Doktori (PhD) értekezés tézisei

HASONLÓSÁG ALAPÚ ÉLETLEN HALMAZOK ÉS ALKALMAZÁSUK AZ ADATBÁNYÁSZATBAN

Nagy Dávid

Témavezető:
Dr. Mihálydeák Tamás

DEBRECENI EGYETEM
Informatikai Tudományok Doktori Iskola

Debrecen, 2020

Tartalomjegyzék

1. Bevezetés és motiváció

Napjainkban az adatok mennyisége robbanásszerűen növekszik. A keletkező adatok azonban gyakran hiányosak esetleg inkonzisztensek. A hiányosságnak számos oka lehet. Például nem ismert az adott érték vagy éppen nem értelmezhető. Inkonzisztenciáról akkor beszélünk, ha az adatok közt valamilyen ellentmondás mutatkozik. Ezen problémák számos nemkívánatos eseményt idézhetnek elő (rossz előrejelzés, nem megfelelő döntéshozatal stb.). Az informatikai tudományokban az ilyen jellegű pontatlanságok elkerülésére számos módszert alkottak.

Az életlen halmazok elmélete egy viszonylag új elmélet, melynek alapjait Pawlak professzor a 80-as években fektette le [3, 4, 5]. A rendelkezésre álló információk alapján számos esetben előfordul, hogy két objektumot nem tudunk megkülönböztetni egymástól. Két objektum megkülönböztethetetlennek tekinthető, ha minden ismert releváns tulajdonságuk megegyezik. A pawlaki terek a rendelkezésre álló adatok alapján fellépő megkülönböztethetetlenségből származó bizonytalanságot teszik kezelhetővé: a megkülönböztethetetlen objektumokról ugyanazt kell mondani, s ez kétségessé/bizonytalanná tesz bizonyos állításokat. A megkülönböztethetetlenség egy ekvivalencia relációval reprezentálható, mely a háttértudásunkat, illetve annak korlátozottságát reprezentálja. Az így kapott ekvivalencia osztályok azokat az objektumokat tartalmazzák, melyek megkülönböztethetetlenek egymástól. Ezeket az osztályokat alaphalmazoknak nevezzük. A megkülönböztethetetlenség hatással van az eleme reláció használatára következésképpen a tartalmazás relációra is, hiszen bizonyos esetekben bizonytalanná teszi a reláció megítélését, életlenné teszi a halmazt azáltal, hogy egy adott objektumra vonatkozó döntés kihat a tőle megkülönböztethetetlen objektumokra vonatkozó hasonló döntésekre. Tulajdonképpen kénytelenek vagyunk bizonyos objektumokat ugyanúgy kezelni. Ezt a bizonytalanságot halmaz-approximációs eszközökkel reprezentáljuk. Egy halmaz alsó közelítése azon objektumokat tartalmazza, melyek biztosan benne vannak a halmazban, míg a felső közelítése pedig azokat, melyek lehetséges, hogy elemei az adott halmaznak. Tehát egy

életlen halmaz két másik halmazzal definiálhatjuk. Ha nagy méretű információs rendszerekből szeretnénk minél több hasznosítható információt kinyerni, akkor elkerülhetetlen a megkülönböztethetetlenség kezelése. Az életlen halmazok elméletében arra szeretnénk választ kapni, hogy hogyan jellemezhetőek bizonyos halmazok, illetve, hogy egyes tulajdonságok által generált halmazba beletartozik-e egy bizonyos objektum vagy sem.

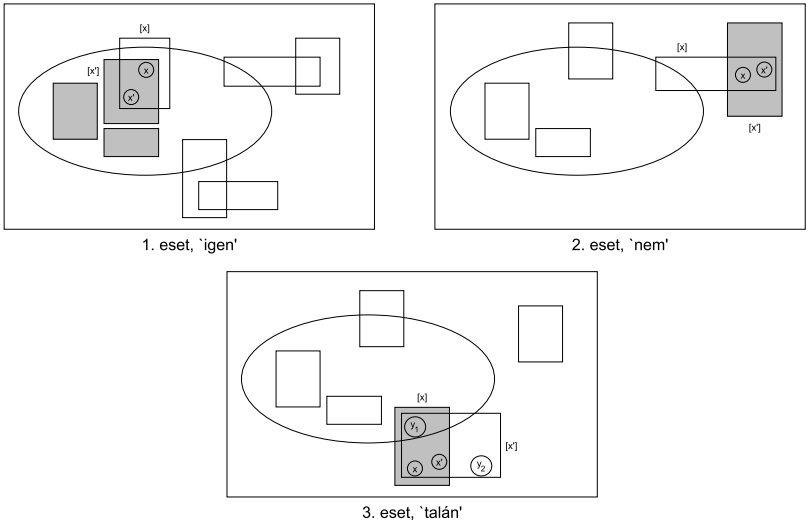
Az adatbányászat az adatok hihetetlen mértékű növekedése miatt az informatikának egy rendkívül fontos és folyamatosan fejlődő ágává vált. Az adatbányászat egy olyan technika, mellyel hasznos információ nyerhető ki automatikus módon nagy mennyiségű adatokból. Célja az adattárak átkutatása annak érdekében, hogy olyan új és hasznos mintázatokat találjanak, amelyek egyébként ismeretlenek maradnának. Az élet számos területén alkalmazhatóak az adatbányászati módszerek. Az életlen halmazok elmélete kulcsfontosságú lehet az adattudományokban [1], hiszen nagy mennyiségű adatok esetén a bizonytalanság megfelelő kezelése elengedhetetlen. Az adatbányászat területén kiválóan alkalmazható az életlen halmazok elmélete. Az adatok előfeldolgozása esetén számos esetben használnak rá épülő módszereket. Előszeretettel alkalmazzák a szabály alapú osztályozásnál, társítási (asszociációs) szabályok generálásánál vagy akár klaszterezésnél is.

2. Eredmények

A gyakorlati alkalmazások során gyakran előfordul, hogy nemcsak a megkülönböztethetetlen objektumokat kell ugyanúgy kezelnünk, hanem azokat az objektumokat is, amelyek egy meghatározott szempontból hasonlóak egymáshoz. (Az adott vizsgálat szempontjából érdektelen különbségeket nem kell figyelembe venni.) Az évek során számos általánosítását dolgozták ki az eredeti pawlaki tereknek [2]. A háttértudást (illetve ennek korlátozottságát) egy halmazrendszerrel jelenítik meg az életlen halmazok elméletében. A halmazrendszer elemeit alaphalmazoknak nevezzük. Az alaphalmazok rendszerében rejlő különbségek különböző approximációs tereket hoznak létre. A lefedő (covering) típusú approximációs terek alaphalmazainak uniója lefedi a vizsgálat hátterében megjelenő teljes objektumkört. Egyes lefedő approximációs terek tolerancia relációt használnak, melyek hasonlóságot reprezentálnak, az ekvivalencia reláció helyett, azonban ezeknek a relációknak az alkalmazása nagyon speciális. A hangsúlyt az egy objektumhoz való hasonlóságra fektetik, nem pedig a hasonlóságra 'általában'. Ez azt jelenti, hogy az alaphalmazok azokat az elemeket tartalmazzák, melyek egy bizonyos objektumhoz hasonlatosak. Ezekben a terekben minden objektum generál egy alaphalmazt.

