Miskolci Egyetem

Gépészmérnöki és Informatikai Kar

Sztochasztikus módszerek alkalmazása optimalizálási és adattisztítási eljárásokban

című PhD értekezés tézisei

Készítette:
Gurka Dezsőné Csizmás Edit
okleveles matematika-fizika-számítástechnika szakos középiskolai tanár

Hatvany József Informatikai Tudományok Doktori Iskola

Doktori iskola vezető:
Prof. Dr. Szigeti Jenő
egyetemi tanár

Témavezető:
Prof. Dr. Kovács László
egyetemi tanár

Társtémavezető:
Dr. Kovács Edith Alice
egyetemi docens

Miskolc, 2023.

Tartalom

1. fejezet

Bevezetés

Egy vagy több esemény bekövetkezésének környezetéről rendelkezésünkre álló információk gyakran hiányosak. Ez történhet például a hiányos adathalmazok miatt vagy például a jövőben történő események bizonytalansága miatt. A bizonytalanság kezelése fontos feladat, célunk olyan modellek kidolgozása, amelyek segítségével szimulálni tudjuk a lehetséges eseményeket, megbecsülhetjük különböző események valószínűségét. Ilyen módon olyan információkhoz juthatunk, amelyek elősegíthetik a "jó" döntések meghozatalát.

Dolgozatomban egyrészt bizonytalanság melletti optimalizálási feladatok megoldását vizsgáltam, különböző módszereket dolgoztam ki és teszteltem. Ezek érintik a kiinduló megengedett megoldások meghatározását és a pontosság szabályozását a valószínűségmaximalizálási feladatok megoldása során. Ezen kívül kidolgoztam többdimenziós eloszlások modellezését kopulákhoz kapcsolódóan és ezek felhasználásából származtatott szimulációkat és szcenáriók generálását. Legfrissebb kutatásaim a hiányzó adatok pótlására irányultak. Ezen a területen kifejlesztettem egy valószínűségi gépi tanulási eljárást, amely a releváns változók alapján becsüli meg a hiányzó értékeket. Egy másik problémakör, ahol szimulációs eljárást terveztem és alkalmaztam a prefix-fa alapú tárolás hatékonyságának költségelemzése volt a fa méretére vonatkozóan.

A dolgozatom a bevezető után négy fejezetet tartalmaz. Mindegyik fejezet tartalmazza a fejezet témájához szigorúan kapcsolódó fogalmakat és a kapcsolatos munkákat, továbbá kiemelésre kerülnek az eredmények és a tézisek.

1.1. Sztochasztikus optimalizálás

Az első fejezet a sztochasztikus optimalizáláshoz kapcsolódik, nevezetesen valószínűség maximalizálással foglalkozom.

Az alkalmazásokban gyakran fordul elő olyan feladat, ahol úgy kell optimalizálnunk, hogy bizonyos feltételek, megkötések egy meghatározott nagy valószínűséggel teljesüljenek. Ezeket a feladatokat nevezzük valószínűségi korlátos feladatoknak. Egy korai klasszikus alkalmazás szavasmarha takarmányozásra Van de Panne és Popp [10] munkája. A vízügyi feladatokban is gyakran alkalmazzák, például víztározók méretezésére és működtetésére úttörő projektekről számol be Prékopa és Szántai az [5, 6] cikkekben. További vízügyi alkalmazások Morgan és munkatársai [3] és Van Ackooij és munkatársai [9] publikációiban található. Vegyipari alkalmazásról számol be Henrion és Moller a [2] cikkben. Egy korai klasszikus energetikai alkalmazás a STABIL modell, amelyet Prékopa és

munkatársai dolgoztak ki [4]. Egy újabb energetikai alkalmazás Van Ackooij [8] cikkében található.

Az infokommunikációs technológiák fejlődésével új alkalmazási területek jelentek meg, például az intelligens hálózatok és a közlekedési rendszerek témakörében. Ezen területek gyors fejlődése szükségessé teszi olyan módszerek kidolgozását, amelyekkel valószínűségi korlátokat tudunk kezelni.

1.2. Kockázat modellezése kopulák felhasználásával

A második fejezetben az adatgenerálás és szimuláció játszik fontos szerepet. Ezek kidolgozásával portfóliók optimalizálásával, illetve portfóliók kockázatának elemzésével foglalkozom.

A 2008-as pénzügyi válság után a matematikai modellek jósága erősen megkérdőjelezhetővé vált, ezért az új alapokra való helyezést célozták meg. Kiderült hogy a ritka, nagy veszteséggel járó események, amelyek tipikusan egyszerre következnek be, pénzügyi válsághoz vezetnek. Ebből következik, hogy fontos a ritka események és az összefüggés struktúrájának modellezése. Ezt a célt tartom szem előtt ebben a fejezetben bevezetett eljárásokkal.

A fejezet a következőképpen épül fel.

Az első alfejezet a felhasznált fogalmakat tartalmazza, szigorúan leszűkítve azokra, amelyeket a szimulációs kutatásban használtam. Ebben a részben a kockázati mutatókra tett hatást is vizsgálom szimulált majd valós portfóliókra is.

