Szövegbányászat Bayes-hálókkal és kiterjesztéseikkel

Antal Péter, Millinghoffer András

Kivonat A Bayes-hálók napjainkra a valószínűségi tudásábrázolás elsődleges eszközeivé váltak. Ennek oka, hogy számítási szempontból hatékonyan valamint emberi felhasználók számára jól értelmezhetően képesek valószínűségi változók együttes eloszlását ábrázolni. E jó tulajdonságaik ellenére a Bayes-hálók tanulása jelenleg is több problémát vet fel: a valós alkalmazási területeken tipikusnak mondható hiányos adatok mellet tanulásuk rendkívül számításigényes. E cikkben a Bayes-hálók alkalmazási lehetőségeit vizsgáljuk orvosi területen, bemutatva egy módszert, mellyel szakirodalmi cikkek szövegbányászata végezhető el, megtanulva a tárgyterület valós világbeli modelljét. Végül ismertetjük az eddig alkalmazott monolitikus Bayes-hálók modellezési korlátait és felvázoljuk azok néhány, a probléma orvoslására bevezetett kiterjesztését.

Alkalmazási terület

Cikkünkben a petefészekrák klinikai modellezését tárgyaljuk, részben kapcsolódó korábbi munkáink miatt [1, 2] és mivel e terület több szempontból is ideális alapját adja az ismertetendő módszereknek. A téma orvosi és társadalmi fontossága mellett (lásd [7]), a modellezési feladat informatikusi szempontból is alkalmas a vizsgálatra, ennek okai:

- A terület régóta az orvosi kutatások fókuszában áll, így jelentős tárgyterületi tudás halmozódott fel, mely mind referenciaként mind kiindulási pontként alkalmazható a tanulás során.
- Egy korábbi adatgyűjtés folytán nagy mennyiségű klinikai mérési adat áll rendelkezésünkre [9], így jól tesztelhetők a különféle tanulási eljárások.
- A terület szakirodalmi feldolgozása is alapos, így az ismertetendő szövegbányászaton alapuló algoritmusokhoz is bőséges adat áll rendelkezésre.

A mérések során a terület egy 16 legfontosabb változóját tartalmazó modelljét vizsgáltuk.

Bayes-hálók

Formálisan egy Bayes-háló egy struktúra-paraméter páros. Egy irányított, körmentes gráf írja le a modell struktúráját: a csomópontok felelnek meg a terület egy-egy változójának, az élek pedig értelmezhetők közvetlen valószínűségi (az októl az okozat felé irányuló) összefüggésként. Minden feltételezésünk szerint diszkrét csomóponthoz egy feltételes valószínűségi tábla (FVT) tartozik, mely a változónak a szülei szerinti feltételes eloszlását írja le.

Kauzális értelmezés

Bár számítási szempontból és az együttes eloszlás modellezését tekintve a gráfstruktúra az abból kiolvasható függetlenségek miatt fontos, az emberi interpretáció esetén az élek, mint közvetlen oksági összefüggések jelennek meg. Ez a szemlélet alapvető a szakértői modell konstruálásakor (amikor a hálót pusztán a szakértői tudás alapján építjük fel), de a gépi tanulás során is felhasználható, az éleket a tárgyterület jellemező jegyeiként szemlélve.

Bayes-hálók tanulási problémái

A tanulás alapvetően a bayesi a posteriori valószínűség alkalmazását jelenti, azaz a statisztikai, megfigyelési adatokat felhasználva, azokat előzetes ismereteinkkel kombinálva alakul ki a tárgyterületet modellek feletti eloszlása. Legegyszerűbb esetben a maximum a posteriori modellt keressük.

Általánosan alkalmazott módszer, hogy a számítások könnyítése érdekében bizonyos feltevéseket teszünk (paraméter függetlenség és modularitás, struktúra modularitás), amelyek

mellett az a posteriori valószínűség lokális jegyek alapján számítható, bár az optimalizálási probléma komplexitása NP-teljes [5].