Legyen U objektumoknak egy univerzuma, $\mathfrak{B}$ az alaphalmazrendszer és l,u az alsó és felső közelítések. Ha arra vagyunk kíváncsiak, hogy $x \in S$ (ahol S a közelítendő halmaz), akkor 3 lehetséges eset lép fel (lásd 2.1. ábra):

- ha $x \in \mathsf{l}(S)$ (azaz x biztosan elem S-nek), akkor $y \in S$ minden $y, y \in [x']$ ahol $x' \in [x]$ és $[x'] \in \{B \mid B \in \mathfrak{B} \text{ és } B \subseteq S\}$;
- ha $x \in \bigcup(\{B \mid B \in \mathfrak{B} \text{ és } B \cap S \neq \varnothing\} \setminus \{B \mid B \in \mathfrak{B} \text{ és } B \subseteq S\})$ (azaz x lehetséges, hogy eleme S-nek), akkor van legalább egy olyan x' és egy alaphalmaz $[x']$ hogy $x \in [x']$, $[x'] \cap S \neq \varnothing$, $[x'] \not\subseteq S$ és y lehet, hogy eleme S-nek minden $y \in [x']$;
- ha $x \in \mathsf{l}(\overline{S})(= U \setminus \mathsf{u}(S))$ (azaz x biztosan nem eleme S-nek), akkor $y \notin S$ bármely $y, x\mathcal{R}y$.



2.1. ábra. A tolerancia reláción alapuló lefedő terek lehetséges esetei

Néhány gyakorlati probléma is megjelenik ezekben a terekben:

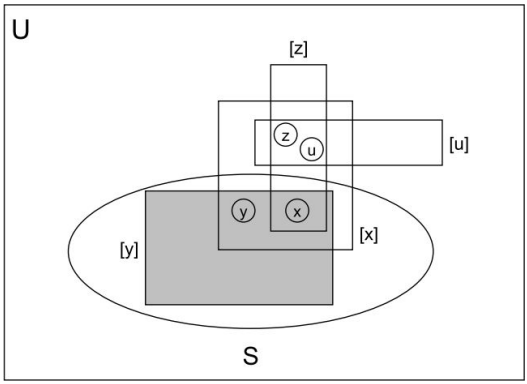
1. Az előző lehetőségek alapján azt mondhatjuk, hogy az alsó és felső közelítések nem zártak az alábbi értelemben (lásd 2.2. ábra):

   (a) Ha $x \in \mathsf{l}(S)$, akkor nem mondhatjuk azt, hogy $[x] \subseteq S$.

   (b) Ha $x \in \mathsf{u}(S)$, akkor nem mondhatjuk azt, hogy $[y] \cap S \neq \varnothing$ minden $y \in [x]$.

2. Előfordulhat, hogy az alaphalmazok száma megegyezik az objektumok számával, azaz a gyakorlati alkalmazásokhoz túl sok alaphalmaz jelenik meg.

A korrelációs klaszterezés egy olyan klaszterező eljárás, mely egy tolerancia relációra támaszkodik. Az eredménye egy partíció és az egyes klaszterek azokat az objektumokat tartalmazzák, melyek a leginkább hasonlítanak. Természetes kérdésként merül fel, hogy nem értelmezhető-e az eredményül kapott partíció az alaphalmazok rendszereként. Kutatásunk során arra következtettünk, hogy érdemes az alaphalmazrendszert a korrelációs klaszterezéssel generálni.



2.2. ábra. A tolerancia reláción alapuló lefedő terekben az alsó és felső közelítések nem zártak

Sikeresen kifejlesztettem egy olyan szoftvert mely segítségével különböző halmazok közelíthetőek a bemutatott térben. A szoftver letölthető az alábbi link segítségével:
`https://github.com/Nagy-David/Similarity-Based-Rough-Sets`.

1. Tézis. ***A korrelációs klaszterezés eredményeként kapott partíció értelmezető alaphalmazrendszerként. Ezzel egy új approximációs tér generálható, mely az objektumok közötti hasonlóságon alapszik és különbözik a már létező terektől, melyek szintén hasonlóságot reprezentáló tolerancia reláción alapszanak. Az approximációs tér neve hasonlóság alapú életlen halmazok*** [1]***.***

A bemutatott tér a következő tulajdonságokkal rendelkezik:

- Az objektumok hasonlósága a tulajdonságaikon alapszik (nem pedig egy bizonyos objektumhoz való hasonlóságon), mely fontos szerepet játszik az alaphalmazok definíciójában.

- Az alaphalmazok páronként diszjunk halmazok, így az alsó és felső közelítések zártak.

[1] A tézisben bemutatott eredmények a 8-as számú referált közleményben olvashatóak.

- Megfelelő számú alaphalmaz jelenik meg (az alkalmazások során elfogadható számú alaphalmazzal tudunk dolgozni).
- Az alaphalmazok mérete nem túl kicsi és nem is túl nagy.

Az alaphalmazrendszer a következőképpen definiálható:

2.1. Definíció.

$$\mathfrak{B} = \{B \mid B \subseteq U, \text{ és } x, y \in B \text{ ha } p(x) = p(y)\},$$

ahol p a korrelációs klaszterezés által generált partíció.

Kereső algoritmusok

A korrelációs klaszterezés problémája belátható időn belül kizárólag keresők és optimalizáló algoritmusok alkalmazásával oldható meg. A különböző algoritmusok azonban különböző partíciókat generálhatnak, így az alaphalmazunk rendszere is változhat. Természetes kérdés tehát, hogy az egyes keresők milyen hatással lehetnek az alaphalmazok szerkezetére. Mivel a korrelációs klaszterezésen alapuló halmazközelítés egy teljen új módszer, így kulcsfontosságú megtalálni a legmegfelelőbb algoritmust. Kutatásunk során a következő, az irodalomban is jól ismert algoritmusokat használtuk:

- Hegymászó algoritmus
- Sztochasztikus hegymászó algoritmus
- Tabu kereső
- Szimulált hűtés
- Párhuzamos szimulált hűtés
- Genetikus algoritmus
- Méhek algoritmusa
- Rovarraj implementáció
- Szentjánosbogár algoritmus

Az algoritmusok összehasonlítására a következő mutatószámokat számoltuk ki:

- Szingleton klaszterek száma
- Az alaphalmazok méretének szórása
- Az alaphalmazok méretének terjedelme
- Az algoritmusok futási ideje

Az alaphalmazok méretén a halmazok számosságát értjük. Ahogy az előző fejezetben is olvasható, a szingleton klaszterek rendkívül kevés információt jelentenek. Minél nagyobb a számuk, annál pontatlanabb információt kapunk a tudásunkról. Minden algoritmus esetén az optimális eset az ha a lehető legkevesebb számú szingletont generálja. Az alaphalmazok méretét is érdemes vizsgálni. A közelítések során az a megfelelő, ha a méretek nem mutatnak túl nagy eltérést, tehát a szórásuk és a terjedelmük minimális. Egy fontos paraméter az algoritmusok futási ideje is. Kulcsfontosságú azokban az esetekben, mikor az objektumok száma nagy, mely a gyakorlati alkalmazások során szinte tövényszerű. Az algoritmusok összehasonlításainál véletlen gráfok segítségével generáltuk a hasonlóságot reprezentáló tolerancia relációkat.