A második alfejezetben bemutatom a VaR és CVaR modellezési lehetőségeit rávilágítva az összefüggési struktúra szerepére. Bemutatom a Gauss-kopula előnyeit, melyről egyébként kevés szó esik és azt, hogy hogyan hatnak a peremek az aszimmetrikus összefüggésre, noha a kopula szimmetrikus. Három-dimenziós eloszlásokat generálok, bemutatva, hogyan hat az összefüggési struktúra, ugyanazon normális peremeloszlások, és korrelációk mellett, majd rámutatok, hogyan változnak az együttes eloszlások a struktúra és a peremek megváltoztatása által, adott korreláció mellett.

A harmadik alfejezet az SSD portfólió optimalizálási problémával foglalkozik. A fejezet első részében beszélek a optimalizálási eljárásról, majd a második fejezetben ismertetem, az általam végzett szcenárió generálásokat és tárgyalom az eredményeket. A kopulákkal való szcenáriógenerálással portfólió optimalizálásban viszonylag kevesen foglalkoztak. Az SSD optimalizálás esetére pedig ez az első ilyen megközelítés.

1.3. Adatpótlás

A harmadik részben a hiányzó adatok pótlására kidolgozott speciális valószínűségekre épülő gépi tanulási eljárást mutatom be.

Az adatpótlás az adattudomány egyik kulcsfontosságú problémaköre. A hiányzó adatok megnövelik a döntéshozók bizonytalanságát, illetve csökkentik a modellek megbízhatóságát.

A valószínűségi gépi tanulás és a neurális hálózatok eszközeit használó, újonnan javasolt megközelítéseket az [S3] cikkben dolgoztuk ki. Az alábbiakban a valószínűségi gépi tanuláshoz tartozó új adatpótlási módszereimet mutatom be. A bevezetett módszerek

hatékonyságát is vizsgáltam, összehasonlítva az R programcsomagban lévő ismert adatpótlási módszerekkel. A hiányzó adatok pótlására kidolgoztam egy valószínűség alapú gépi tanulási módszert, melynek keretében attribútum csökkentési eljárásokat is kidolgoztam. A módszerek hatékonyságának összehasonlítására kidolgoztam két mutatót, az egyiket a pontosságra, a másikat a végrehajtási időre vonatkozóan. Az algoritmusokat R nyelven implementáltam. A módszerek összehasonlítására kísérleteket végeztem több, különböző tulajdonságú adathalmazon.

A fejezet felépítése a következő. Az első részben rövid áttekintést adok az adatpótláshoz kapcsolódó eredményekről és ezen belül elhelyezem a jelen kutatásomat. A második alfejezetben az adatok előkészítése kerül bemutatásra, ahol a kategorikus és folytonos attribútumokat diszkrét attribútumokká alakítom, viszonylag kis értékkészlettel. A harmadik alfejezetben bemutatom a releváns attribútum-kiválasztási módszereket, amelyek az adatpótlási módszerem alappillére. Az ezeken alapuló adatpótlási algoritmusokat a negyedik részben ismertetem. Az ötödik részben először két általánosan is használható módszert mutatok be, amely alapján össze lehet hasonlítani az adatpótlási módszerek pontossági eredményeit és a végrehajtási időket különböző attribútumok és adatkészletek esetén. Majd ismertetem az adatkészleteket, és illusztrációként bemutatok néhány adatpótlási eredményt, amelyeket az R-ben használt módszerekkel hasonlítok össze. Az utolsó szakasz következtetéseket tartalmaz.

1.4. Prefix-fa felépítés költségelemzése

A negyedik részben a prefix-fa alapú tárolás költségelemzésével foglalkozom a fa méretére vonatkozóan.

Nagy mennyiségű adat tárolása, elemzése során gyakran alkalmazott struktúrák a listák és a halmazok. A különböző feladatok megoldása során fontos, hogy a tárolást és a keresést hatékony algoritmusokkal oldjuk meg. Többféle tárolási lehetőséget szoktak alkalmazni, mint például a relációs adatbázisok, listák, kereső fák és a prefix-fák. A fastruktúra hatékony tárolást és adatkezelést biztosít a különböző adatlistákhoz vagy adatkészletekhez. A prefix-fa egy speciális fastruktúra az adatelemek rendezett listáinak tárolására, széles körben használt adatstruktúra számos alkalmazási területen. A prefix-fában a közös előtagrésszel rendelkező listák ugyanazon az útvonalon osztoznak. Mivel a prefix-fa számos adatmanipulációs algoritmusban részt vehet, a fa felépítésének költségbecslése fontos összetevője a teljes adatmanipulációs algoritmus költségfüggvényének. Jelen fejezetben a prefix-fa alapú tárolás hatékonyságának azon részével foglalkozom, amely egy véletlenszerű objektumlistákból álló bemeneti halmazra nyújt költségelemzést a fa méretére vonatkozóan. Az elemzés analitikus és szimulációs módszereket egyaránt tartalmaz, és a bemutatott fő eredmény egy gamma-eloszláson alapuló közelítő függvény.

