tokenek között két éltípust különböztetünk meg: egyrészt, hogy megtartsuk a tokenek sorrendiségét és szekvenciához tartozását, az egyes tokeneket összekötöttük az őket megelőző és a rákövetkező tokennel; másrészt az előző pontban említett hasonlóság reprezentálására szolgálnak. Ez a módszer szoros összefüggésben áll a skip-chain CRF-fel [3], ami azt a tényt használja ki, hogy ha egy token többször fordul elő a dokumentumban, akkor az előfordulások nagy valószínűséggel ugyanabból az osztályból származnak, ezért kombinálja az azonos előfordulások jellemzőit, és olyan címkézésre törekszik, amely az ismétlődő tokeneket azonosnak tekinti. Ezzel szemben a gráfalapú részben felügyelt tanulási módszerek nemcsak az azonos előfordulások, hanem az aktuális tokenhez leghasonlóbb tokenek jellemzőit is képesek felhasználni azáltal, hogy a gráfban ezek a tokenek szomszédsági kapcsolatban állnak.

A gráf pontjainak felcímkézését egyszerű propagáló algoritmusokkal végeztük [1]. Propagálás során az a célunk, hogy a tanítópéldák jelöléseit eljuttassuk a szomszédos gráfpontokon keresztül a jelöletlen pontokhoz, a példákat összekötő élek súlyait figyelembe véve.

3 Módszer

KNN-gráfot használtunk, amelyben egy adott pontból csak a K leghasonlóbb szomszédba megy el. A KNN-gráf felépítésének időigénye ($O(n \cdot \log n)$) kisebb, mintha teljes gráfot építenénk fel ($O(n^2)$), és tárigénye is kevesebb ($O(n^2)$ helyett $O(K \cdot n)$). Mindezek ellenére a KNN-gráf használata újszerű megközelítés, hiszen a publikált rendszerek jelentős része teljes gráfokat használ. Érdeemes megjegyezni, hogy – ennek következményeként – a magukat kimondottan nagy adatbázisokon működőnek valló algoritmusok is csak néhány ezer pontra működnek elfogadható ideig [2].

A gráf pontjai közt értelmezett hasonlósági metrikát a Hamming-távolságból származtattuk: két token jellemzővektorát véve nem az eltérések, hanem az egyezések számát tekintettük. A gráf építése során a jellemzők súlyozásra kerültek az alapján, hogy az adott jellemző csak az osztályozandó tulajdonnevek (CC), a tulajdonnevek és nem tulajdonnevek (NC) vagy csak nem tulajdonnevek között fordul elő (NN). Minden jellemzőre összeszámoltuk, hogy hányszor fordul elő az egyes csoportokban. A gráf legközelebbi szomszédjának keresésekor két pont jellemzőinek metszetét véve a hasonlóságok (w) megadására kétféle módszert használtunk:

1. Csak azokat a jellemzőket vettük számításba, amelyek a tulajdonnevek között szerepeltek:

$$w = CC. \quad (0)$$

2. A hasonlóságot az alábbi csoportok gyakoriságát felhasználó képlet segítségével adtuk meg:

$$w = CC*(1-NC)*(1-NN). \quad (1)$$

Az általunk használt algoritmus a label propagation volt [1]. Ez minden iterációban frissíti a pontok címkéjét a következő képlet szerint:

$$y_i^{t+1} = \frac{\sum_{j \in K_i} (w_{ij} * y_j^t)}{\sum_{j \in K_i} w_{ij}}. \quad (2)$$

(2)-ben y_i^{t+1} jelöli az i . pont címkéjét a $(t+1)$. iterációban; K_i az i . pontból kimenő élek végpontjainak halmazát; w_{ij} pedig az i . pontból a j . pontba tartó él súlyát. Az iterációk után minden ponthoz y_i szerint rendelünk címkét.