2. Tézis. ***Az előző kritériumok alapján a szimulált hűtés algoritmus produkálta a legelfogadhatóbb eredményeket véletlen gráfok esetén*[2].**

Hasonlóság alapú életlen halmazok annotációval

A szingleton klaszterek rendkívül kevés információt hordoznak, hiszen az elemük kizárólag önmagával tekinthető hasonlatosnak. Ezek az elemek tehát minden esetben egyéni döntéseket igényelnek. A szingletonok elhagyásával egy parciális approximációs tér generálható. Előfordulhat azonban, hogy egy objektum azért nem tartozik egyetlen klaszterhez sem, mert az adott háttértudás alapján a rendszer nem találta hasonlónak egyetlen objektumhoz sem. Ez nem jelenti minden esetben azt, hogy az objektum kizárólag önmagához hasonlít, hanem a megfelelő információhiány miatt (pl. zaj) a rendszer képtelen volt egyetlen klaszterhez is besorolni őt. A felhasználók általában rendelkeznek valamilyen háttértudással az adott adatról, melyet használhatnának arra, hogy az egyes szingleton klaszterek

[2] A tézisben bemutatott eredmények az 1-es számú referált közleményben olvashatóak.

elemét besorolják valamely alaphalmazba (nem szingleton klaszter). Ezt a folyamatot annotációnak nevezzük.

Legyen S a közelítendő halmaz, $\{x\}$ egy szingleton és B egy alaphalmaz. A B alaphalmazzal a következő esetek történhetnek az annotáció után, ha $B \subseteq \mathsf{l}(S)$:

- Ha $x \in S$, akkor $B' = \{x\} \cup B$ és $B' \subseteq \mathsf{l}(S)$ Így az S halmaz közelítése pontosabbá válik.
- Ha $x \notin S$, akkor $B' = \{x\} \cup B$ és $B' \subseteq \mathsf{u}(S)$ de $B' \not\subseteq \mathsf{l}(S)$ Így az S-hez való bizonytalanság növekszik.

A B alaphalmazzal a következő esetek történhetnek az annotáció után, ha $B \subseteq \mathsf{u}(S)$:

- Ha $x \in S$, akkor $B' = \{x\} \cup B$ és $B' \subseteq \mathsf{u}(S)$
- Ha $x \notin S$, akkor $B' = \{x\} \cup B$ és $B' \subseteq \mathsf{u}(S)$

A B alaphalmazzal a következő esetek történhetnek az annotáció után, ha $B \subseteq \mathsf{u}(S) \setminus \mathsf{l}(S)$:

- Ha $x \in S$, akkor $B' = \{x\} \cup B$ és $B' \subseteq \mathsf{u}(S) \setminus \mathsf{l}(S)$
- Ha $x \notin S$, akkor $B' = \{x\} \cup B$ és $B' \subseteq \mathsf{u}(S) \setminus \mathsf{l}(S)$

Mindenkét esetben a felső közelítés és a határterület nagyobb lesz. Elmondható, hogy az annotáció függ a közelítendő halmaztól. Hasznos lehet tehát a következő:

- $x \in S$, akkor a felhasznló csak azokból a B alaphalmazokból választhasson, melyekre $\mathsf{l}(S)$.
- $x \notin S$, akkor a felhasznló csak azokból a B alaphalmazokból választhasson, melyekre $\mathsf{l}(\overline{\mathsf{u}(S)})$, ahol $\overline{\mathsf{u}(S)}$ a felső közelítés komplementere.

Az annotáció sorrendjét is érdemes vizsgálni. Ha az x_1, x_2 két különböző szingletonhoz tartozó elem, melyeket ugyanabba a B alaphalmazba szeretnénk besorolni, akkor a következő kérdést kell megválaszolni. Az x_2 objektumot az x_1 után is releváns lehet besorolni B-be?

- Ha a válasz igen, akkor a két objektum felcserélhető. Ezt azt jelenti, hogy az x_1, x_2 objektumok között valamilyen hasonlóság van, mely a hasonlóságot reprezentáló tolerancia relációban van elrejtve.
- Ha a válasz nem, akkor a két objektum nem felcserélhető. Ez azt jelenti, hogy x_1 irrelevánsá teszi x_2 besorolását B-be.

Az annotáció támogatása

Számos esetben előfordulhat, hogy egy felhasználó rendelkezik olyan sajátos tudással/információval, melyet az adatok nem jelenítenek meg, így a rendszer nem ismeri ezeket. Emiatt igen hasznos lehet, ha lehetősége van a felhasználónak felülvizsgálnia a rendszer által hozott döntést azáltal, hogy egyes szingleton klaszterek elemeit valamely alaphalmazba beilleszti. Az annotáció egy ilyen folyamat, hiszen egy lehetséges kommunikációs csatornát biztosít a felhasználó és a rendszer között. Természetesen az egész folyamat a felhasználó szaktudásán alapszik. Ez nem azt jelenti azonban, hogy ne lehetne valamilyen segítséget biztosítania a rendszernek. Munkánk során két technikát mutattunk be. Az első egy grafikus módszer, mely a hasonlóságon alapuló tolerancia relációk egy vizuális reprezentációját próbálja megadni. Ha ebben a rendszerben két objektum közel van, akkor ez azt jelenti, hogy őket akár hasonlóként is lehet kezelni. Ha egy szingleton klaszter eleme közel van egy másik klaszterhez, akkor elképzelhető, hogy össze kellene vonni őket. A másik technika egy matematikai eljárás, mely minden klaszter esetén azokat az elemeket kívánja megtalálni, mely az adott klaszter egészét reprezentálja. A számítások során kizárólag a reprezentatív elemekket érdemes figyelembe venni. Így rengeteg idő és erőforrás megtakarítható. Egyes esetekben előfordulhat, hogy egy szingleton eleme több alaphalmazba is beleilleszthető lenne. Ilyenkor érdemes lehet abba az alaphalmazba besorolni az objektumot, melynek reprezentánsa a leginkább hasonlít az adott elemhez. Így nem szükséges minden alaphalmazbeli elemmel összehasonlítást végezni.

Approximációs párok a hasonlóság alapú életlen halmazokon

A hasonlóság alapú életlen halmazok egy olyan approximációs tér, mely a korrelációs klaszterezés által generált partíción alapszik. Egy tetszőleges halmaz alsó közelítése azon alaphalmazok uniója, melyek részhalmazai a közelítendő halmaznak. Ezen alaphalmazok meghatározásánál minden elemüket figyelembe kell venni, mely egy időigényes folyamat lehet nagy elemszám esetén. Az reprezentatív elemek hasznossága azokban esetekben rejlik, mikor az objektumok száma nagyon nagy. Hasznosnak tűnhet a reprezentánsok jó tulajdonságait a halmazközelítések esetén is használni. Legyen $B \in \mathfrak{B}$ egy alaphalmaz és $REP(B)$ a reprezentásainak halmaza. A reprezentatív elemeken alapuló approximációs párok a következőképpen definiálhatóak:

- $\mathsf{l}_\mathsf{r}(S) = \bigcup\{B \mid B \in \mathfrak{B} \text{ és } \forall x \in REP(B) : x \in S\}$;
- $\mathsf{u}_\mathsf{r}(S) = \bigcup\{B \mid B \in \mathfrak{B} \text{ és } \exists x \in REP(B) : x \in S\}$.

Így egy halmaz alsó közelítése azon alaphalmazok uniója lesz, melynek minden reprezentánsa eleme a közelítendő halmaznak. Egy alaphalmaz a felső közelítésbe tartozik, ha legalább egy reprezentatív eleme benne van a közelítendő halmazban. Természetesen az alsó közelítés bizonyosságából veszítünk, de az objektumok számának növekedésével rendkívül sok erőforrást takaríthatunk meg, mely egy olyan fontos tulajdonság, ami miatt érdemes lehet a bizonyosságot feladni.

