Opponensi vélemény

Abonyi János

Technológiai fejlesztési célú adatbányászati technikák

c. MTA doktori értekezéséről

1. Témaválasztás, módszertani kérdések

Az értekezés a termék- és technológia fejlesztés hatékonyságának növelésének motivációjával indokolja az értekezés alapját képező kutatások témaválasztását. Megállapíthatjuk, hogy a viszonylag szerény cím mögött valójában igen széles körben alkalmazható és alkalmazott, nagyon fontos modell identifikációs kulcstechnológiák terén elért jelentős eredmények rejlenek. A termék- és technológiafejlesztés mellett az irányítás- és döntéselmélet, illetve ezek alkalmazásai szinte minden területén komoly jelentősége van azoknak az utóbbi évtizedekben egyre jobban a tudományos közvélemény érdeklődésének középpontjába kerülő kutatási eredményeknek, amelyek sorába Abonyi János tézisei, illetve az ezek mögött rejlő, a gyakorlatban is sok esetben sikeresen alkalmazott, új módszertani eredmények illeszkednek. A vizsgált technológiák és ipari folyamatok (példánkban vegyipari és építőipari területek szerepelnek) tipikusan sok állapotváltozótól függnek, analitikus leírásuk nem teszi lehetővé a kezelhető számítási bonyolultságú pontos modellezést, illetve, gyakran tartalmaznak olyan elemeket, amelyek a modellező mérnök és matematikus szemszögéből nemdeterminisztikus viselkedésként jelennek meg. A bonyolultság, a zajjal terheltség, a nemdeterminisztikus elemek külön-külön is indokolják a hagyományos modellezési technikákkal szemben a számítási intelligencia eszköztárához való fordulást. A Jelölt eredményei nagymértékben támaszkodnak a fuzzy rendszerek elméletére és módszertanára, a csoportosítási algoritmusokra (klaszterezés), és kisebb mértékben az evolúciós és azokkal rokon eljárásokat is bevonja vizsgálódásai körébe. Megjegyezzük, hogy ez utóbbiak körében a már klasszikusnak számító genetikus algoritmusok korszerűbb és hatékonyabb változatai kívül esnek a Jelölt vizsgálódási körén, így nem hivatkozik a genetikusnál bizonyítottan jobb konvergenciájú bakteriális, a több lokális optimummal rendelkező problémák esetében a viszonylag lassú globális optimumot nyújtó evolúciós algoritmusokkal szemben az evolúciós-memetikus eljárásokra, melyek a különösen nagy állapotváltozó szám esetén lassan konvergáló evolúciós technikákkal szemben kombinált evolúciós és gyors lokális keresési eljárások egymásba ágyazott együttesén alapulnak. Érinti a rajintelligencián alapuló megközelítéseket is, de csak kisebb intenzitással.

A témaválasztás igen időszerű, a vizsgált problémakör alkalmazhatósága nagyon széles, a fent jelzett kritikai megállapításoktól függetlenül a Jelölt által elvégzett kutatás volumene,

az elért eredmények sokasága és jelentősége (melyet az intenzív nemzetközi idézettség és a Jelölt eredményeire épülő további kutatások sokasága bizonyít) impozáns. A tézisek eredményeit közlő publikációk mennyisége és minősége, a független hivatkozások száma, a szélesebb szakmai-kutatási területen belül kiemelkedő munkásságot indikál.

2. Az eredmények összefoglalása, néhány kritikai megjegyzés

A következőkben összefoglalom az értekezés fő eredményeit, a tézisek sorrendjét és logikáját követve, az egyes téziscsoportokkal kapcsolatban néhány általános kritikai megjegyzést is téve:

1. E téziscsoport lényege az 1.3 alatt ismertetett idősorszegmentálási algoritmus. Ez az eljárás hagyományos statisztikai módszereket alkalmaz, a hiba-kovariancia mátrix alapján történnek az állapotbecslések, melyek a Jelölt által javasolt dinamikus rendszermodellt támogatják. Ez a módszer bizonyos mértékig heurisztikus, illetve szakértő általi közreműködést is igényel. A bemutatott alkalmazási példa, mely feltehetően e modell kialakításának is motiválója volt, a polietilén előállítására alkalmas nagyüzemi technológia. Az 1.1 altézis megfogalmazása túlságosan „szoft” ez mintegy a következő altézisek bevezetése értelmezhető. Ez az altézis lényegesen új tudományos gondolatot nem tartalmaz.
Az 1.2 altézis beépíti az intelligens számítási elemeket, a neurális hálózatos és a fuzzy modelleket. Ennek az eltézisnek a tudományos újdonsága, egy klaszterezésen alapuló Takagi-Sugeno fuzzy részmodelleket tartalmazó, a Jelölt által szemi-mechanisztikusnak nevezett modell, melynek paramétereit spline-simításon alapuló technikával lehet becsülni. Az 1.3 altézis a Kálmán-szűrő alkalmazása és az idősor-szegmentáció új algoritmusa, amely sokváltozós esetben is alkalmazható.

Ezen téziscsoport tudományos értékét elsősorban az 1.3 altézis új algoritmusa jelenti. Ezzel természetesen nem kívánom csökkenteni a módszertani javaslatok jelentőségét, melyek az első altézisekben kerültek megfogalmazásra.

2. Ebben a téziscsoportban alkalmas metrika definiálásával a Jelölt a Gath-Geva klaszterezési eljárás egy célszerű módosított változatával olyan identifikációs módszert javasol, amely közvetlenül alkalmas Takagi-Sugeno (-Kang) típusú fuzzy modellek identifikálására. A 2.1 altézisben szereplő eredmény a lokális lineáris részmodellek meghatározását a legkisebb négyzetek módszere felhasználásával végzi.

Megjegyzem, hogy széleskörű elterjedtsége ellenére a Takagi-Sugeno típusú modellek interpolációs viselkedése az antecedensek átfedési intervallumaiban nem kedvező, a lineáris

szakaszok közötti átvezetések többszörös inflexiós pontokat tartalmaznak. Bár kimutatható, hogy ez a modelltípus is elméletileg univerzális tulajdonságú, felvetődik annak a kérdése, hogy szerencsésebb volna a CRI-Mamdani modellcsalád irányában keresni a gyakorlati alkalmazású modellek identifikációjának módszereit. (A két modelltípus csak asszimptótikusan ekvivalens.) Ez a megjegyzés természetesen nem csökkenti a Jelölt kutatási eredményeinek értékét.

Igen fontos az, hogy a 2-3 altézis szerint a javasolt új módszer képes a főkomponensek és szegmensek számának automatikus meghatározására. Osztályozásnál, szegmentálásnál az egyik igen kritikus kérdés a komponensek számának meghatározása, amely téves érték esetén az eredményeket akár teljesen értékelhetetlenné is teheti. A teljes második téziscsoport eredményeinek érvényessége és gyakorlati alkalmazhatósága is függ a klaszteregyesítési eljárás jóságától. Az alkalmazott Kelly-féle módszer erre kétségkívül lehetőséget kínál, bár az irodalom alternatívákat is felsorakoztat. E megközelítés filozófiai háttere statisztikai jellegű, mind oly sok fuzzy modellt alkalmazó adatfeldolgozási eljárásnál. Az a kérdés természetesen mindig feltehető, vajon a lehetőségi mérték használata a valószínűségi helyett milyen irányban változtatná meg a kapott eredményeket. (A lényegi különbség a valószínűség-elmélet additivitási axiómája, melyet a lehetőség-elmélet az egymásba skatulyázott elemi események miatt a szub-additivitás irányában változtat meg.) A legnagyobb sajátértékekből nyerhető főkomponens szám alátámasztja a Jelölt által javasolt módszer helyességét. Természetesen a klaszteranalízisben elterjedten használt kovariancia mátrixokra épülő szeparabilitás eleve a valószínűségi mérték talaján történő megközelítést feltételez.