2. fejezet

Új tudományos eredmények

2.1. Sztochasztikus optimalizálás

A 2.1. fejezetben a Fábián Csaba vezette kutatócsoport által, valószínűség maximalizálására kidolgozott epigráf-közelítő módszert mutattam be. Először a valószínűséggel megfogalmazott modellek és megoldó eljárások rövid történetét tekintettem át, majd a módszer alapjául szolgáló p-efficiens pont megközelítéseket.

Az általunk kidolgozott epigráf-közelítő módszer szemléletes bemutatását tartalmazó fejezet után részletes kifejtésre került a módszer, mégpedig a fejlesztés időbeli sorrendjében, ahogy egyre hatékonyabb és gyorsabb lett a megoldás. Ezen belül először bemutattam az oszlopgenerálási eljárást és a megoldó felépítését, amelyen az első kísérleteket végeztük. Mivel nagyon pontosan akartunk számolni, nagyon lassú volt a megoldás. A kutatás ezen részében a mesterfeladattal kapcsolatban a kezdőmegoldás keresést és a feladat inicializálását dolgoztam ki, valamint a megoldó (solver) futtató keretrendszerét és mesterfeladatot megoldó részt Matlabban implementáltam.

A következő fejezetben bemutatásra került az a javítás, hogy az új próbapontokat csak közelítőleg számítottuk ki, viszont a függvényértéket pontosan. A megoldóban elvégeztem a módosításokat. Az általam végzett kísérletek tapasztalatai azt mutatták, hogy még mindig hosszú volt a futási idő. A fejezet végén ismertettem az elméleti alátámasztását annak, hogy miért elég, ha a próbapontokat csak közelítőleg számoljuk ki.

A következő fejezetben az újabb javítás arra vonatkozott, hogy a véletlenített eljárás, azaz a gradiens becslések, nagy valószínűséggel kellő pontosságúak legyenek, ugyanakkor a függvényértékeket se számoljuk ki pontosan, hanem csak kellő pontossággal. Ami jelen esetben azt jelentette, hogy a pontosságot annak függvényében adjuk meg menet közben, hogy mennyire közelítettük meg a megoldást. A pontosság dinamikus szabályozására eljárást dolgoztam ki és kísérleteket végeztem a módszer tesztelésére. A kísérleti tapasztalatokat és a véletlenített módszer működésének elméleti alátámasztását is ismertettem.

Az utolsó fejezetben a véletlenített eljárásban alkalmazott pontosság beállítására végzett kísérleteimet részletesen mutattam be. A kísérletek eredményeképpen egy gyakorlatban jól használható megoldó eljárást sikerült kidolgozni.

1. Tézis:
A valószínűség maximalizálási feladatok megoldásához alkalmazott oszlopgenerálás első lépéseként kidolgoztam egy eljárást megoldható mesterfeladat előállítására.

Ezen felül a mesterfeladatot úgy inicializáltam megfelelő oszlopok hozzávételével, hogy a megoldó eljárás megfelelően javító irányba induljon el.

A megoldónak a mesterfeladat részét és a kísérletekhez a futtató keretrendszert implementáltam, valamint kísérleteket végeztem különböző feladatokon.

Kidolgoztam egy eljárást a Genz-kód számítási pontosságának dinamikus szabályozására. Olyan megoldást sikerült kidolgozni, amellyel megoldás közben a gradiens becslés nagy valószínűséggel kellő pontosságú lett, és a függvényérték a szükséges pontosságot nem meghaladóan, de kellően pontos lett. Az eljárás tesztelésére és a megfelelő paraméterek megtalálására kísérleteket végeztem.

A kísérletek eredményeképpen egy gyakorlatban jól használható megoldó eljárást sikerült kidolgozni.

Kapcsolódó publikációk: [S4], [S5], [S8], [S11], [S10], [S12]

2.2. Kockázat modellezése kopulák felhasználásával

A fejezet első részében bemutattam a többváltozós valószínűségi eloszlás modellezésének lehetőségeit kopulák segítségével és ezeknek hatását a kockázatra. Rávilágítottam arra, hogy a széles farokeloszlásokat nemcsak a speciális összefüggéssel rendelkező kopulák használata okozhatja, hanem az alkalmazott peremeloszlások is. Az aszimmetriát is elő lehet állítani, akár Gauss-kopula és különböző peremek segítségével.

Ha a feltételes függetlenségi reláció kihasználásával jó közelítést kapunk, akkor a marginális párokhoz még többféle kopulát használhatunk egyszerre. Valójában ez lesz az úgynevezett vine-kopulák használatának alapja a többváltozós kopulák modellezésében.

A vizsgált kopulákhoz hozzávéve a teljes vine-struktúrát kiderült, hogy az adja meg a legjobb közelítést, ennek ellenére, mivel a paraméterei száma nagyon magas, nem igazán előnyös használni, az adatokra való túltanulás miatt. Helyette a sokkal inkább az első szinten elvágott (Truncated) fa kopulát célszerű használni.

