Opponensi vélemény

Bolla Marianna
Clustering Graphs and Contingency Tables with Spectral Methods
MTA doktori értekezéséről

Az értekezés angol nyelven íródott, 114 oldal terjedelmű, Bevezetést, három fejezetet, valamint Irodalomjegyzéket tartalmaz. A hozzá tartozó magyar nyelvű tézisek (Gráfok és kontingenciatáblák klaszterezése spektrális módszerekkel) 23 oldalas. A jelölt a kandidátusi értekezése után publikált 1 saját könyvet és 23 saját cikket sorol fel, ez utóbbiak közül 10 társszerzőkkel közös. A saját műveken túl 102 egységet tartalmaz az irodalomjegyzék.

Az első fejezet a többszempontú vágásokat és a spektrummal kapcsolatos reprezentációkat tartalmazza. Ennek első alfejezete a gráfokra koncentrál. Élsúlyozott összefüggő gráfok esetén a Laplace-mátrix segítségével megadja a k-dimenziós reprezentációt. Erre támaszkodva alsó becslést ad a gráf súlyozott k-vágásaira. Ezután él- és csúcs-súlyozott összefüggő gráfokat tárgyal. Abban az esetben, amikor a csúcs-súlyok az általánosított fokszámok, megadja a k-dimenziós reprezentációra és a k-vágásokra vonatkozó fenti tételek megfelelőit. Ezekben a tételekben a normált Laplace-mátrix sajátértékei játszanak szerepet. Az izoperimetrikus számra (Cheeger-konstansra) vonatkozó Cheeger-egyenlőtlenséget is megjavítja. A Laplace-mátrixon kívül használja a (Newman-Girvan-féle) modularitási mátrixot is. Az összefüggő súlyozott gráf normált modularitási mátrixának sajátértékei és a független csúcshalmazok mérete, valamint a gráf többrészes (multipartite) volta közötti kapcsolatokat leírja. Definiálja a Newman-Girvan-féle modularitást, amellyel a gráf csúcsainak klaszterein belüli kapcsolatoknak a véletlenszerűnél erősebb (community structure), illetve gyengébb (anticommunity structure) volta jellemezhető.

A második alfejezetben kontingenciatáblákat vizsgál. A korrespondencia elemzés feladatával, azaz a két változó egyidejű osztályozásával foglalkozik, eszközként szintén a korrespondencia elemzésben megszokott szinguláris felbontást használja. A gráfokra korábban igazolt tételeinek megfelelő eredményeit közli. Nevezetesen egy reprezentációs tételt és egy tételt az osztályzást jellemző mérőszám alsó korlátjáról.

A harmadik alfejezetben tovább általánosítja a tárgyalást. Nevezetesen, együttes eloszlások reprezentációit vizsgálja. A feltételes várható érték operátor segítségével adja meg a minimális költségű k-dimenziós reprezentációt. Szintén ebben az alfejezetben tárgyalja a reprodukáló magú Hilbert-tereket. Valójában azt – az alkalmazásokban fontos – dolgot magyarázza el, hogy sok esetben a feladat egy magasabb dimenziós térbe (feature space) történő transzformálással lineárissá tehető. Sőt a transzformáció helyett magfüggvényekkel érdemes dolgozni. Ezt a módszert széles körben használják a neurális hálózatok esetén, különösen a Vapnik-féle SVM-nél.

A második fejezet „Véletlenség kezelése nagyméretű hálózatokban és klaszterezés kis diszkrepanciával". Ennek első alfejezete szimmetrikus mátrixokkal foglalkozik. A kiindulás egy P szimmetrikus valószínűségi mátrix ún. felfújtja: B_n. Először belátja, hogy B_n sajátértékei n nagyságrendűek. Ezután egy W_n Wigner-zajjal perturbálja B_n-et: $A_n = B_n + W_n$. A W_n spektrálnormájára vonatkozó Füredi-Komlós-tétel és a Weyl-féle perturbációs tétel segítségével A_n és B_n sajátértékeinek az eltérését becsüli meg. Ezután az A_n élsúlyokkal rendelkező gráfot vizsgálja. Az A_n ún. strukturális sajátértékeihez tartozó sajátvektorokkal felírt reprezentáció esetén becsüli a gráf k-varianciáját. A gráf normált Laplace-mátrixának a sajátértékei elhelyezkedését is leírja. Az A_n mátrixra kirótt alkalmas feltételek esetén explicit konstrukciót is ad az A_n-et megközelítő B_n felfújt mátrix előállítására.