Kauzalitás a szakirodalomban

Módszerünk alapötlete, hogy a modellt meghatározó struktúrát a fogalmak szakirodalmi cikkekben való együttes felbukkanása alapján tanuljuk. Ez az eljárás a Bayes-hálók kauzális értelmezésével indokolható, azaz az éleket közvetlen oksági összefüggésekként értelmezzük, és a cikkek alapján is ugyanilyen összefüggéseket várunk el. Abból a feltevésből indulunk ki, miszerint az irodalmi publikációkban azok a változók fognak együtt megjelenni, melyek a valóságban is szoros összefüggésben állnak (bővebben lásd [3]).

Szakirodalmi adatok származtatása

Az adatok előállítása során az elsőként a tárgyterület (petefészekrák klinikai diagnosztikájának) a fogalmai kerültek azonosításra [9]. Ezt követte a cikkek gyűjteményének meghatározása. Majd egy indikátorvektort rendelünk a cikkekhez, ami a cikkekben szereplő releváns fogalmak jelenlétét jelzi az adott írásban, felhasználva a fogalmakhoz-változókhoz rendelt rövid szöveges leírások és a cikkek statisztikai hasonlóságát. A módszer részletes leírását lásd [2]-ben.

Generatív modellek

Célunk tehát olyan Bayes-hálók építése, melyek a szakirodalmi adatok keletkezését modellezik. Erre a célra két modellosztályt vizsgáltunk, melyeket az optimális struktúra kiválasztásával illesztünk az irodalmi adatokhoz.

A számítások egyszerűsítése érdekében felhasználtuk, hogy egy korábbi kutatásból elérhető a változók egy okozati sorrendezése (mely meghatározza, hogy e csomópont szülője csak egy előtte álló csomópont lehet), valamint hogy a csomópontok szülői halmazának méretére egy felső korlát adható.

A tranzitív modell

E modellben a használt változók egy az egyben megfelelnek a valós modell változóinak, vagyis a formát tekintve ugyanazt a végső modellt használjuk az irodalomból való tanítás során is, majd az itt kapott struktúrát visszük tovább.

Az intranzitív modell

Az intranzitív osztályban, minden fogalomhoz két csomópont tartozik, egy, mely az adott változó említéséhez kapcsolódó szándékot jelképezi, és egy másik, mely a változó cikkben való megjelenését reprezentálja. E két csoport egy páros gráf két csomópontosztályát adja, az élek a „szándék-változóktól” a „megjelenés-változókhoz” futnak, melyek azt jelzik, hogy a szülőre vonatkozó említési szándék maga után vonhatja a gyermek megjelenését. A tanulás végeztével a valós modellhez illeszkedő struktúrát kapunk az egy-egy fogalomhoz tartozó két változó összevonásával.

Mérési eredmények

A tanulások eredménye a vizsgált struktúrák közül kiválasztott maximális a posteriori valószínűségű struktúra, ezeket egy a területet ismerő szakértő által adott referenciahálóval vetettük össze, a lehetséges csomópontpárok kauzális viszonya szerint. Az ilyen

modell szakértő	**független**	**megzavart**	**kauzális út**	**kauzális él**
független	14	14	12	12
megzavart	6	14	0	2
kauzális út	44	48	24	14
kauzális él	14	6	4	12

modell szakértő	**független**	**megzavart**	**kauzális út**	**kauzális él**
Független	44	0	0	8
Megzavart	14	8	0	0
kauzális út	82	18	20	10
kauzális él	8	4	2	22