A második részben egy valós adathalmazon, amely egy három egyenlő súlyozású eszközből álló portfólióból áll, bemutattam a függőségi modellezéshez használt különböző kopulák hatásait. Továbbá megmutattam, hogy ebben az esetben a feltételes függetlenségen alapuló approximáció még jobban közelíti a valós adatokat.

Egy másik valós adathalmazon, amely hat egyenlő súlyozású eszközből álló portfólióból áll, bemutattam a függőségi modellezéshez használt különböző kopulák hatásait, a teljes vine-kopulával való szimulációval kiegészítve.

Bemutattam, hogy a kopulákkal modellezett összefüggések különbözőképpen hatnak a kockázati mértékekre. Különböző kísérleteket végeztem a függőség különböző aspektusainak illusztrálására.

A harmadik részben pedig összehasonlítunk két SSD dominanciára épülő portfólió optimalizálási módszert, azzal hogy szcenáriókat generáltam egy tanuló mintából kiindulva.

Az alapcél egy olyan portfóliót találni, aminek hozam eloszlása minél jobban dominálja az index hozam eloszlását.

Két optimalizálási eljárást alkalmaztunk, skálázatlan [7] és skálázott [1], valamint ezek eredményeit tárgyaltuk in sample és out of sample, továbbá összehasonlítottuk a tanulóhalmaz generálás nélküli adatokból számolt optimális portfólióval.

A legfontosabb eredmény, hogy az általam generált minta alapján meghatározott portfóliók robusztusak. A nemskálázott eljárás többnyire megtartja a dominanciáját out of sample is. A skálázott modell esetében sok feltételt teljesít, de van néhány, amit nem, amelyek az extremitásokban vannak. Miközben a skálázatlan modell védekezik az extrém vesztességektől, a skálázott modell jobban leköveti az index hozam eloszlását. A szcenáriógenerálás hozzájárult a dominancia megtartásához az out of sample esetben. A skálázott eljárást kifejezetten támogatta a szcenáriógenerálás.

2. Tézis:

Szimulációt terveztem többdimenziós eloszlások modellezési lehetőségeire, különböző kopulákat felhasználva. Tárgyaltam a Gauss-kopula esetét, és megmutattam, hogy többféle kapcsolatot tud egyszerre modellezni, továbbá alkalmas aszimmetrikus eloszlások modellezésére is, különböző farokösszefüggésekkel. Ezekre a szakirodalomban nem térnek ki. Bemutattam 2- és 3-dimenziós szimulált kopulák felhasználásával ugyanazon normális peremek mellett a Student-, Clayton-, Gumbel-, Frank-kopula hatásait, majd bemutattam, hogy különböző peremek milyen hatást váltanak ki, ami az aszimmetria és a farokösszefüggéségeket illeti.

Kapcsolódó publikációk: [S6], [S2], [S18]

3. Tézis:

Mivel a kockázati mutatók a ritka eseményekhez kötődnek, ezeket farokösszefüggések jellemezik. Szimulációt terveztem és bemutattam a szimulált adatokon, hogyan befolyásolja az összefüggéseket leíró kopula a kockázati mutatókat. Az is igazolódott, hogy a közel-függetlenség csökkenti a portfólió kockázatát. Valós 3, majd 6 értékpapírból álló egyenletes portfólió esetén, kétirányú kutatást végeztem. Az egyik arra vonatkozott, hogy milyen típusú kopula illeszkedik legjobban az adatokra, a másik pedig a kockázati mutatókra való hatására vonatkozott. A kísérleteket elvégezve arra jutottam, hogy a sokparaméterű kopulák illeszkednek jobban, vagyis a vine- és a Gauss-kopula. Ugyanakkor jó eredményeket hozott a feltételes függetlenségeket tartalmazó (CI) kopula közelítés is, amely sokkal kevesebb paraméterrel rendelkezik, mint a vine-kopula. A portfólióból jól látszanak a Gumbel-, illetve a Clayton-kopula specifikus farokösszefüggései. Rámutattam, hogy kis méretű portfóliók esetén, a kockázat modellezésében komoly szerepe van a kopulatípusnak, vagyis az összefüggés modellezésnek, továbbá a generált mintanagyságnak is.

Kapcsolódó publikációk: [S6], [S2], [S18]

4. Tézis:
Portfólió optimalizáláshoz készítettem szcenáriógenerálási eljárást és algoritmust, amely segítségével több adatot tudtam szolgáltatni a két portfólió optimalizálási eljárásnak. A szimulációs eljárást Matlabban implementáltam. Az eredmények biztatóak, mivel a tanuló mintából tanult és szimulált adatok a Gauss-kopula és lognormális peremeloszlások segítségével elősegítettem az optimális portfólió megválasztását. A skálázott és skálázatlan esetben is a portfólió várhatóértéke nőtt és a szórás csökkent. A különbség az, hogy a skálázott esetben mintha a hozamot szemléltető hisztogramot jobbra toltuk volna, azaz minden egyenletesen jobbra tolódott. A skálázatlan esetben pedig mintha levágtuk volna a "bal farkát". Ezért a skálázatlan megoldás olyan befektetőknek jó, akik félnek az extrém kis valószínűséggel előforduló extrém nagy veszteségtől, míg a skálázott azoknak jó, akik inkább az általános befektetői hozzáállásnak felelnek meg.