Bár a bíráló egyes vizsgálatai is alátámasztják a Gath-Geva klaszterezés hatékonyságát, mégis szerencsésnek tartottam volna, ha a Jelölt hivatkozott, illetve összehasonlító vizsgálatokat végzett volna az irodalomban igen elterjedt és különösen korszerű továbbfejlesztett változataiban meglehetősen hatékonynak mondható fuzzy c-means (FCM) klaszterezési eljárásra, illetve ennek alternatív alkalmazásával kapcsolatban. Bár közismert, hogy ez az eljáráscsoport csak viszonylag homogén méretű és alakú klaszterek esetén eredményez jó felosztást, továbbá csak az osztályok számának előzetes ismerete mellett ad korrekt eredményeket, általában véve ez az eljáráscsalád gyors és jó konvergenciája, az eredményként létrehozott tagsági függvények könnyű kezelhetősége miatt kétségkívül a legelterjedtebb ilyen jellegű eljárás.

3. A harmadik téziscsoport az a priori információk (stabilitás, erősítés, időállandó, stb.) fuzzy modellek paramétereire vonatkozó korlátok formájában történő megfogalmazásával kapcsolatos. A két altézis megállapításai filozófiai értelemben nem

meglepőek – pontosabb a modell és predikciós képessége, ha több rendelkezésre álló információ kerül beépítésre. A Jelölt módosított fuzzy modellre tesz javaslatot. A 4.2 identifikációs példa tapasztalatai nem támasztják alá az a priori információ jelentőségét a modellhatékonyság szempontjából. Tény azonban, hogy különböző a priori korlátok alkalmazása más és más modellparamétereket eredményez.

Mivel a mai napig nem született meg a fuzzy szabálybázisos modellek egzakt funkcionálkalkulusa, nehéz eldönteni, vajon egy bizonyos modellosztályon belül létezik-e egyáltalán optimális fuzzy modell. Általában megállapítható azonban, hogy fuzzy szabálybázisos modellidentifikáció esetében a legkritikusabb kérdés az identifikáció alapjául szolgáló adatok eloszlása, kiterjedése. Remélhető, hogy a jövőben olyan tételek kerülnek kimondásra, amelyek a szabálybázisos modellek által reprezentált explicit függvények tulajdonságai és a felhasznált adategyüttesek jellegzetességei között állapítanak meg szükségszerű összefüggéseket.

A fuzzy modellekkel kapcsolatos fenti megállapítások értelemszerűen vonatkoznak a harmadfokú spline-ok illesztésével kapcsolatos 3.2 altézis megállapításaira is.

4. A negyedik téziscsoport két altézisének összekötő eleme az alkalmazott genetikus algoritmikus megközelítés. A 4.1 altézis a modellek egyszerűsítésére vonatkozó eljárásra vonatkozik, ahol – a korábbi téziscsoportokhoz hasonlóan – kulcsszerepet játszik a legkisebb négyzetének módszere. Az alkalmazott genetikus programozási elem kétségkívül hasznos lehet, ám megjegyzem, hogy a szakirodalomban már régóta elterjedtek a genetikusnál jobb konvergenciatulajdonságú evolúciós és az evolúciót más, hagyományosabb optimalizálási eljárásokkal kombináló módszerek (bakteriális algoritmus illetve programozás, memetikus módszerek).

A 4.2 altézis a részecskeraj-alapú optimalizációra is épít. Ez utóbbival kapcsolatban az elmúlt években nagyszámú új eredmény született. A módszer hatékonyságát sok szerző támasztotta alá. Ez az altézis interaktív lehetőséget vet fel, mely bizonyos értelemben ismét a heurisztika irányába viszi a Jelölt érdeklődését.

Bár ezen altézisek értéke nem vitatható, meg kell állapítsam, hogy a Jelölt az evolúciós és rokon algoritmusok területén meglehetősen esetlegesen válogatott a megvizsgált módszerek között és így az eredmények kevéssé tekinthetők általános érvényűnek és jelenlegi ismereteink szerint optimálisnak.