3. Tézis. ***Az annotációs folyamat segítségével a felhasználók a saját tudásokat implementálhatják a rendszerbe interaktívvá téve azt. Az annotáció eredménye egy új approximációs tér, melyben csökkentett a parcialitás mértéke a kevesebb, alaphalmazba nem tartozó objektum miatt. A bemutatott reprezentatív elemkre támaszkodó approximációs pár pedig csökkentheti a futási idejét és az erőforrásigényét az approximációs folyamatnak különösen nagy mennyiségű adatok esetén*** [3]***.***

Halmaz központú approximációs párok

Ebben az alfejezetben további két, a reprezentatív elemekre épülő approximációs pár kerül bemutatásra.

A két approximációs pár $\langle \mathsf{l}, \mathsf{u}_1 \rangle$ és $\langle \mathsf{l}, \mathsf{u}_2 \rangle$ a következőképpen definiálható:

$$\mathsf{l}(S) = \bigcup_{\substack{x \in REP(S) \\ [x] \subseteq S}} \{[x]\}$$

$$\mathsf{u}_1(S) = \bigcup_{x \in REP(S)} \{[x]\}$$

$$\mathsf{u}_2(S) = \bigcup_{\substack{x \in REP(S) \\ \|[x] \cap S\| > \|[x] \setminus S\|}} \{[x]\}$$

Ebben a térben egy alaphalmaz azon objektumokat tartalmazza, mely az adott reprezentatívhoz hasonlatos, tehát minden reprezentatív egy

[3] A tézisben bemutatott eredmények a 4-es és 7-es számú referált közleményekben olvashatóak.

alaphalmazt fog generálni. Mivel a reprezentatívok a közelítendő halmaztól függnek, ezért az alaphalmazok rendszere is függni fog tőle. Így ha a közelítendő halmaz változik, akkor az alaphalmazok is vele együtt fognak változni (az alaphalmazok relatívak a közelítendő halmazhoz képest).

Az approximációs pároknak számos érdekes tulajdonsága létezik, melyeket érdemes megvizsgálni. Kutatásunk során a következőket tanulmányoztuk:

2.2. Definíció.

- *Monotonitás:* l *és* u *monotonok ha* $S \subseteq S'$*, akkor* $\mathsf{l}(S) \subseteq \mathsf{l}(S')$ *és* $\mathsf{u}(S) \subseteq \mathsf{u}(S')$
- *Gyenge approximációs tulajdonság:* $\forall S \in 2^U : \mathsf{l}(S) \subseteq \mathsf{u}(S)$
- *Erős approximációs tulajdonság:* $\forall S \in 2^U : \mathsf{l}(S) \subseteq S \subseteq \mathsf{u}(S)$
- l *normalitása:* $\mathsf{l}(\varnothing) = \varnothing$
- u *normalitása:* $\mathsf{u}(\varnothing) = \varnothing$

4. Tézis. ***Az alsó és felső közelítések normalitása valamint a gyenge approximációs tulajdonság mindkét approximációs pár esetén teljesülnek. A monotonitás pedig egyik párra sem teljesül. Az erős approximációs tulajdonság kizárólag*** $\langle \mathsf{l}, \mathsf{u}_1 \rangle$***-re érvényes***[4]***.***

[4] A tézisben bemutatott eredmények a 3-as számú referált közleményben olvashatóak.

3. Összefoglalás

A dolgozat legfontosabb eredménye egy olyan teljesen új approximációs tér kifejlesztése volt, mely egy tolerancia reláción alapul és az alaphalmazok a korrelációs klaszterezés eredményeként jönnek létre. A bemutatott tér (hasonlóság alapú életlen halmazok) számos jó tulajdonsággal rendelkezik és különbözik a tolerancia reláción alapuló lefedő terektől. A legfontosabb különbség a kifejlesztett tér és ezen lefedő terek között az, hogy az általunk bevezetett tér az objektumok közötti valódi hasonlóságon alapszik, míg az említett lefedő terek csak egy bizonyos objektumhoz való hasonlóságon. A szingleton klaszterek rendkívül kevés információt hordoznak, hiszen az eleme csak önmagához tekinthető hasonlatosnak. Ha a szingleton klasztereket nem tekintjük alaphalmaznak, akkor egy parciális approximációs teret generálhatunk. A térnek egy hasznos továbbfejlesztése lehet, ha megengedünk egy lehetséges interakciót a felhasználó és a rendszer között. Ezt a folyamatot annotációnak neveztük. Tulajdonképpen a felhasználó a saját tudását implementálhatja a rendszerbe azáltal, hogy egyes szingleton klaszterek elemét valamely alaphalmazba illeszti be. Ezáltalán csökken a parcialitás mértéke is. Bemutattunk két olyan módszert is, mely segítséget nyújt az annotációs folyamatban. Az első egy vizualizációs technika, mellyel könnyedén lehet hasonlóságot reprezentáló tolerancia relációkat ábrázolni. A reprezentációban két objektum közelsége a két objektum hasonlóságára utal, azaz ha egy szingleton klaszter elem közel van valamely alaphalmazhoz, akkor érdemes lehet abba az alaphalmazba beilleszteni az objektumot. Elkészítettünk továbbá olyan algoritmusokat, mellyel minden klaszterből ki lehet választani azon elemeket, melyek a leginkább hasonlítanak a többihez. Az ilyen objektumokat reprezentatív elemeknek neveztük. A módszer jó tulajdonsága, hogy megfelelő számú (nem túl sok és nem túl kevés) reprezentánst válogat ki, így csökkentve a számolás során figyelembe veenedő objektumok számát. A reprezentatív elemeket az annotáció támogatására is alkalmaztuk. A reprezentatív elemekre támaszkodva több új approximációs-párt is definiáltunk. A hasonlóság alapú életlen halmazokra támaszkodó gráf-approximációs folyamattal pe-

dig egy olyan algoritmus került bemutatásra, mellyel egy adathalmaz két attribútumhalmaza közti szorosságot lehet vizsgálni, és az adatok előfeldolgozásában használatos ún. jellemző kinyerésben is alkalmazható.

Short Thesis for the degree of doctor of philosophy (PhD)

SIMILARITY BASED ROUGH SETS AND ITS APPLICATIONS IN DATA MINING

by Dávid Nagy

Supervisor:
Dr. Tamás Mihálydeák

UNIVERSITY OF DEBRECEN
Doctoral School of Informatics

Debrecen, 2020

1 Introduction and Motivation

Nowadays the amount of data is growing exponentially. In some cases, data are often incomplete or inconsistent. Incompleteness can happen if some value is unknown, unassigned or even inapplicable. Inconsistency occurs when the data are contradictory. These issues can cause some undesirable events (bad prediction, inappropriate decision making, etc). In computer science, there are numerous ways to handle these kinds of inaccuracies.

Rough set theory can be considered as a rather new field in computer science. Its fundamentals were proposed by professor Pawlak in the 80's [3, 4, 5]. In many cases, based on the available knowledge, two objects cannot be distinguished from each other because all of their known properties are the same. The Pawlakian spaces handle the uncertainty arising from indiscernibility based on the available data. The same statement must be stated for the objects that are indiscernible from each other and this makes some statements uncertain/doubtful. This indiscernibility can be modeled by an equivalence relation which represents our background knowledge or its limits. The equivalence classes gained this way contain objects that are indiscernible from each other. These classes are called base sets. The relation can affect the membership relation by making the judgment on this relation uncertain. It also makes a set vague because a decision about a certain object has an effect on the decisions about all the objects that are indiscernible from the given object. This uncertainty can be represented by set-approximation tools. The lower approximation of a set contains objects that surely belong to the set, and the upper approximation contains objects that possibly belong to the set. So a rough set can be defined by two sets. If we want to extract as much useful information as possible from large-scale information systems, then it is inevitable to handle indiscernibility. Rough set theory tries to answer how certain sets can be characterized or if a given object belongs to a set generated by some property.