Kapcsolódó publikációk: [S1], [S9], [S13], [S14], [S19], [S15], [S16], [S17], [S21]

2.3. Adatpótlás

Ebben a fejezetben attribútum kiválasztáson alapuló valószínűségi gépi tanulási adatpótlási módszereket mutattam be. A módszereket diszkrét adatokra fejlesztettem ki, de kategorikus és folytonos adatokra is alkalmazhatók. A folytonos adatokat az előfeldolgozási fázisban diszkretizálom. Ha az imputált változó folytonos, eredményként egy intervallumot kapunk, amelyhez az imputált érték tartozik (egy adott szakterület szakértői számára ez néha nagyon hasznos információ lehet), vagy az imputált érték a kapott intervallumhoz tartozó értékek mediánja vagy átlaga. Az információelméleti attribútum kiválasztáson alapuló módszerek kategorikus változókra is alkalmazhatók, mivel csak a valószínűségi eloszlástól függnek, nem pedig a felvett értékektől.

A fejezetben különböző adatadatpótlási módszerek hatékonysági elemzését mutattam be, az egyszeri adatpótlási megközelítésre összpontosítva. A bevezetett módszereimet összehasonlítottam négy, az R programcsomagokban megvalósított módszerrel.

A különböző módszerek különböző adathalmazokon történő összehasonlítása érdekében bevezettem két mutatót. Az egyik az általános relatív pontosság, amely a hiányzó adatok különböző módszerekkel való pótlásának pontosságát összehasonlíthatóvá teszi különböző adathalmazokon. A másik az általános relatív végrehajtási idő, amely a hiányzó adatok különböző módszerekkel való pótlásának végrehajtási idejét teszi összehasonlíthatóvá különböző adathalmazokon.

Az eljárást implementáltam.

Az elvégzett összehasonlító tesztek fő következtetése, hogy a javasolt módszereim minden kategóriában jobbnak bizonyultak. A valószínűségi gépi tanulás esetében az információtartalom-alapú módszereim bizonyultak a legjobbnak, de fontos megjegyezni, hogy a hagyományos korrelációs együtthatókon alapuló jellemzőválasztás is hatékony volt. Az időköltséget tekintve a valószínűségi gépi tanulási módszereim jellemzően gyorsabbak voltak.

Az elvégzett elemzések és tesztek azt mutatják, hogy a javasolt adatpótlási valószínűségi gépi tanulási technikák egyértelműen előnyösebbek a szokásos gépi tanulási technikáknál.

5. Tézis:

Attribútum kiválasztáson alapuló valószínűségi gépi tanulási adatpótlási módszereket vezettem be. A módszereket diszkrét adatokra fejlesztettem ki, de kategorikus és folytonos adatokra is alkalmazhatóak.

A fejezetben különböző adatadatpótlási módszerek hatékonysági elemzését mutattam be, az egyszeri adatpótlási megközelítésre összpontosítva. A bevezetett módszereimet összehasonlítottam négy, az R programcsomagokban megvalósított módszerrel.

A különböző módszerek különböző adathalmazokon történő összehasonlítása érdekében bevezettem két mutatót. Az egyik a pontosságot, a másik a végrehajtási időt jellemzi.

A valószínűségi gépi tanulás esetében az információtartalom-alapú attribútumcsökkentési módszereim bizonyultak a legjobbnak, de a hagyományos korrelációs együtthatókon alapuló jellemzőválasztás is hatékony volt. Az időköltséget tekintve a valószínűségi gépi tanulási módszereim jellemzően gyorsabbak voltak.

Az elvégzett elemzések és tesztek azt mutatják, hogy a javasolt adatpótlási valószínűségi gépi tanulási technikák egyértelműen előnyösebbek a vizsgált beépített gépi tanulási technikáknál.

Kapcsolódó publikációk: [**S3**]

2.4. Prefix-fa felépítés költségelemzése

Az elvégzett elemzésben analitikus és kísérleti úton vizsgáltam a prefix-fa méretfüggvényét a bemeneti adatok méret- és értékeloszlásától való függésben. A vizsgálati eredmények azt mutatják, hogy a szimuláción alapuló kísérletek jó közelítést adnak az analitikus költségfüggvényekhez. A kísérleti költségfüggvények normalizálása után megállapítottam, hogy a normalizált költségfüggvények hatékonyan közelíthetők gammaeloszlással. A kutatási projekt következő fázisában az elemzés célja egy hatékony módszer kidolgozása a legjobban illeszkedő gammaeloszlás skála- és alakparamétereinek kiszámítására.

6. Tézis:

Adott alaphalmazhoz tartozó prefix-fa költségbecslésére kidolgoztam egy hatékony szimuláción alapuló közelítő eljárást. Az eljárás bemenő adathalmaz alapparamétereihez (elemszám, listák száma, elemek előfordulásának valószínűségei) regressziós paraméterbecslés mellett meghatároztam a prefix-fa várható költségét. Az elvégzett teszt vizsgálatok alapján a módszer igen pontos (4-5 %-os pontosság) eredményt szolgáltat egy alacsony számítási költség mellett.