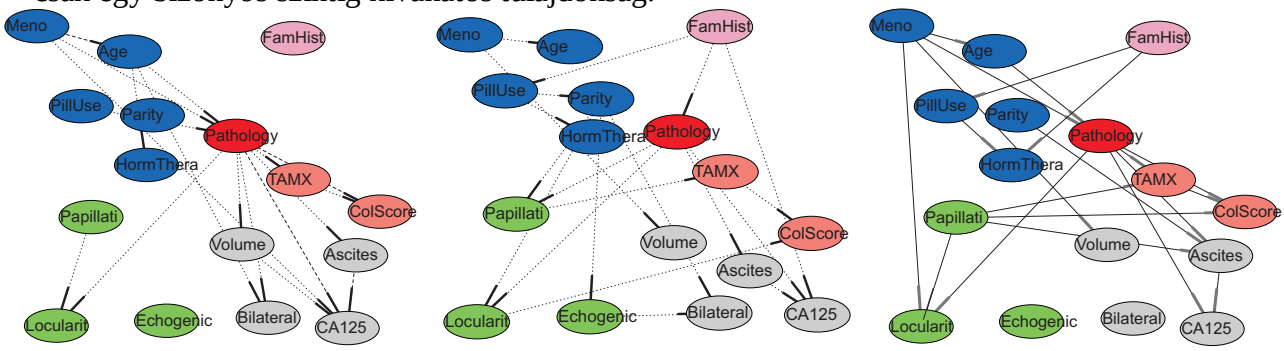
1. táblázat: csomópontpár-viszonyok mátrixai

viszonyok a következők lehetnek (gyengülő sorrendben):

- kauzális él – a két csomópont között él fut
- kauzális út – a csomópontok között irányított út van
- megzavart – a változók rendelkeznek közös őssel
- a fentiek egyike sem

Két struktúra hasonlóságát jól jellemzi ha egy mátrixban ábrázoljuk, hogy a referenciahálóhoz képest egy adott típusú kapcsolatból hány alakult át és milyenné (lásd 1. táblázat). Minél magasabbak egy ilyen mátrix átlóelemei a többihez képest, annál jobban hasonlít a két struktúra.

A mátrixok vizsgálatából kiderül, hogy mindkét tanulás jól közelíti a szakértői modellt, emellett az intranzitív modell valamivel pontosabb, bár ennek az ára a ritkább struktúra, ami csak egy bizonyos szintig kívánatos tulajdonság.



1. ábra: a szakértői, a tranzitív és az intranzitív háló

A fenti vizsgálatokkal kapcsolatban megjegyzendő, hogy elhanyagolja az irodalmi adatok szekvenciális tulajdonságát, azaz, hogy minden bejegyzéshez (cikkhez) egy adott dátum köthető. Eszerint, ahogy alakulnak a területről alkotott elképzelések és változnak a publikációs szokások, azaz változnak a fogalmak együttes megjelenésének valószínűségei, úgy változik az időben a maximum a posteriori irodalmi modell is. Egyetlen modell keresése szintén túlzott egyszerűsítés, hisz a bayesi technikák erejét többek közt az adja, hogy az utólagos következtetéseket nem csupán egyetlen modell alapján, hanem az összes modell feletti a posteriori eloszlás szerint hozhatjuk meg. További előrelépést jelent, ha csak releváns strukturális jegyek a posteriori valószínűségét elemezzük. Ezek kutatása jelenleg is folyik [4], a továbbiakban pedig a strukturális jegyek elemzésének általánosabb, a következő fejezetben ismertetett keretrendszerekbe foglalt módjait kívánjuk vizsgálni.

A monolitikus Bayes-hálók korlátai

Bár a fenti mérések eredményei biztatóak, és igazolják, hogy hasonló területen eredményesen alkalmazhatók az irodalomra támaszkodó heurisztikus módszerek, már számos alkalmazási területen bebizonyosodott, hogy a modelltér méretének növekedésével (különösen hiányos adatok esetén) hamar elérjük a rendelkezésre álló számítási kapacitás korlátait. A probléma gyökerei a hagyományos (ún. monolitikus) Bayes-hálók modellezési képességeiben keresendők. Itt a modell szempontjából minden változó teljesen függetlenül viselkedik a többitől, azaz az esetleg hasonló szerkezetű, viselkedésű struktúrák (amelyek pl. az orvosi területeken igen gyakoriak) kialakulása csak ad hoc módon kényszeríthető egy tanulás során, vagyis a részstruktúrák nem. A helyzetet illusztrálhatjuk a propozicionális és az elsőrendű logikák viszonyával, ahol az elsőrendű predikátumok kifejező erejét nulladrendű kijelentések halmazával tudjuk helyettesíteni: a monolitikus hálókra igaz hasonló tulajdonságok mindegyikét külön-külön rögzíteni kell.