3. Értékelés, megjegyzések

Az eredmények összefoglalása során minden egyes téziscsoport kapcsán megfogalmaztam néhány olyan általános kérdést vagy megjegyzést, amelyek az egyes téziscsoportok kapcsán felvetik további mélyebb, az itt kapott eredményeket az adott részterület szakirodalmának kontextusában pontosabban elhelyező további vizsgálatok lehetőségét. Különösen az igen konzisztens 2. téziscsoport mutat be olyan kutatási eredményeket, amelyek akár egy vagy több teljes értekezés volumenével összemérhető további vizsgálatok lehetőségeit vetik fel. A 4. téziscsoport pedig csupán érint egy, az előbbinél még nagyobb volumenű kérdéskört. Nem vitatva a teljes értekezés egészén végighúzódó vezérfonalat, az egyes téziscsoport között meglévő bizonyos mértékű összefüggést, mégis úgy látom, az alkalmazott eszköztár igen széles és az alkalmazás mélysége nem homogén. Az értekezés ilyen szerkezetét elsősorban a téziscsoportok mindegyikének a modellidentifikációhoz való kapcsolódása, de még inkább a Jelölt igen nagyszámú és jelentős részeredményt tartalmazó igen komoly fórumokon megjelent közleményei és azoknak a szakmai közvélemény által történt pozitív fogadtatása, az élénk hivatkozások indokolták. Az értekezés volumene, az abban megfogalmazott eredmények kétségkívül elérik, sőt meghaladják, egy MTA Doktora értekezés szintjét. A téziscsoportokat egy sorba fűző gondolati lánc azonban lazának tűnik és az egyes érintett szűkebb kutatási irányok mindegyikében sok nyitott kérdés marad fent. Módszertani szempontból összeköti a „hard” téziseket a statisztikai modellek, mindenekelőtt a hiba-kovarianciamátrixok, valamint a TS-típusú fuzzy és bizonyos neurális modellek ismételt alkalmazása. A klaszterezés és az evolúciós eljárások sok tekintetben egymás alternatívájaként merülhetnek fel, de egyetlen esetben sem történt meg ezen alternatívák összehasonlító értékelése, nem világos, miért az egyiket vagy másikat alkalmazza a Jelölt egyik-másik eljárásában, illetve miért nem vizsgálja meg a másik alternatívát, míg egy másik részkutatásban ezt éppen fordítva teszi.

A következőkben néhány apróbb kritikai észrevételt teszek:

- 4. 5. oldal (1.2 3.a) A radiálbázisos neurális modellek valóban igen szorosan összefüggenek a fuzzy szabálybázisos rendszerekkel. Ez utóbbiak azonban formálisan általánosabbak, így óvatosabban alkalmaznám az ekvivalencia kifejezését.
- 9. oldal – A 2.1 ábrán bemutatott és a következő bekezdésben ismertetett modell egy részben vitatható értelmezést feltételez, amely a tudásalapú viselkedést élesen elkülöníti a szabályalapútól. A „szabály” kifejezés természetesen használható egy erősen szűkítő értelemben is, ám az értekezésben újra meg újra

kulcsszerepet játszó fuzzy szabályok éppen egy tudásbázis alkotóelemeiként jelennek meg, ezért a szabályok kizárása a tudás-poolból nem szerencsés.

