R-alapú algoritmikus keretrendszer változótípus-felismerésre és asszociáció-elemzésre az EDA fázisban

Horváth Ádám Béla

horvath.adam@kgk.uni-obuda.hu

Absztrakt: A feltáró adatelemzés (Exploratory Data Analysis, EDA) a modern adatbányászati és adattudományi folyamatok kulcsszakasza, amelynek során a kutató nem pusztán tisztítja és leírja az adatokat, hanem azonosítja a változók közötti lehetséges összefüggéseket is. A gyakorlatban azonban két, egymást erősítő probléma jelenik meg: (1) a változók típusainak (kategorikus, diszkrét, folytonos) gyakran csak szemantikus, nem pedig szintaktikus módon megkülönböztethető jellege, valamint (2) a nagydimenziós adatbázisokban előálló kombinatorikus robbanás, amikor több száz változó között kell asszociációkat szűrni. A tanulmány egy egységes, R-nyelven megvalósított algoritmikus keretrendszert mutat be, amely két fő funkciót integrál: (i) egy vektor-szintű modul automatikusan osztályozza a változótípusokat heurisztikus küszöbértékek alapján, majd a típusoknak megfelelő statisztikai eljárást választ (χ^2-próba, Pearson-korreláció, Kendall-tau, Mann–Whitney-teszt, Kruskal–Wallis-próba); (ii) egy adatkeret-szintű modul nagyszámú változópárra hajt végre asszociáció-elemzést kizárólag base R függvényekkel, és csak a szignifikáns kapcsolatokat tartalmazó, egységes struktúrájú táblázatot ad vissza. A keretrendszer a CRISP-DM módszertani keret „Data Understanding" fázisához illeszkedik, és célja a reprodukálható, skálázható és könnyen automatizálható EDA támogatása.

Kulcsszavak: feltáró adatelemzés, EDA, változótípus-osztályozás, korreláció, R programozási nyelv, CRISP-DM

Bevezetés

A feltáró adatelemzés (EDA) klasszikus felfogásban a „józan ész statisztikai megtestesülése": olyan eszköztár, amelynek segítségével a kutató a puszta számhalmazt értelmezhető mintázatokká, hipotézisekké és későbbi modellezési lépések kiindulópontjává alakítja. A CRISP-DM folyamatmodellben a Data Understanding fázis hivatott ezt a szerepet betölteni: itt történik az adatok üzleti és módszertani kontextusának feltárása, a leíró statisztikák előállítása, valamint az első asszociációk azonosítása a változók között.

A gyakorlati adattudományi projektekben – különösen nagymintás, több száz attribútumot tartalmazó kérdőíves vagy tranzakciós adatbázisok esetén – két gyakori nehézség jelentkezik:

1. Változótípus-azonosítás. Szintaktikailag az egész számokat tartalmazó vektor lehet Likert-skálás, valóban diszkrét numerikus, vagy aggregált folytonos változó is; a statisztikai eljárás megválasztása mégis alapvetően ettől függ.

2. Nagyszámú páros összehasonlítás. Ha n változóval dolgozunk, a páronkénti összehasonlítások száma C_n^2 = n(n-1)/2, ami már n ≈ 100 mellett is több ezer potenciális kapcsolatot jelent. Ezek kézi áttekintése, tesztelése és értékelése reálisan nem kivitelezhető.

A jelen tanulmány célja egy olyan egységes R-alapú keretrendszer bemutatása, amely

- automatikusan azonosítja a változótípusokat heurisztikus szabályok alapján,

- a típusoknak megfelelő eljárást választ a lehetséges statisztikai tesztek közül,

- egységes szerkezetben adja vissza az eredményeket,

- és nagy dimenziószám esetén képes a kizárólag szignifikáns asszociációk „szűrt” táblázatos előállítására.

A keretrendszer kizárólag a base R csomag függvényeire épül, így elkerüli a külső csomagfüggőségek miatti reprodukálhatósági problémákat.