Az OOBN és a PRM keretrendszerek

Az objektumorientált Bayes-hálók (object-oriented Bayesian networks – OOBN, lásd [8]) és valószínűségi relációs modellek (probabilistic relational models – PRM, lásd [6]) a fenti problémát próbálják megoldani, a Bayes-hálók kiterjesztéseivel.

Az OOBN-ek az ismert programozási paradigma mintájára objektumokat vezet be, melyek között a különféle relációk (öröklődés, rész-egész) biztosítanak egyfajta újrafelhasználhatóságot. Az objektumok önálló Bayes-háló részletek, melyek csak meghatározott (ún. bemeneti és kimeneti) változókon keresztül kapcsolódhatnak egymáshoz ezáltal elszigetelve egymástól a koncepcionálisan nem összefüggő változókat, ez pedig a következtetési algoritmus gyorsítását eredményezi, ami kardinális a tanulás sebességét illetően is.

A PRM modellek a változókat relációs adatbázisok változóinak feleltetik meg, a valószínűségi modell pedig a táblák sorain belüli összefüggéseket írják le, egy-egy sor pedig a vonatkozó változóhalmaz egy behelyettesítésének felel meg. E modellosztály alapvető újítása az, hogy a többi tábla kulcsaira vonatkozó valószínűségek által a strukturális viszonyokat is valószínűségi alapokra helyezhetjük.

Összegzés

E cikkben megvizsgáltuk a szakirodalmi publikáció két generatív modelljét, melyek alkalmazhatók a modellezendő terület strukturális viszonyainak tanulására. A szakértői referenciahálóval való összevetés megmutatta, hogy az irodalomból kinyerhető ismeretek összemérhetőek a szakértők által adott modellek által reprezentálttal.

A fenti sikerek ellenére az efféle heurisztikák nem adhatnak általános megoldást a modellezés korlátainak problémájára, ezért a további kutatási iránynak az OOBN és a PRM keretrendszerek felé kell mutatnia, vagyis a struktúratanulást ezekben a rendszerekben azok erősségeihez illeszkedve kell végrehajtani.

Irodalomjegyzék:

[1] P. Antal, G. Fannes, Y. Moreau, B. De Moor: Bayesian Applications of Belief Networks and Multilayer Perceptrons for Ovarian Tumor Classification with Rejection, Artificial Intelligence in Medicine, vol 29, pp 39-60, 2003

[2] P. Antal, G. Fannes, Y. Moreau, D. Timmerman, B. De Moor: Using Literature and Data to Learn Bayesian Networks as Clinical Models of Ovarian Tumors, Artificial Intelligence in Medicine, vol 30, pp 257-281, 2004

[3] P. Antal, A. Millinghoffer: Learning Causal Bayesian Networks from Literature Data, Proceedings of the 3rd Int. Conf. on Global Research and Education, Inter-Academia'04, 2004.

[4] P. Antal: Bayesian network based analysis of data, literature and expertise for classification of ovarian tumor, doktori disszertáció, előkészületben

[5] D. M. Chickering: A Transformational Characterization of Equivalent Bayesian Network Structures, Proceedings of 11th Conference on Uncertainty in Artificial Intelligence (UAI-1995), pp 87-98, 1995

[6] L. Getoor, N. Friedman, D. Koller, B. Taskar: Learning Probabilistic Models of Relational Structure, Proceedings of 18th International Conf. on Machine Learning, pp 170-177, 2001

[7] R. Harris, A. S. Whittemore, J. Itnyre and the Collaborative Ovarian Cancer Group: Characteristics Relating to Ovarian Cancer Risk (I, II, III, IV), American Journal of Epidemiology, vol 136, pp 1175-1220, 1992

[8] D. Koller, A. Pfeffer: Object-Oriented Bayesian Networks, Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI-97), pp 302-313, 1997

[9] D. Timmerman, L. Valentin, T. H. Bourne, W. P. Collins, H. Verrelst, I. Vergote: Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group, Ultrasound Obstetrics Gynecology, vol 16, pp 500-505, 2000