Kapcsolódó publikációk: [**S7**], [**S20**]

3. fejezet

Összefoglalás

Dolgozatomban olyan témákkal foglalkoztam, amelyek valamilyen módon a bizonytalanság kezeléséhez kapcsolódnak. Új eljárásokat dolgoztam ki, szimulációkat terveztem és implementáltam, valamint ezeken kísérleteket futtattam. Négy fejezetben mutattam be a kutatásaimat.

Az első fejezetben ismertetett kutatás során, a Fábián Csaba vezette kutatócsoportban, valószínűség maximalizálási feladat epigráf-közelítő módszerrel való megoldását valósítottuk meg. Ennek egyik részfeladataként a mesterfeladat kezdőmegoldásának megkeresésére és inicializálására dolgoztam ki eljárásokat, valamint a módszer lényegét képező véletlenített eljárás keretében a pontosság szabályozására dolgoztam ki egy módszert. Az eljárásokat Matlabban implementáltam, és kísérleteket végeztem az eljárások tesztelésére.

A második fejezetben portfólió optimalizálással, illetve kockázatelemzéssel foglalkoztam. Adatgenerálás és szimuláció, illetve valós pénzügyi adatok segítségével megmutattam, hogy kopulákkal jól modellezhetők a portfóliók VaR és CVaR kockázati mutatói, de a különböző kopulák különböző módon befolyásolják ezeket a mutatókat. A szimulációkat R nyelven implementáltam, és különböző paraméterekkel kísérleteket végeztem. Egy másik kutatás keretében valós adatokhoz készítettem szcenáriógenerálási eljárást és algoritmust, amely segítségével több adatot tudtam szolgáltatni kétféle − skálázott és skálázatlan − portfólió optimalizálási eljárásnak, ennek segítségével a portfóliók robusztusabbá váltak. A Gauss-kopulával való adatgenerálást megvalósító algoritmust és a szimulációt végző programokat Matlabban implementáltam.

A harmadik részben a hiányzó adatok pótlására kidolgoztam egy speciális valószínűségekre épülő gépi tanulási eljárást. Bemutattam az eljárás részét képező attribútum kiválasztáshoz használt módszereket. Az új módszert összehasonlítottam ismert módszerekkel, különböző arányú hiányzásokkal, különböző adathalmazokon. A módszerek összehasonlíthatósága érdekében bevezettem két mutatót, az egyiket a pontosság, másikat a végrehajtási idő összehasonlítására. Az algoritmust és a szimulációkat megvalósító programot R nyelven implementáltam.

A negyedik részben a prefix-fa alapú tárolás költségelemzésével foglalkoztam a fa méretére vonatkozóan. Kidolgoztam egy szimulációs eljárást a költség becslésére, majd az eredményeket felhasználva gamma függvénnyel közelítettem a költségfüggvényt. A szimulációt végző programot és az illesztést Matlabban implementáltam.

Summary

The information we have about the environment is often incomplete. This may be due to incomplete data sets or because of the uncertainty of future events. This is why dealing with uncertainty is an important task, and our goal is to develop models that allow us to simulate possible realizations and estimate the probability of different events. This way, we can obtain information that can help us make the "right" decisions. This motivation led us to different fields like probability theory, data science, machine learning, stochastic optimization, and programming. The topics of the present dissertation are related to uncertainty management. I developed new procedures, designed simulations, and ran experiments in order to deal with the problem of uncertainty, I also implemented codes in order to apply them. The research is presented in four chapters.

In the first chapter, I gave an introduction to the probability maximization problem using an epigraph approximation, a method that was developed and implemented by the research team led by Csaba Fábián. As a sub-task of this work, I developed procedures for finding and initializing the starting solution of the master problem and a method for controlling the accuracy in the context of the randomized procedure which is the essence of the method. I implemented the procedures in Matlab and performed experiments to test the procedures.

In the second chapter, I dealt with portfolio optimization and risk analysis. Using data generation and simulation and real financial data, I showed how copulas can be used in modeling and their effects on the VaR and CVaR of portfolios. I showed how different copulas affect these indicators. I implemented the simulation procedures in R and performed experiments with different parameters. As further research, I created a scenario generation procedure and algorithm for real data, which allowed me to provide more input data to two different − scaled and unscaled − portfolio optimization procedures, with the help of which the portfolios became more robust. I implemented the Gaussian copula data generation algorithm and simulation programs in Matlab.

In the third part, I developed a machine learning procedure based on feature selection to impute the missing data with a value with high probability. I presented methods used for attribute selection as part of the procedure. Then I compared the new method with known methods with different missingness rates on different data sets. To make the methods comparable, I introduced two metrics, one to compare the accuracy and the other to compare execution time. I implemented the imputation algorithms and the comparison methods in R.