Elméleti háttér

A. EDA és a CRISP-DM „Data Understanding” fázisa

Az EDA hagyományosan olyan lépésekből áll, mint az eloszlások, középértékek, szóródási mutatók becslése, outlierek azonosítása, valamint a változók közötti kapcsolatokat célzó elsődleges vizsgálatok. A CRISP-DM modell ezt a folyamatot explicit módon beágyazza egy projektciklusba, hangsúlyozva, hogy a későbbi modellezési fázisok megbízhatósága alapvetően a Data Understanding fázis minőségétől függ. Az adatminőség és az információs rendszerek teljesítményének megbízható értékelése alapvető feltétele a Data Understanding fázis sikerének, ahogy azt Szikora (2009) is hangsúlyozza. A kvantitatív dimenzióban kiemelt szerepe van a változótípusok azonosításának és a megfelelő statisztikai eljárások hozzárendelésének. Téves eljárásválasztás esetén a szoftver ugyan eredményt szolgáltat, de az interpretáció félrevezető lehet.

B. Változótípusok és heurisztikus osztályozás

A statisztikai szakirodalomban a változókat gyakran a kategorikus–numerikus tengely mentén, azon belül pedig nominalitás, ordinálitás, diszkrétség és folytonosság szerint különítik el. A gyakorlatban azonban ez a felosztás többnyire szemantikus, míg a szoftverek szintaktikus információkkal (pl. egész szám, karakter, double) dolgoznak.

A bemutatott keretrendszer a következő, egyszerűen implementálható heurisztikákat alkalmazza:

1. Kis kategóriaszámú vektor (kategorikus):

- ha az egyedi értékek száma 2 → bináris kategorikus változó;

- ha az egyedi értékek száma 3–10 → többkategóriás kategorikus változó.

E küszöbérték a szakirodalomban javasolt, 10 körüli határértékhez igazodik.

2. Diszkrét vs. folytonos numerikus:

- kiszámítjuk az egyedi értékek és a vektorméret arányát;

- ha ez az arány $< 0{,}05$ → diszkrét;

- ha $\geq 0{,}05$ → folytonos.

Ez a megoldás különösen hasznos olyan esetekben, amikor a folytonos mennyiségek aggregált, egész szám formában kerülnek rögzítésre (pl. ezer forintban mért pénzügyi adatok).

A fenti szabályok végrehajtását a keretrendszerben egy v_cls() nevű segédfüggvény végzi, amely egy kóddal (1–4) és egy verbális leírással jelöli a változótípus kategóriáját.

C. A lehetséges statisztikai eljárások áttekintése

A keretrendszer első verziója tudatosan egy reprezentatív eljárást rendel minden típus-kombinációhoz, elkerülve a túlzottan összetett M:N megfeleltetéseket:

- Folytonos × Folytonos:

Pearson-féle korreláció, a lineáris kapcsolat klasszikus paraméteres mutatója.

- Kategorikus × Kategorikus:

χ^2-próba, amely a két változó közötti függetlenséget vizsgálja kontingenciatáblában.

- Folytonos × Diszkrét:

Kendall-féle rangkorreláció (τ), amely kevésbé érzékeny a rangszintek „lépcsőzetes” szerkezetére.

- Kategorikus × Folytonos/Diszkrét:

- két kategória esetén: Mann–Whitney-féle U-próba (Wilcoxon rangösszeg-próba),

- három vagy több kategória esetén: Kruskal–Wallis-próba, a egyutas ANOVA nemparaméteres analógja.

A konkrét tesztfüggvények – cor.test(), chisq.test(), wilcox.test(), kruskal.test() – eltérő struktúrájú objektumokat adnak vissza (eltérő számú output-elemmel), ami egységes riportolás esetén további absztrakciót igényel.

Kapcsolódó munkák és meglévő R-eszközök

A statisztikai környezetek – különösen az R programozási nyelv – gazdag eszköztára már ma is számos lehetőséget kínál asszociáció-elemzésre, azonban ezek többnyire széttöredezett, részleges megoldásokat adnak.