Propagálás során a gráf élsúlyait a szomszédos pontok címkéjének a.priorijával is normáltuk (CMN) [1]. Ezzel a módszerrel igyekeztünk kiküszöbölni azt a problémát, hogy a szekvenciákban előforduló pozitív példák (tulajdonnevek) száma lényegesen kisebb, mint a negatív példáké. Az alábbi képlet szerint normalizáltunk:

$$\hat{w}_{ij} = w_{ij} + \frac{\lambda}{p(y_j)}. \quad (3)$$

(3)-ban \hat{w}_{ij} az élsúly normalizáció utáni értékét; $p(y_j)$ az y_j -nek megfelelő címke a.priori értékét jelenti; λ pedig egy normalizációs tényező, ahol $\lambda \in [0;1]$.

4 Kísérletek, tapasztalatok

Kísérleteinket a Reuters hírkorpuszon végeztük. A korpusz tanítóhalmazának 3000 pontját választottuk ki véletlenszerűen. A tanítóadatbázisban négyféle tulajdonnév-osztály került felcímkezésre (személyek, szervezetek, helyek, egyéb). Egy tesztelés alatt csak egy adott osztályra fókuszáltunk, a többi osztályba tartozó tokeneket ekkor nem kezeltük tulajdonnévként. A tesztekhez a korpusz kiértékelő adatbázisának 3000 véletlenszerűen választott pontját használtuk fel. A gráfban a K értékét 10-nek választottuk meg. Az algoritmusok kiértékelése során használt referenciaalgoritmusnak a CRF valószínűségi tanulót használtuk [3]. Kétféle paraméterrel kísérleteztünk:

1. A mondatokon belül szomszédos tokenek közötti éleket (tokenélek) súlyoztuk egy konstans értékkel.
2. CMN esetén a λ értékét változtattuk

A referencia legjobb eredményei a személynevekre adódtak. Ebben az esetben 62.1%-os F-measure értéket kaptunk. A legrosszabbul pedig a szervezeteket címkézte fel a CRF, ahol az F-measure 2.7%-os lett. Az eltérés valószínűleg a tanítóhalmazban található pozitív példák száma miatt tapasztalható.

A részben felügyelt tanulás eredményei 100 iteráció után a következőképpen alakultak erre a két osztályra. Személynevek esetén a címkepropagálás szignifikánsan alulmaradt a CRF-fel szemben. A legjobb eredményt (F-measure = 19.5%) akkor értük el, amikor a tokenélek súlyait 0.5-re, a CMN normalizációs tényezőjét pedig 0.0-re állítottuk. Utóbbi azt jelenti, hogy a CMN normalizáció egyáltalán nem segített a személyneveknél. Bár a szervezetek esetében a gráfalapú módszerek legjobb eredménye csupán 4.1%-os F-measure-t eredményezett, a CRF-hez képest mégis javulást értünk el. Ebben az esetben a CMN segített az eredmény javításában, a λ értéke 0.05 volt; a tokenélek súlya pedig az előző esethez hasonlóan 0.5.

5 Konklúzió

Összességében azt mondhatjuk, hogy bár a gráf alapú módszereink kis adatbázisok esetén bizonyos esetekben jobban működnek, mint a szekvenciajelölők, a legtöbb esetben ez nem mondható el azok jól megfogalmazható matematikai háttérére. Ezért további kísérleteket folytatunk a CMN-nel történő normalizálásra és a tokenélek súlyának nem konstans értékű megválasztására. A jövőbeli terveink között szerepel továbbá a K értékének és a címkézett pontok száma hatásának vizsgálata.

Köszönetnyilvánítás

A kutatást – részben – a TEXTREND projekt (Jedlik Ányos program) keretében az NKTH támogatta.

Hivatkozások

1. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi Supervised Learning. 11. fejezet, The MIT Press (2006)
2. Farkas R.: Részben felügyelt tanulási módszerek a tulajdonnév felismerésben. In: V. Magyar Számítógépes Nyelvészeti Konferencia (2007) 166-176
3. Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields for Relational Learning. In: Getoor, L., Taskar, B. (eds.): Introduction to Statistical Relational Learning, The MIT Press (2007)
4. Zhu, X.: Semi-Supervised Learning Literature Survey. Technical Report Computer Sciences 1530, University of Wisconsin-Madison (2005)