Data mining became a very important and popular field in computer

science due to the incredible increase in data. With its help, useful information can be extracted automatically from a large amount of data. Its goal is to search for new and useful patterns, which could otherwise remain unknown, in data repositories. Data mining methods can be applied in many areas of life. Rough set theory can be crucial in data sciences [1], because handling the uncertainty is necessary in case of a large amount of data. There numerous fields in data mining in which rough set theory can be successfully applied. There are many existing data pre-processing techniques based on rough set theory. In many data mining techniques, it also proved to be very useful. For example in decision rule induction, association rule mining, clustering, etc.

2 Results

In practical applications not only the indiscernible objects must be handled in the same way but also those that are similar to each other based on some property. (Irrelevant differences for the purpose of the given applications should not be taken into account.) Over the years, many new approximation spaces have been created as the generalization of the original Pawlakian space [2]. In rough set theory, the background knowledge (or its limit) is represented by a family of sets. The members of this family are called base sets. The differences between the various systems of base sets generate different approximation spaces. The union of the base sets of the covering type approximation spaces covers every object of the investigated system. Some covering approximation spaces use tolerance relations, which represent similarity, instead of equivalence relations, but the usage of these relations is very special. It emphasizes the similarity to a given object and not the similarity of objects 'in general'. This means that a base set contains objects which are similar to a distinguished object. In these spaces, each object generates a base set. In these spaces, if one is interested in whether $x \in S$ (where S is the set to be approximated), then there are three possible answers (see Fig. 2.1):

- if $x \in \mathsf{l}(S)$ (i.e. x is surely a member of S), then $y \in S$ for all $y, y \in [x']$ where $x' \in [x]$ and $[x'] \in \{B \mid B \in \mathfrak{B} \text{ and } B \subseteq S\}$;
- if $x \in \bigcup(\{B \mid B \in \mathfrak{B} \text{ and } B \cap S \neq \varnothing\} \setminus \{B \mid B \in \mathfrak{B} \text{ and } B \subseteq S\})$ (i.e. x is possibly a member of S), then there is an x' and a base set $[x']$ such that $x \in [x']$, $[x'] \cap S \neq \varnothing$, $[x'] \not\subseteq S$ and y may be a member of S for all $y \in [x']$;
- if $x \in \mathsf{l}(\overline{S})(= U \setminus \mathsf{u}(S))$ (i.e. x is surely not a member of S), then $y \notin S$ for all $y, x\mathcal{R}y$.

where U is the universe of objects, $\mathfrak{B}$ is the system of base sets and l,u are the lower and upper approximation respectively.

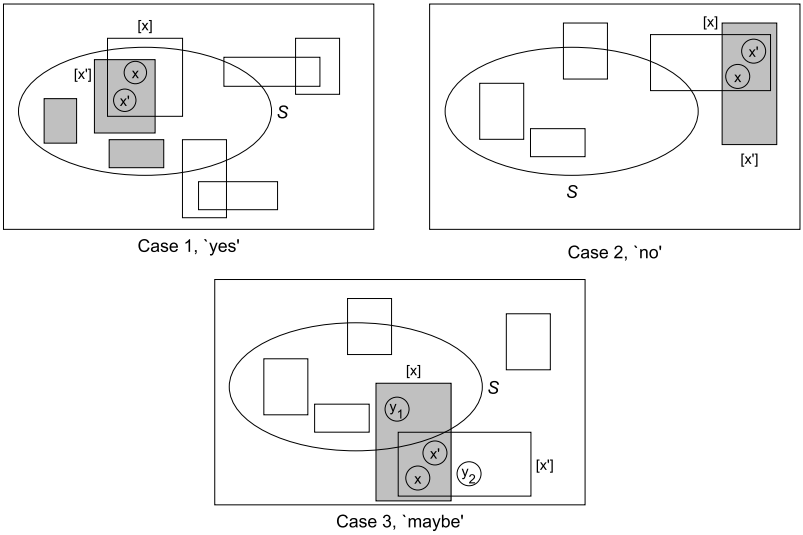


Figure 2.1. Some base sets in covering (based on a tolerance relation) cases

Some practical problems of covering approximation spaces based on a tolerance relation:

1. The former answers show that generally the lower and upper approximations are not closed in the following sense (see Fig. 2.2):

   (a) If $x \in \mathsf{l}(S)$, then it cannot be said that $[x] \subseteq S$.

   (b) If $x \in \mathsf{u}(S)$, then it cannot be said, that $[y] \cap S \neq \varnothing$ for all $y \in [x]$.

2. It can happen that the number of base sets equals the number of objects, so there can be too many base sets for practical applications.

Correlation clustering is a clustering technique that is based on a tolerance relation. Its result is a partition and the clusters contain objects that are mostly similar to each other. So a natural question arises: can the partition represent the system of base sets? In our research, we found that it is worth to generate the system of base sets by the correlation clustering.

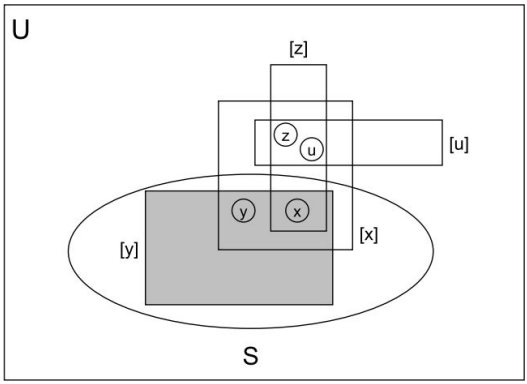


Figure 2.2. In covering (based on a tolerance relation) the lower and upper approximations are not closed

I developed a program that helps us approximate different types of sets using the proposed new approximation space. The software can be downloaded from
`https://github.com/Nagy-David/Similarity-Based-Rough-Sets`.

Thesis 1. ***The partition gained from the correlation clustering can be understood as a system of base sets. As a result, a new approximation space appears which is based on the similarity of objects and it is different from the existing approximation spaces that are based on a tolerance relation representing similarity. This approximation space is called similarity based rough sets*[1]*.***

The proposed approximation space has the following features:

- The similarity of objects relying on their properties (and not the similarity to a distinguished object) plays a crucial role in the definition of base sets.
- The system of base sets consists of disjoint sets, so the lower and upper approximations are closed.
- Only the necessary number of base sets appears (in applications we have to use an acceptable number of base sets).

[1] The results described in this thesis was based on publication no. 8.

- The size of base sets is not too small, or too big.

The definition of the system of base sets is the following:

Definition 2.1.

$$\mathfrak{B} = \{B \mid B \subseteq U, \text{ and } x, y \in B \text{ if } p(x) = p(y)\},$$

where p is the partition gained from the correlation clustering.

Search algorithms

In a reasonable time, correlation clustering can only be solved by using search algorithms. However, each algorithm can provide different partitions. So the system of base sets can also be different. It is a natural question to ask how the search algorithms can affect the structure of the base sets. As the approximation based on correlation clustering is a completely new way of approximating sets, so it is crucial to use the best possible search algorithm. The following list shows the used algorithms in the experiments.