A cor.test() függvény részletes kimenetet (becslés, tesztstatisztika, szignifikanciaszint, konfidenciaintervallum) biztosít, és több módszert (Pearson, Spearman, Kendall) támogat, ugyanakkor egyidejűleg csak egy változópárra alkalmazható, ezért nagydimenziós adatbázisoknál rendkívül nehézkes a használata.

A cor() függvény vektorosított módon, tömör korrelációs mátrixot ad, de nem tartalmaz tesztstatisztikákat és p-értékeket, így a szignifikancia-szűrés kézzel vagy további csomagokkal oldható meg.

A Hmisc csomag rcorr() függvénye, illetve az easystats csomag correlation() eljárása már képes mintanagyságot és szignifikanciaszinteket is riportálni, mindazonáltal külső csomagfüggőségeket vezetnek be, és továbbra sem adnak egységes, típusfüggetlen kimeneti sémát.

A szakirodalom szintén rámutat a spurius korrelációk és a túl sok összehasonlítás problémájára: ha nagyszámú változópárt vizsgálunk, a véletlenszerűen szignifikáns találatok aránya elkerülhetetlenül nő, így elengedhetetlen a szisztematikus szűrés és a módszertanilag indokolt eljárásválasztás.

A bemutatott keretrendszer ezekre a korlátokra reagál:

- egységesíti a kimeneti struktúrát,

- minimalizálja a csomagfüggőségeket (base R),

- támogatja a nagy sebességű, tömeges asszociáció-elemzést,

- és a változótípus-felismerést integrálja a tesztválasztás logikájába.

A fejlesztett keretrendszer felépítése

A keretrendszer két, egymással összekapcsolódó modulból áll:

1. Vektor-szintű modul: két tetszőleges vektor (változó) között választ és hajt végre statisztikai próbát a típusok alapján, egységes listastruktúrában visszaadva az eredményt.

2. Adatkeret-szintű batch-modul: egy adatkeret meghatározott oszloptartományaiban minden releváns változópárt tesztel, és csak a szignifikáns eredményeket gyűjti egy „tiszta” táblázatba.

A. Előzetes ellenőrzés: v_test()

Az első lépés egy alapvető minőségellenőrzés:

- a két vektor hossza megegyezik,

- egyikben sincs NA,

- legalább két elem áll rendelkezésre.

A v_test() függvény logikai (TRUE/FALSE) értéket ad vissza. Sikertelen ellenőrzés esetén a keretrendszer azonnal leáll a vizsgált párra, egyértelmű hibaüzenetet ad, és így elkerüli a nehezen dekódolható, részben hibás kimeneteket.

B. Változótípus-felismerés: v_cls()

A v_cls() függvény a II. fejezetben ismertetett heurisztikák szerint:

- megszámolja az egyedi értékeket,

- meghatározza az egyedi értékek és a vektorméret arányát,

- ennek alapján kódol:

- 1: bináris kategorikus,

- 2: többkategóriás kategorikus (3–10 kategória),

- 3: diszkrét numerikus,

- 4: folytonos numerikus.

A függvény egy listát ad vissza:

- egy kódot (amely if-struktúrákban egyszerűen használható),

- egy szöveges leírást, amely a későbbi riportolásban szerepel.

C. Statisztikai eljárás választása és egységes eredménystruktúra: analyse()

Az analyse() függvény a keretrendszer „magja”:

1. Meghívja a v_test() függvényt; hiba esetén megszakítja a futást.

2. Meghatározza a két bemeneti vektor típusát a v_cls() segítségével.

3. A típusok kombinációja alapján kiválasztja a megfelelő eljárást:

- χ^2-próba (kategorikus × kategorikus),

- Pearson-korreláció (folytonos × folytonos),

- Kendall-tau (folytonos × diszkrét),

- Mann–Whitney vagy Kruskal–Wallis (kategorikus × numerikus).