In the fourth chapter, I dealt with the cost analysis regarding the tree size of prefix-tree-based storage. I developed a simulation procedure to estimate the cost and then used the results to approximate the cost function with a gamma function. I implemented the simulation procedure and the model fitting in Matlab.

Saját publikációk

[**1. Könyvfejezet**]

[*angolul*]

[S1] C.I. Fábián, G. Mitra, D. Roman, V. Zverovich, T. Vajnai, E. Csizmás, O. Papp. Portfolio choice models based on Second-order Stochastic Dominance measures: An overview and a computational study. In: M. Bertocchi, G.Consigli, M.A. H. Dempster (szerk.). *Stochastic Optimization Methods in Finance and Energy: New Financial Products and Energy Market Strategies*. New York (NY), Amerikai Egyesült Államok: Springer-Verlag London Ltd, pp. 441-469., 2011. DOI: 10.1007/978-1-4419-9586-5_18 **SCOPUS Független idéző: 6**

[*magyarul*]

[S2] Csizmás Edit, Kovács Edith: Kopulák a kockázatszámításban. In: Johanyák Zsolt Csaba, Kovács Lóránt, Pásztor Attila, Ferenczy Tibor, Weltsch Zoltán, Tóth Ákos, Dobjánné Antal Elvira (szerk.). *Kutatás és innováció 2021 : GAMF Közlemények tanulmánykötete*. Kecskemét, Magyarország. pp. 383-388. 2021.

[**2. Folyóirat cikk**]

[*angolul*]

[S3] Edit Csizmás, Edith Alice Kovács, László Kovács. Data Imputation Methods based on feature selection. **Under Review**, in: *Knowledge-Based Systems*, 2023.

[S4] C.I. Fábián, E. Csizmás, R. Drenyovszki, T. Vajnai, L. Kovács and T. Szántai. A randomized method for handling a difficult function in a convex optimization problem, motivated by probabilistic programming. *Annals of Operations Research*, To appear in S.I.: Stochastic Modeling and Optimization, in memory of András Prékopa (editors: E. Boros, M. Katehakis, A. Ruszczyński). pp. 1-32. 2019. DOI: 10.1007/s10479-019-03143-z. **SCOPUS Q1 IF=4.820 Független idéző: 2**

[S5] Csaba I Fábián, Edit Csizmás, Rajmund Drenyovszki, Wim van Ackooij, Tibor Vajnai, Lóránt Kovács and Tamás Szántai. Probability maximization by inner approximation. *Acta Polytechnica Hungarica*, 15(1):105–125, 2018. DOI: 10.12700/APH.15.1.2018.1.7 **SCOPUS Q2 IF=1,711 Független idéző: 5**

[*angolul hazai*]

[S6] Edit Csizmás, Edith Kovács. The effect of the dependence structure on risk measures. *GRADUS* 8 : 3 pp. 157-171. 2021. DOI: 10.47833/2021.3.CSC.004

[S7] Csizmás Edit, Kovács László. Cost analysis of the prefix tree data structure. *PRODUCTION SYSTEMS AND INFORMATION ENGINEERING*. 8 pp. 39-49. 2019. DOI: 10.32968/psaie.2019.003

[*magyarul*]

[S8] Edit Csizmás, Rajmund Drenyovszki, Tibor Vajnai, Lóránt Kovács and Csaba Fábián. Valószínűség maximalizálás. *GRADUS*, 5(1):128–133, 2018.

[S9] Vajnai Tibor, Csizmás Edit, Fábián Csaba. Véletlen vektorparaméterek szimulációja. *A GAMF KÖZLEMÉNYEI* 23, pp. 131-140. 2009.

[**3. Konferencia**]

[*angolul*]

[S10] Rajmund Drenyovszki, Edit Csizmás, Csaba Fabian, Lorant Kovacs, Tamas Szantai. A probabilistic formulation of a demand-side management problem, and its solution with a randomized scheme. *VOCAL Optimization Conference : Advanced Algorithms*. Budapest, 2022. május 25-27., Előadás

[S11] E. Csizmás, T. Vajnai, C.I. Fábián. Experiments with randomized method for probability maximization. In: Bojan, Lalić (szerk.) *Proceedings of TEAM 2018 : 9th International Scientific and Expert Conference: 10-12th October 2018*, pp. 237-240. 2018.

[S12] Fábián Csaba, Csizmás Edit, Drenyovszki Rajmund, Vajnai Tibor, Kovács Lóránt, Szántai Tamás. A randomized method for probabilistic problems. *AGTEDU 2018*. Kecskemét : 2018. november 15., Előadás

[S13] Csaba Fabian, Edit Csizmas, Tibor Vajnai. Modeling uncertainty for stochastic optimisation. In: Andrea Ádámné Major, Lóránt Kovács, Zsolt Csaba Johanyák, Róbert Pap-Szigeti (szerk.): *Proceedings of TEAM 2014 : 6th International Scientific and Expert Conference of the International TEAM Society*. Kecskemét, pp. 174-178. 2014.