- Hill Climbing Algorithm
- Stochastic Hill Climbing Algorithm
- Tabu Search
- Simulated Annealing
- Parallel Tempering
- Genetic Algorithm
- Bees Algorithm
- Particle Swarm Optimization
- Firefly Algorithm

To compare the algorithms, the following values were computed:

- Number of singleton clusters
- Standard deviation of the base set sizes
- The range of the base set sizes

- Execution time of the algorithm

Here, the cardinality of a base set is referred to as its size. As previously mentioned, singleton clusters mean little information. The greater their number is, the more unclear the information based on our knowledge becomes. For a search algorithm, the most optimal is if it provides the least number of these clusters in order to have a precise result of the system. The sizes of the base sets are also worth to be checked. For set approximations, it is more suitable if the sizes do not vary much. So the standard deviation of the base set sizes should be minimized as well as the range of the base set sizes. An important parameter is the execution time of the search algorithms. It is especially crucial when there are a huge number of objects.

In the comparisons, we used random graphs to generate the tolerance relations (representing similarity).

Thesis 2. ***Based on the aforementioned criteria, the simulated annealing algorithm provided the most acceptable result in the case of random graphs*[2].**

Similarity based rough sets with annotation

Singleton clusters represent very little information because its member is only considered to be similar to itself. So they always require an individual decision. By discarding the singletons, a partial system of base sets is generated. However, it can sometimes happen that an object does not belong to any cluster because the system could not consider it similar to any other objects based on the background information. This does not mean that this object is only similar to itself, but without proper information (e.g. due to noise) the system could not insert it into any cluster. The users usually possess some background knowledge of the given data. They can use this knowledge to help the system by inserting the members of some singletons into base sets (non-singleton clusters). This process is called annotation.

Let S be the set to be approximated, $\{x\}$ be a singleton gained from the correlation clustering and B be a base set. The following cases can happen with the base set B after the annotation if $B \subseteq \mathsf{I}(S)$:

[2]The results described in this thesis was based on publication no. 1.

- If $x \in S$, then $B' = \{x\} \cup B$ and $B' \subseteq \mathsf{l}(S)$ This way the approximation of the set S becomes more precise.
- If $x \notin S$, then $B' = \{x\} \cup B$ and $B' \subseteq \mathsf{u}(S)$ but $B' \not\subseteq \mathsf{l}(S)$ This increases the uncertainty relative to the set S.

The following cases can happen with the base set B after the annotation if $B \subseteq \mathsf{u}(S)$:

- If $x \in S$, then $B' = \{x\} \cup B$ and $B' \subseteq \mathsf{u}(S)$
- If $x \notin S$, then $B' = \{x\} \cup B$ and $B' \subseteq \mathsf{u}(S)$

The following cases can happen with the base set B after the annotation if $B \subseteq \mathsf{u}(S) \setminus \mathsf{l}(S)$:

- If $x \in S$, then $B' = \{x\} \cup B$ and $B' \subseteq \mathsf{u}(S) \setminus \mathsf{l}(S)$
- If $x \notin S$, then $B' = \{x\} \cup B$ and $B' \subseteq \mathsf{u}(S) \setminus \mathsf{l}(S)$

In both cases, the upper approximation and the boundary region become larger. It can be said that the annotation depends on the set to be approximated. It could be useful if:

- $x \in S$, then the user could only choose from those B base sets which are in $\mathsf{l}(S)$.
- $x \notin S$, then the user could only choose from those B base sets which are in $\mathsf{l}(\overline{\mathsf{u}(S)})$, where $\overline{\mathsf{u}(S)}$ denotes the complement of the upper approximation.

The order of the annotation is also worth to be checked. If the members x_1, x_2 of 2 different singletons were to be inserted into the same base set B, then the following question needs to be answered. Is it still relevant to insert x_2 into B after putting x_1 into B?

- If the answer is yes, then the two members are interchangeable. This means that x_1, x_2 has some sort of similarity that was hidden in the tolerance relation (based on similarity) .
- If the answer is no, then the two members are not interchangeable. This means that annotating x_1 makes it irrelevant to insert x_2 into B.

Tools of the annotation

In many cases, a user may possess some knowledge/information which cannot be extracted from the data making the system unable to recognize them. So it can sometimes be useful if the users have an option to reconsider the decision made by the system by inserting the member of a singleton into a base set. The annotation is such a process as it provides a possible communication channel between the user and the system. Naturally, the entire process is based on the user's expertise. However, this does not mean that there should not be any help or suggestions provided by the system itself. We introduced two techniques. The first one is a graphical method which tries to give a visual representation of the tolerance relation based on similarity. If two objects are close in this representation, then this indicates that those two objects should be treated as similar. If a member of a singleton is close to another cluster, then maybe they should be merged. The second method is a mathematical way that aims to find those members in each cluster that can represent the entire cluster. During the computations, only the representative members should be considered. This way a lot of time and resources can be saved. It is sometimes possible that there are more than one suitable base sets into which the user should insert the member of a singleton. In this case, the recommended base set should be the one whose representative member is the most similar to the member of the given singleton. In this way, there is no need to compare it to each member of each base set.

Approximation pairs based on similarity based rough sets

Similarity based rough sets is an approximation space which is based on the partition generated by correlation clustering. The lower approximation of a set S is the union of those base sets that are subsets of S. In order to get these base sets, every point in each base set must be considered. It can be a time-consuming task if the number of points is high. The effectiveness of the representatives lies in situations when the number of objects is very large. It can be practical to use the power of representatives in the approximation process. For each base set, let us consider only its representatives. Let $B \in \mathfrak{B}$ be a base set, and $REP(B)$ be the set of its representatives. The approximation functions are defined as the following:

- $\mathsf{l}_\mathsf{r}(S) = \bigcup\{B \mid B \in \mathfrak{B} \text{ and } \forall x \in REP(B) : x \in S\}$;

- $\mathsf{u_r}(S) = \bigcup\{B \mid B \in \mathfrak{B} \text{ and } \exists x \in REP(B) : x \in S\}$.

This way, the lower approximation of a set S becomes the union of those base sets for which every representative is a member of S. A base set belongs to the upper approximation if at least one of its representatives is in the set S. Naturally, the certainty of the lower approximation might be lost, but as the number of points is increasing a lot of resources can be saved. This a very important feature for which it can be worth giving up the certainty property.

Thesis 3. ***With the help of the annotation process, the user can put their own knowledge into the system making it interactive. The result of the annotation is a new approximation space in which the partiality is decreased due to the fewer number of objects that do not belong to any base sets. The proposed approximation pair can decrease the execution time and resource need of the approximation process especially if the number of objects huge*** [3]**.**

Set-based approximation pairs

In this subsection, two other new possible approximation pairs are proposed based on the representatives.

The two proposed approximation pairs can be given as $\langle \mathsf{l}, \mathsf{u}_1 \rangle$ and $\langle \mathsf{l}, \mathsf{u}_2 \rangle$, where

$$\mathsf{l}(S) = \bigcup_{\substack{x \in REP(S) \\ [x] \subseteq S}} \{[x]\}$$

$$\mathsf{u}_1(S) = \bigcup_{x \in REP(S)} \{[x]\}$$

$$\mathsf{u}_2(S) = \bigcup_{\substack{x \in REP(S) \\ \|[x] \cap S\| > \|[x] \setminus S\|}} \{[x]\}$$

In this space, a base set contains objects that are similar to a given representative, so each generates a base set. The representatives depend on the set to be approximated and so does the system of base sets. As a result, the base sets change as the set changes (they are relative to the set).

[3] The results described in this thesis was based on publication no. 4 and 7.

There are many interesting properties that can be checked on approximation pairs. In our research, the following properties are examined:

Definition 2.2.