4. A kiválasztott R-függvény visszatérési objektumából egységes mezőnevek szerint épít fel egy listát, pl.:

- variable_1, type_variable_1,

- variable_2, type_variable_2,

- method (eljárás neve),

- estimated_value / estimated_calc (pl. korrelációs együttható vagy W-statisztika),

- statistic_value / statistic_calc (pl. t-, χ^2-, z-érték),

- df (szabadságfok, ha értelmezhető),

- p_value.

Az így képzett struktúra egyrészt felhasználóbarát, másrészt egyszerűen programozható (pl. data.frame-be rendezés, rangsorolás p-érték vagy hatásnagyság szerint).

D. Batch-feldolgozás adatkeret szinten

A keretrendszer második modulja egy adatkeret-szintű függvény, amelynek főbb paraméterei:

- az elemzendő adatkeret,

- az első és második változóhalmaz oszlopindex-tartománya,

- a szignifikanciaszint (pl. $\alpha = 0,05$),

- opcionálisan a hiányzó értékek kezelési módja (pl. complete.cases()).

A modul

1. Bejárja a megadott oszloptartományok összes lehetséges párosítását két egymásba ágyazott ciklus segítségével, elkerülve az ön-összehasonlításokat és – kiterjesztés esetén – a duplikált párokat (A,B) vs. (B,A).

2. Minden párra meghívja az analyse() függvényt, majd ellenőrzi, hogy a kapott p-érték ≤ α-e.

3. Az első szignifikáns eredmény esetén létrehoz egy adatkeretet a standardizált mezőkből; a további találatokat soronként fűzi hozzá.

4. A futás végén egy „tiszta", áttekinthető táblázatot ad vissza, amely kizárólag a szignifikáns asszociációkat tartalmazza (változónevekkel, típusinformációval, módszernévvel, statisztikával, p-értékkel).

Ez a megoldás közvetlenül támogatja a nagydimenziós EDA-t: a kutató nem a teljes korrelációs mátrixot lapozza végig, hanem egy célzott listát kap azokról a párokról, amelyek statisztikailag értelmezhető kapcsolatot mutatnak.

Esettanulmányok és kísérleti eredmények

A. Két Likert-skálás változó esete

Egy korábbi kutatásban két, ötfokú Likert-skálán mért kérdőívtétel kapcsolatát vizsgáltuk:

- K31_1: az IKT-eszközök modernségének szubjektív megítélése,

- K31_2: az IKT-túlterhelés érzete.

A következő eredmények születtek:

- Pearson-korreláció: $r \approx 0{,}043$; $df = 496$; $p \approx 0{,}328$ → nincs szignifikáns lineáris kapcsolat.

- χ^2-próba: $\chi^2 \approx 104{,}3$; $df = 16$; $p \approx 5{,}36 \cdot 10^{-15}$ → a két változó között szignifikáns összefüggés mutatható ki a kategóriastruktúrában.

Ez az egyszerű példa rámutat arra, hogy szintaktikailag ugyanazon adathalmaz (egész számok 0 és 1 között normalizált formában) a különböző eljárások számára eltérő típusú bemenetet jelenthet. A keretrendszer heurisztikái alapján mindkét vektor kategorikusnak minősül, ezért automatikusan a χ^2-próbát választja, és egységes szerkezetű outputot ad a felhasználónak.

B. Nagydimenziós kérdőíves adatbázis

A batch-modult egy több mint száz változót tartalmazó kérdőíves adatbázison teszteltük, ahol a cél az volt, hogy gyorsan azonosítsuk azokat a változópárokat, amelyek legalább egy előre definiált küszöbértéknél (pl. $\alpha = 0{,}01$) szignifikáns kapcsolatot mutatnak.

A manuálisan végrehajtott EDA-val összehasonlítva:

- a keretrendszer
- rövid idő alatt több ezer páros összehasonlítást tudott elvégezni,
- csak a releváns eredményeket adta vissza,
- a változótípusokat automatikusan kezelte;
- a kézi megközelítés
- hajlamos a véletlenszerűen kiválasztott tesztfüggvényekre,
- nehezen dokumentálható és reprodukálható,
- jóval időigényesebb.