A második fejezet második alfejezetében zajos kontingencia-táblákat vizsgál. A zaj nélküli kontingencia-táblázat egy P (nem szimmetrikus) valószínűségi mátrix felfújtja: $B_{m\times n}$. A zajos kontingencia-tábla ennek egy Wigner-zajjal való perturbáltja: $A_{m\times n} = B_{m\times n} + W_{m\times n}$. $A_{m\times n}$ és $B_{m\times n}$ szinguláris értékeinek az eltérését becsüli meg. A szinguláris vektorok eltérését is vizsgálja. $A_{m\times n}$ normalizáltjának a szinguláris értékeinek elhelyezkedését is leírja.

A harmadik alfejezetben a diszkrepancia és a spektrum kapcsolatát vizsgálja. Egy kontingencia-tábla esetén a sorok és az oszlopok klaszterei közötti kapcsolatok „homogén" voltának mérésére a k-részes diszkrepanciát vezeti be. Ennek segítségével felső becslést ad a normált kontingencia-táblázat k-adik szinguláris értékére. Megfordítva, a szinguláris értékek segítségével felső becslést ad k-részes diszkrepanciára. Mindkét fenti eredmény megfelelőjét belátja irányított élsúlyozott gráfokra is.

A harmadik fejezet első alfejezetében gráfparaméterek tesztelhetőségét vizsgálja. Ez lényegében annyit jelent, hogy amikor a gráf méreténél fogva nem megfigyelhető, akkor annak egy viszonylag kicsi részéből tudunk-e következtetni az egész valamelyik paraméterére. Ennek érdekében belátja, hogy amennyiben élsúlyozott gráfok egy sorozata konvergens, akkor normált modularitás mátrixának sajátértékei és sajátalterei konvergálnak a limesz grafon megfelelő objektumaikoz. A harmadik fejezet második alfejezetében a Lovász-Sós-féle általánosított kvázirandom gráf sorozatokat vizsgál. Egy sejtést állít fel arra vonatkozóan, hogy a kvázirandom tulajdonság milyen más tulajdonságokkal (pl. a szomszédsági mátrix, illetve a normált modularitás-mátrix sajátértékeinek és sajátvektorainak tulajdonságaival) ekvivalens. Ezt a kérdéskört a disszertáció beadása után, a Bolla: „Generalized quasirandom properties of expanding graph sequences", arXiv: 1508.04369v6, 2017, kéziratában tisztázza: az említett tulajdonságok pontosítása után belátta azok ekvivalenciáját. A harmadik fejezet harmadik alfejezetében homogén, illetve inhomogén blokkmodellben javasol EM-algoritmust a paraméterek becslésére.

A doktori disszertáció Bolla Mariann munkásságának jelentős részét, lényegében a kandidátusi disszertáció után született eredmények többségét mutatja be. Kiemelendő, hogy egymással szoros összefüggésben lévő, egymásra épülő részekből álló, egységes műről van szó. A szerző részletesen bemutatja a szakirodalmi előzményeket, a megoldandó problémák fontosságát, az alkalmazott módszereket, továbbá elmagyarázza az eredményeket is. Ezek az ismertetések nem egyszerűen „összekötő szövegek”, hanem élvezetes szakmai útmutatók. A mű egységes jelölésrendszert és terminológiát használ.

A feldolgozott téma: gráfok klaszterezése (és az ennek általánosításaként tekinthető kontingencia táblák klaszterezése) aktuális, jelentős, a kutatók figyelmének középpontjában álló terület. A nagyméretű gráfok vizsgálata az utóbbi évtizedekben általában a figyelem középpontjában van a hálózatok (gyakorlati) jelentőségének felismerése miatt. A kontingencia táblák elemzése pedig genetikai vizsgálatokhoz kapcsolódik. Bolla Marianna egyik fő törekvése, hogy gyakorlati problémák megoldásának elméleti hátterét felderítse. Vizsgálatai kapcsolódnak élvonalbeli kutatók (pl. N. Alon, M.E.J. Newman, Lovász László, T. Sós Vera,...) eredményeihez.

A szerző a matematika különböző területeinek eszközeit széles körűen alkalmazza a bizonyításokban: rutinszerűen használja a lineáris algebra klasszikus módszereit, a valószínűségszámítás és a statisztika számos tételét is alkalmazza, továbbá járatos a gráfelmélet újabb eredményeinek használatában is.

A disszertáció hatalmas anyagot ölel fel, a főbb tételeket bizonyítja is, a többinél a szerző eredeti publikációira utal. A bizonyításoknak egy részét részletesen ellenőriztem, azokban hibát nem találtam. A műben csak néhány elírást találtam, pl.:

20. oldal, 25. sor: gyökjelek hiányoznak;

60. oldal, alulról 2. sor: talán a „*normalized* noisy contingency table” szinguláris értékeiről van szó.

A doktori munka tudományos eredményeit elegendőnek tartom az MTA doktori cím megszerzéséhez, a nyilvános védés kitűzését javaslom. Javaslom az MTA doktora cím odaítélését.

Debrecen, 2017. december 10.

Dr. Fazekas István
MTA doktora