- *Monotonicity:* l *and* u *are said to be monotone if* $S \subseteq S'$ *then* $\mathsf{l}(S) \subseteq \mathsf{l}(S')$ *and* $\mathsf{u}(S) \subseteq \mathsf{u}(S')$
- *Weak approximation property:* $\forall S \in 2^U : \mathsf{l}(S) \subseteq \mathsf{u}(S)$
- *Strong approximation property:* $\forall S \in 2^U : \mathsf{l}(S) \subseteq S \subseteq \mathsf{u}(S)$
- *Normality of* l*:* $\mathsf{l}(\varnothing) = \varnothing$
- *Normality of* u*:* $\mathsf{u}(\varnothing) = \varnothing$

Thesis 4. ***The normality of the lower and upper approximation and the weak approximation property hold for both approximation pairs. The monotonicity does not hold for either pair. The strong approximation property holds only for*** $\langle \mathsf{l}, \mathsf{u}_1 \rangle$ [4]**.**

[4] The results described in this thesis was based on publication no. 3.

3 Summary

The main result of my research was developing a completely new approximation space which was based on a tolerance relation representing similarity. In this space, the system of base sets is generated by the correlation clustering. Correlation is a clustering technique that is also based on a tolerance relation. The proposed space has many good qualities and it is different from the existing covering spaces induced by a tolerance relation (based on similarity). The main difference between our space and these covering spaces is that ours considers the real similarity among the objects while the covering spaces generated by a tolerance relation only consider similarity to a distinguished member. Singleton clusters represent very little information because their members could not be considered as similar to any other objects. If the singleton clusters are removed from the system of base sets, then a partial approximation space appears. A useful improvement can be if we allow a possible interaction between the system and the user. We call this process annotation. The user can use their background knowledge to help the system by inserting the members of some singletons into base sets (non-singleton clusters). With the help of the annotation, the users can put their own knowledge into the system. It also decreases the partiality by decreasing the number of singletons. We proposed two techniques which offer some assistance in the annotation process. The first one is a graphical method which tries to give a visual representation of the tolerance relation based on similarity. The closeness of two objects indicates similarity between them. If a member of a singleton is close to a base set, then maybe they should be merged. We introduced algorithms which select those objects from each cluster that are similar to most of the elements in the cluster. These objects are called representatives. The feature of the method is that it selects the appropriate number (not too few or too many) of representatives and thus decrease the number of objects to be considered in the computations. We used the representatives in the annotation process as well. We developed several new approximation pairs based on the representatives. We also used the

similarity based rough set technique in graph-approximation. The proposed method is good for determining the correlation between two sets of attributes and it can be used in feature selection process as well.

A Tézisek Irodalomjegyzéke / References of the Theses

[1] Bello, R., and Falcon, R. *Rough Sets in Machine Learning: A Review*. Springer International Publishing, Cham, 2017, pp. 87–118.

[2] Mihálydeák, T. Logic on similarity based rough sets. In *Rough Sets* (Cham, 2018), H. S. Nguyen, Q.-T. Ha, T. Li, and M. Przybyła-Kasperek, Eds., Springer International Publishing, pp. 270–283.

[3] Pawlak, Z. Rough sets. *International Journal of Parallel Programming 11*, 5 (1982), 341–356.

[4] Pawlak, Z., et al. Rough sets: Theoretical aspects of reasoning about data. *Systern Theory, Knowledge Engineering and Problem Solving, Kluwer Academic Publishers, Dordrecht, 199l 9* (1991).

[5] Pawlak, Z., and Skowron, A. Rudiments of rough sets. *Information sciences 177*, 1 (2007), 3–27.

Publikációs lista / List of Publications

Referált közlemények / Refereed Publications

1. **Different Types of Search Algorithms for Rough Sets** (Dávid Nagy, Tamás Mihálydeák and László Aszalós), *In Acta Cybernetica*, volume 24, pp. 105-120, 2019.

2. **Finding the representative in a cluster using correlation clustering** (Dávid Nagy, László Aszalós and Tamás Mihálydeák), *In Pollack Periodica*, volume 14, pp. 15- 24, 2019.

3. **Iterative Set Approximations Based on Tolerance Relation** (László Aszalós and Dávid Nagy), In Rough Sets (Tamás Mihálydeák, Fan Min, Guoyin Wang, Mohua Banerjee, Ivo Düntsch, Zbigniew Suraj and Davide Ciucci, ed.), Springer International Publishing, pp. 78-90, 2019.

4. **Approximation Based on Representatives** (Dávid Nagy and László Aszalós), In Rough Sets (Tamás Mihálydeák, Fan Min, Guoyin Wang, Mohua Banerjee, Ivo Düntsch, Zbigniew Suraj and Davide Ciucci, ed.), Springer International Publishing, pp. 91-101, 2019.

5. **Selecting representatives** (László Aszalós and Dávid Nagy), *In Communication Papers of the 2019 Federated Conference on Computer Science and Information Systems (Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki, eds.)*, Annals of Computer Science and Information Systems, PTI, volume 20, pp. 13-19, 2019.

6. **Visualization of tolerance relations** (László Aszalós and Dávid Nagy), *In Proceedings of the 10th International Conference on Applied Informatics (Gábor Kusper and Roland Király ed.)*, pp. 15-22, 2018.

7. **Similarity Based Rough Sets with Annotation** (Dávid Nagy, Tamás Mihálydeák and László Aszalós), In Rough Sets (Hung Son Nguyen, Quang-Thuy Ha, Tianrui Li and Małgorzata Przybyła-Kasperek, ed.), Springer International Publishing, pp. 88-100, 2018.

8. **Similarity Based Rough Sets** (Dávid Nagy, Tamás Mihálydeák and László Aszalós), In Rough Sets (Lech Polkowski, Yiyu Yao, Piotr Artiemjew, Davide Ciucci, Dun Liu, Dominik Ślęzak and Beata Zielosko, ed.), Springer International Publishing, pp. 94-107, 2017.

9. **Conjectures on phase transition at correlation clustering of random graphs** (László Aszalós, János Kormos and Dávid Nagy), *In Annales Univ. Sci. Budapest., Sect. Comp.*, pp. 37-54, 2014.

10. **Graph Approximation on Similarity Based Rough Sets** (Dávid Nagy, Tamás Mihálydeák and László Aszalós), *In Pollack Periodica* [ACCEPTED FOR PUBLICATION]

DEBRECENI EGYETEM
EGYETEMI ÉS NEMZETI KÖNYVTÁR
H-4002 Debrecen, Egyetem tér 1, Pf.: 400
Tel.: 52/410-443, e-mail: publikaciok@lib.unideb.hu

Nyilvántartási szám: DEENK/44/2020.PL
Tárgy: PhD Publikációs Lista

Jelölt: Nagy Dávid
Neptun kód: LF1CXV
Doktori Iskola: Informatikai Tudományok Doktori Iskola
MTMT azonosító: 10056962

A PhD értekezés alapjául szolgáló közlemények

Idegen nyelvű tudományos közlemények hazai folyóiratban (1)

1. **Nagy, D.**, Aszalós, L., Mihálydeák, T. S.: Finding the representative in a cluster using correlation clustering.
   Pollack Period. 14 (1), 15-24, 2019. ISSN: 1788-1994.
   DOI: http://dx.doi.org/10.1556/606.2019.14.1.2

Idegen nyelvű konferencia közlemények (7)