A végeredmény egy olyan rangsorolható táblázat lett, amely közvetlenül felhasználható további modellezés (pl. regresszió, klaszterezés) előkészítésére, illetve vizualizációk (pl. hálódiagramok, korrelációs hőtérképek) generálására.

Jövőbeli kutatási irányok

A bemutatott keretrendszer prototípus, amely már jelen formájában is csökkenti a kutatói terhelést, ugyanakkor több ponton továbbfejleszthető. A következő irányok különösen ígéretesek:

1. Átfogóbb bemeneti validálás.

- az adatkeret-azonosítók, oszloptartományok, vektorméretek és értékkészletek ellenőrzése;
- „fail-fast” stratégia részleges vagy inkonzisztens bemenetek esetén.

2. Hiányzó adatok és hatékony mintanagyság.

- explicit hiányzó-adat politika (listwise vs. pairwise törlés),
- az effektív mintanagyság riportolása minden változópárra.

3. Duplikáció-kontroll.

- {A,B} és {B,A} párok azonos kezelése,

- ismételt futások során azonos párok újbóli tesztelésének elkerülése.

4. Spurius és közvetett kapcsolatok diagnosztikája.

- részkorrelációk, kondicionálás és modellezési megközelítések integrálása,

- jelölés azoknál a kapcsolatoknál, ahol közös ok vagy mediáció valószínűsíthető.

5. Eljárás-ajánlás és bővített procedúrakészlet.

- további tesztek (pl. Spearman-korreláció, loglineáris modellek, lineáris vegyes modellek) integrálása,

- döntési logika finomítása, ha több eljárás is módszertanilag indokolható.

6. Konfigurálható, adaptív küszöbök.

- az egyedi-érték és egyedi-arány küszöbök felhasználói paraméterként való kezelése,

- érzékenységvizsgálat a küszöbök változtatására.

7. Kódoptimalizálás és nyílt forrású terjesztés.

- a kód átláthatóságának megőrzése mellett mesterséges intelligencia alapú refaktorálási javaslatok óvatos integrálása,

- verziókövetett, dokumentált GitHub-tárhely fenntartása a kutatóközösségi együttműködés támogatására.

Következtetések

A tanulmány egy olyan, R-nyelven megvalósított keretrendszert mutatott be, amely egyesíti a változótípus-felismerés és az asszociáció-elemzés logikáját a feltáró adatelemzés fázisában. A megoldás

- automatizálja a változótípusok heurisztikus azonosítását,

- a típusok kombinációja alapján kiválasztja a megfelelő statisztikai eljárást,

- egységes kimeneti struktúrát biztosít a különböző tesztekhez,

- és adatkeret-szinten képes tömeges, szignifikancia-alapú asszociáció-szűrésre.

A keretrendszer illeszkedik a CRISP-DM folyamatmodellhez, támogatja a reprodukálhatóságot és a módszertani fegyelmet, miközben csökkenti a nagydimenziós adatbázisokkal dolgozó kutatók manuális terhelését. A jövőbeli fejlesztések – különösen a spurius kapcsolatok diagnosztikája, a döntési logika bővítése és az adaptív küszöbök – további lépést jelenthetnek egy teljesen automatizált, mégis átlátható EDA-pipeline irányába.

Köszönetnyilvánítás

A kutatás az EKÖP-24-KDP-2 azonosítójú projekt keretében valósult meg, a Kulturális és Innovációs Minisztérium, valamint a Nemzeti Kutatási, Fejlesztési és Innovációs Alap támogatásával, a 2024-2.1.2 Egyetemi Kutatói Ösztöndíj Program – Kooperatív Doktori Program finanszírozási konstrukcióban.

Hivatkozások

[1] J. W. Tukey, Exploratory Data Analysis. Reading, MA, USA: Addison-Wesley, 1977.