[S14] Tibor Vajnai, Olga Papp, Edit Csizmás, Csaba I. Fábián. Using scenario generation for decision making under uncertainty. In: Dominika Lehocká, Ján Cárach, Lucia Knapčíková, Sergej Hloch (szerk.): *Proceedings of the 5th International Scientific and Expert Conference of the International TEAM Society : (Technique, Education, Agriculture & Management)*. Presov, Szlovákia : International TEAM Society Press, pp. 150-153. 2013.

[S15] O. Papp, E. Csizmás, T. Vajnai, C.I. Fábián, G. Mitra, D. Roman, V. Zverovich. A comparison of copula-based scenario generation methods. In: L. Suhl, G. Mitra, C. Lucas, A. Koberstein, L. Beckmann (szerk.). *Applied Mathematical Optimization and Modelling : Extended Abstracts of the APMOD 2012 Conference*. Paderborn, Németország : Universität-Gesamthochschule Padernborn, pp. 89-94. 2012.

[S16] Olga Papp, Edit Csizmás, Csaba I. Fábián, Tibor Vajnai. A comparison of scenario generation methods with risk-averse decisions. In: Marija Živić, Tomislav Galeta (szerk.). *TEAM 2012 : Proceedings of the 4th International Scientific and Expert Conference : Mechanical Engineering Faculty*. pp. 223-226. 2012.

[S17] Csaba Fabian, Tibor Vajnai, Gautam Mitra, Diana Roman, Olga Papp, Edit Csizmas. A computational study on the utility of scenario generation methods in portfolio optimisation. *Veszprém Optimization Conference : Advanced Algorithms VOCAL 2010*. Veszprém, 2010., Előadás

[*magyarul*]

[S18] Csizmás Edit, Kovács Edith. A hozamok összefüggésének hatása a kockázati mutatókra. In: Sziklai R. Balázs (szerk.) *XXXIV. Magyar Operációkutatási Konferencia: Absztraktok könyve*. Budapest, Magyarország : Gazdaságmodellezési Társaság. pp. 44. 2021., Előadás

[S19] Csizmás Edit, Vajnai Tibor. Kopula függvények használata véletlenek szimulálására. In: Dobjánné Antal Elvira, Nagy Péter (szerk.). *Matematikát, Fizikát és Informatikát Oktatók 42. Országos Konferenciája MAFIOK 2018 : Konferencia absztraktok* Kecskemét, Magyarország. pp. 26. 2018., Előadás

[S20] Csizmás Edit, Kovács László. Prefix-fa felépítés költségelemzése Monte-Carlo szimulációval. In: Kovács László, Piller Imre (szerk.). *Doktoranduszok Fóruma 2016 : Gépészmérnöki és Informatikai Kar szekciókiadványa*. Miskolci Egyetem, pp. 37-42. 2017.

[S21] Papp Olga, Csizmás Edit, Fábián Csaba, Vajnai Tibor. Kísérletek szcenárió generálási eljárásokkal. In: Ferencz, Árpád; Borsné, Pető Judit; Lipócziné, Csabai Sarolta; Kovács, Lóránt (szerk.) AGTEDU 2011 : a Magyar Tudomány Ünnepe alkalmából rendezett 12. tudományos konferencia pp. 369–374. 2011.

Irodalomjegyzék

[1] Csaba I Fábián, Gautam Mitra, Diana Roman, and Victor Zverovich. An enhanced model for portfolio choice with ssd criteria: a constructive approach. *Quantitative Finance*, 11(10):1525–1534, 2011.

[2] R. Henrion and A. Möller. Optimization of a continuous distillation process under random inflow rate. *Computer & Mathematics with Applications*, 45:247–262, 2003.

[3] D.R. Morgan, J.W. Eheart, and A.J. Valocchi. Aquifer remediation design under uncertainty using a new chance constraint programming technique. *Water Resources Research*, 29:551–561, 1993.

[4] A. Prékopa, S. Ganczer, I. Deák, and K. Patyi. The STABIL stochastic programming model and its experimental application to the electrical energy sector of the Hungarian economy. In M.A.H. Dempster, editor, *Stochastic Programming*, pages 369–385. Academic Press, London, 1980.

[5] A. Prékopa and T. Szántai. Flood control reservoir system design using stochastic programming. *Math. Programming Study*, 9:138–151, 1978.

[6] A. Prékopa and T. Szántai. On optimal regulation of a storage level with application to the water level regulation of a lake. *European Journal of Operations Research*, 3:175–189, 1979.

[7] Diana Roman, Ken Darby-Dowman, and Gautam Mitra. Portfolio construction based on stochastic dominance and target return distributions. *Mathematical Programming*, 108:541–569, 2006.

[8] W. van Ackooij. Decomposition approaches for block-structured chance-constrained programs with application to hydro-thermal unit commitment. *Mathematical Methods of Operations Research*, 80:227–253, 2014.

[9] W. van Ackooij, R. Henrion, A. Möller, and R. Zorgati. Joint chance constrained programming for hydro reservoir management. *Optimization and Engineering*, 15:509–531, 2014.

[10] C. van de Panne and W. Popp. Minimum-cost cattle feed under probabilistic protein constraints. *Managment Science*, 9:405–430, 1963.