2. **Nagy, D.**, Aszalós, L.: Approximation Based on Representatives.
   In: Rough Sets : International Joint Conference, IJCRS 2019. Eds.: Tamás Mihálydeák, Fan Min, Guoyin Wang, Mohua Banerjee, Ivo Düntsch, Zbigniew Suraj, Davide Ciucci, Springer, Cham, 91-101, 2019, (Lecture Notes in Computer Science, ISSN 0302-9743 ; 11499.) ISBN: 9783030228149

3. **Nagy, D.**, Mihálydeák, T. S., Aszalós, L.: Different Types of Search Algorithms for Rough Sets.
   Acta Cybern. 24 (1), 105-120, 2019. ISSN: 0324-721X.
   DOI: http://dx.doi.org/10.14232/actacyb.24.1.2019.8

4. Aszalós, L., **Nagy, D.**: Iterative Set Approximations Based on Tolerance Relation.
   In: Rough Sets : International Joint Conference, IJCRS 2019. Eds.: Tamás Mihálydeák, Fan Min, Guoyin Wang, Mohua Banerjee, Ivo Düntsch, Zbigniew Suraj, Davide Ciucci, Springer, Cham, 78-90, 2019, (Lecture Notes in Computer Science, ISSN 0302-9743 ; 11499.) ISBN: 9783030228149

5. Aszalós, L., **Nagy, D.**: Selecting representatives.
   In: Communication Papers of the 2019 Federated Conference on Computer Science and Information Systems. Eds.: Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki, Polskie Towarzystwo Informatyczne, Warszawa, 13-19, 2019, (Annals of Computer Science and Information Systems, ISSN 2300-5963 ; 20.) ISBN: 9788395541636

6. **Nagy, D.**, Mihálydeák, T. S., Aszalós, L.: Similarity based rough sets with annotation.
In: Rough Sets. Ed.: Hung Son Nguyen, Quang-Thuy Ha, Tianrui Li, Małgorzata Przybyła-Kasperek, Springer International Publishing, Cham, 88-100, 2018, (Lecture Notes in Computer Science, ISSN 0302-9743 ; 11103.) ISBN: 9783319993676

7. Aszalós, L., **Nagy, D.**: Visualization of tolerance relations.
In: Proceedings of the 10th International Conference on Applied Informatics. Ed.: Terdik György, Kovácsnai Gergely, Tómács Tibor, Eszterházy Károly Egyetem, Eger, 15-22, 2018. ISBN: 9786155621727

8. **Nagy, D.**, Mihálydeák, T. S., Aszalós, L.: Similarity based rough sets.
In: Rough Sets : proceedings, part II. Eds.: Lech Polkowski, Yiyu Yao, Piotr Artiemjew, Davide Ciucci, Dun Liu, Dominik Ślęzak, Beata Zielosko, Springer, Cham, 94-107, 2017, (Lecture Notes in Computer Science, ISSN 0302-9743 ; 10314.) ISBN: 9783319608396

További közlemények

Idegen nyelvű tudományos közlemények hazai folyóiratban (1)

9. Aszalós, L., Kormos, J., **Nagy, D.**: Conjectures on phase transition at correlation clustering of random graphs.
Ann. Univ. Sci. Bp. Rolando Eötvös Nomin., Sect. Comput. 42, 37-54, 2014. ISSN: 0138-9491.

A DEENK a Jelölt által az iDEa Tudóstérbe feltöltött adatok bibliográfiai és tudománymetriai ellenőrzését a tudományos adatbázisok és a Journal Citation Reports Impact Factor lista alapján elvégezte.

Debrecen, 2020.02.21.

UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/44/2020.PL
Subject: PhD Publikációs Lista

Candidate: Dávid Nagy
Neptun ID: LF1CXV
Doctoral School: Doctoral School of Informatics
MTMT ID: 10056962

List of publications related to the dissertation

Foreign language scientific articles in Hungarian journals (1)

1. **Nagy, D.**, Aszalós, L., Mihálydeák, T. S.: Finding the representative in a cluster using correlation clustering.
   Pollack Period. 14 (1), 15-24, 2019. ISSN: 1788-1994.
   DOI: http://dx.doi.org/10.1556/606.2019.14.1.2

Foreign language conference proceedings (7)

2. **Nagy, D.**, Aszalós, L.: Approximation Based on Representatives.
   In: Rough Sets : International Joint Conference, IJCRS 2019. Eds.: Tamás Mihálydeák, Fan Min, Guoyin Wang, Mohua Banerjee, Ivo Düntsch, Zbigniew Suraj, Davide Ciucci, Springer, Cham, 91-101, 2019, (Lecture Notes in Computer Science, ISSN 0302-9743 ; 11499.) ISBN: 9783030228149

3. **Nagy, D.**, Mihálydeák, T. S., Aszalós, L.: Different Types of Search Algorithms for Rough Sets.
   Acta Cybern. 24 (1), 105-120, 2019. ISSN: 0324-721X.
   DOI: http://dx.doi.org/10.14232/actacyb.24.1.2019.8

4. Aszalós, L., **Nagy, D.**: Iterative Set Approximations Based on Tolerance Relation.
   In: Rough Sets : International Joint Conference, IJCRS 2019. Eds.: Tamás Mihálydeák, Fan Min, Guoyin Wang, Mohua Banerjee, Ivo Düntsch, Zbigniew Suraj, Davide Ciucci, Springer, Cham, 78-90, 2019, (Lecture Notes in Computer Science, ISSN 0302-9743 ; 11499.) ISBN: 9783030228149

5. Aszalós, L., **Nagy, D.**: Selecting representatives.
   In: Communication Papers of the 2019 Federated Conference on Computer Science and Information Systems. Eds.: Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki, Polskie Towarzystwo Informatyczne, Warszawa, 13-19, 2019, (Annals of Computer Science and Information Systems, ISSN 2300-5963 ; 20.) ISBN: 9788395541636

DEBRECENI EGYETEM Nemzeti Könyvtár

UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

6. **Nagy, D.**, Mihálydeák, T. S., Aszalós, L.: Similarity based rough sets with annotation.
In: Rough Sets. Ed.: Hung Son Nguyen, Quang-Thuy Ha, Tianrui Li, Małgorzata Przybyła-Kasperek, Springer International Publishing, Cham, 88-100, 2018, (Lecture Notes in Computer Science, ISSN 0302-9743 ; 11103.) ISBN: 9783319993676

7. Aszalós, L., **Nagy, D.**: Visualization of tolerance relations.
In: Proceedings of the 10th International Conference on Applied Informatics. Ed.: Terdik György, Kovácsnai Gergely, Tómács Tibor, Eszterházy Károly Egyetem, Eger, 15-22, 2018. ISBN: 9786155621727

8. **Nagy, D.**, Mihálydeák, T. S., Aszalós, L.: Similarity based rough sets.
In: Rough Sets : proceedings, part II. Eds.: Lech Polkowski, Yiyu Yao, Piotr Artiemjew, Davide Ciucci, Dun Liu, Dominik Ślęzak, Beata Zielosko, Springer, Cham, 94-107, 2017, (Lecture Notes in Computer Science, ISSN 0302-9743 ; 10314.) ISBN: 9783319608396

List of other publications

Foreign language scientific articles in Hungarian journals (1)

9. Aszalós, L., Kormos, J., **Nagy, D.**: Conjectures on phase transition at correlation clustering of random graphs.
Ann. Univ. Sci. Bp. Rolando Eötvös Nomin., Sect. Comput. 42, 37-54, 2014. ISSN: 0138-9491.

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

21 February, 2020

DEBRECENI EGYETEM
Egyetemi és Nemzeti Könyvtár