[2] F. Provost and T. Fawcett, Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. Sebastopol, CA, USA: O'Reilly Media, 2013.

[3] S. Dhummad, "The Imperative of Exploratory Data Analysis in Machine Learning," Scholars J. Eng. Technol., vol. 13, no. 1, pp. 30–44, Jan. 2025.

[4] J. S. Saltz, "CRISP-DM for Data Science: Strengths, Weaknesses and Potential next Steps," in Proc. IEEE Int. Conf. Big Data, 2021, pp. 2337–2344.

[5] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," Procedia Comput. Sci., vol. 181, pp. 526–534, 2021.

[6] M. Cazacu and E. Titan, "Adapting CRISP-DM for Social Sciences," BRAIN – Broad Res. Artif. Intell. Neurosci., vol. 11, no. 2 (Suppl. 1), pp. 99–106, 2020.

[7] S. S. Mayya, A. D. Monteiro, and S. Ganapathy, "Types of Biological Variables," J. Thoracic Dis., vol. 9, no. 6, pp. 1730–1733, 2017.

[8] A. Hazra and N. Gogtay, "Biostatistics Series Module 3: Comparing Groups: Numerical Variables," Indian J. Dermatol., vol. 61, no. 3, p. 251, 2016.

[9] A. P. Field, J. Miles, and Z. Field, Discovering Statistics Using R, 2nd ed. London, U.K.: SAGE, 2017.

[10] N. S. Matloff, Probability and Statistics for Data Science: Math + R + Data. Boca Raton, FL, USA: CRC Press, 2020.

[11] H. Wickham and G. Grolemund, R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Beijing, China: O'Reilly, 2017.

[12] K. Okoye and S. Hosseini, R Programming. Singapore: Springer Nature, 2024.

[13] C. J. Wendt and G. B. Anderson, “Ten Simple Rules for Finding and Selecting R Packages,” PLOS Comput. Biol., vol. 18, no. 3, e1009884, Mar. 2022.

[14] S. Wattanakriengkrai et al., “GitHub Repositories with Links to Academic Papers: Public Access, Traceability, and Evolution,” J. Syst. Softw., vol. 183, 111117, Jan. 2022.

[15] S. Konakanchi, “Artificial Intelligence in Code Optimization and Refactoring,” Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 11, no. 2, pp. 1197–1211, Mar. 2025.

[16] M. Bakucz, Á. Horváth, and A. Bozóti, “The Role of Guest Cards in Developing ‘Slow Tourism’ in Austria and Hungary,” in Competitiveness Analysis of Settlements with Spas in Certain Hungarian and Central European Regions, M. Bakucz, Ed. Pécs, Hungary: Univ. Pécs, 2016, pp. 63–70.

[17] A. B. Horváth, “Research on ICT Innovation in Hungarian SMEs,” Business & IT, vol. 13, no. 1, pp. 203–212, 2023.

[18] Á. B. Horváth, “Different Approach of the Digital Transformation at SME,” Acta Polytechnica Hungarica, vol. 20, no. 9, pp. 145–164, 2023.

[19] J. Aldrich, “Correlations Genuine and Spurious in Pearson and Yule,” Statist. Sci., vol. 10, no. 4, pp. 364–376, 1995.

[20] C. S. Calude and G. Longo, “The Deluge of Spurious Correlations in Big Data,” Found. Sci., vol. 22, no. 3, pp. 595–612, 2016.

[21] W. P. Maddison, “Testing Character Correlation Using Pairwise Comparisons on a Phylogeny,” J. Theor. Biol., vol. 202, no. 3, pp. 195–204, 2000.

[22] A. Soetewey, “Correlation Coefficient and Correlation Test in R,” Stats and R blog, May 28, 2020. Available: https://statsandr.com

[23] Szikora, P. (2009). Measured Performance of an Information System. In 7th International Conference on Management, Enterprise and Benchmarking, Budapest (pp. 267-272).