- 18. oldal – A 3.1 első bekezdésében szereplő állítása a Takagi-Sugeno modellt, mint a legnagyobb figyelmet kapott fuzzy modelltípust állítja be. Ez tévedés, mivel a Mamdani-modellcsalád az előbbinél kétségkívül elterjedtebb. Ugyanakkor e két modell asszimptotikusan azonos viselkedést mutat, mely kérdést a Jelölt nem veti fel. A 19. oldalon hivatkozott eredmények (31-36) tipikusan 10-20 évesek. Ezek a publikációk (leszámítva a klasszikus, 1993-as [31] cikket) egy kutatási vonulatot dokumentálnak, ám sok alternatív és esetenként akár eredményesebb vizsgálatra az értekezésben nem történik hivatkozás, illetve ezeket a vonulatokat a Jelölt nem vizsgálja.
- 20. oldal – Meglepő, hogy a Gustavson-Kessel és a Gath-Geva klaszterezési algoritmusok mellet még csak nem is említi a jelölt a Bezdek által kidolgozott és sokszorosan továbbfejlesztett FCM-eljárást.
- 33. oldal – Érdekes tény, hogy a 3.2 táblázat teszteredményei az FMID kivételével szinten teljesen azonosak. Mennyiben támasztja ez alá tehát egyik vagy másik módszer előnyét?
- Meglepő, hogy a 68-71. oldalakon szereplő példában a 4.2 táblázatban látható identifikált modellparaméterek nagymértékben különböznek egymástól. Ennek hatása természetesen a 4.5 ábrán látható lépcsős válaszfüggvények különbözőségében is megnyilvánul. Bár a Jelölt a 70. oldalon magyarázatot fűz a paraméterek eltérő voltához, ezt az identifikációs technikát nem érzem teljesen megnyugtatónak és nem érzékelem a matematikai értelemben vett stabilitás avagy robusztus viselkedés garantált voltát. Tud-e a Jelölt erre nézve valamilyen analitikus indoklást adni?
- 87. oldal – Nem vitatva a polinomiális modellek elterjedtségét, felvetem, hogy a polinomiális függvények igen rossz stabilitási tulajdonságai miatt nem volna-e érdemesebb megvizsgálni kedvezőbb tulajdonságú, robosztusabb függvényosztályokat is (mint például a racionális törtfüggvények osztályát).
- 85-93. oldal – Mint az általános megjegyzések között is említettem, az evolúciós (és részecskeraj) algoritmusok alkalmazásával foglalkozó fejezet túlságosan esetleges és nem tartalmaz rendszeres összehasonlító vizsgálatokat. Az itt szereplő állítás, mely a genetikus algoritmusok korszerűségét és növekvő népszerűségét illeti erősen vitatható, és az irodalomjegyzék semmiképpen sem támasztja alá ezt a kijelentést.

- Meglepő az, hogy a 2010-ben elkészített értekezés 157 hivatkozása között 2002-nél frissebb dátumú alig szerepel (1 db 2004-es, 2 db 2003-as, 1 db 2005-ös „megjelenés alatt” jelölésű, és 2 db 2006-os részben nem publikált tanulmány – az önhivatkozásokat leszámítva). A hivatkozások döntő többsége azonban az 1980-as évek végéről és az 1990-es évekből való. Miért nincsenek friss irodalmi hivatkozások, vajon ismeri-e a szerző az újabb eredményeket?

Az értekezés külalakja mintaszerű, az angolság jó. Az érthetőséget csak néhány esetben zavarja nyelvtani hiba, mint például a 49. oldal első sorában „neither… does not…”.

4. Összefoglaló

Az értekezés kétségkívül eléri, sőt meghaladja egy átlagos MTA doktori értekezés volumenét és színvonalát.

Az egy téziscsoporton belül különösen az 1.3 altézis tekinthető jól megfogalmazható új tudományos eredménynek. A 2. téziscsoport a tudományos eredmények egy csokrát tartalmazza, melyek kétségkívül az értekezés legnagyobb értékét képviselik. A 3. téziscsoportot csak fenntartással tudom új tudományos eredményként kezelni, hiszen az ott megfogalmazott állítások kevéssé újszerűek, bár a Jelölt konkrét vizsgálatokkal, futtatásokkal támasztotta alá állításait. Ezért ez inkább alkalmazási jellegű illusztrációs háttér az egész értekezés témakörének bemutatására. A 4. téziscsoport igen érdekes kérdéseket vet fel és itt részeredményeket is tartalmaz, bár a fentiekben megfogalmazott fenntartásaimat kénytelen vagyok újra felvetni.

Összefoglalva a Jelölt tézisértékű új kutatási eredményeket ért el, ezeket elegáns módon publikálta és rájuk számos érdemi független hivatkozást is kapott. A fentiek alapján – kérdéseim és kritikai megjegyzéseim érdemi megválaszolása esetén – az MTA Doktora cím odaítélését támogatom.

Kóczy T. László

MTA Doktora