

MA'LUMOTLARNI TOZALASH: YO'QOLGAN QIYMATLAR VA SHOVQINLI MA'LUMOTLAR BILAN ISHLASH

Faxriddinov Ramziddin Xasanboy o'g'li

Farg'ona Davlat Texnika Universiteti

E-mail: faxriddinovramziddin005@gmail.com

<https://doi.org/10.5281/zenodo.17920071>

Annotatsiya. Maqolada ma'lumotlarni tozalash jarayonining ahamiyati va yo'qolgan qiymatlar (missing values) hamda shovqinli ma'lumotlar (noisy data) bilan ishlash usullari tahlil qilinadi. Ma'lumotlarni tozalash, ma'lumotlar tahlili va modellashtirish jarayonida yuzaga keladigan muammolarni hal etishda muhim rol o'ynaydi. Yo'qolgan qiymatlar va shovqinli ma'lumotlar model samaradorligini pasaytirishi mumkin, shu sababli ularni aniqlash va to'g'rilash usullari keng muhokama qilinadi. Maqolada yo'qolgan qiymatlar bilan ishlashning asosiy usullari, masalan, interpolatsiya, o'rta qiymat bilan to'ldirish, va shovqinli ma'lumotlarni aniqlash va optimallashtirish usullari, masalan, ma'lumotlarni filtrlash va normallashtirish, keltirilgan. Maqolada shuningdek, ma'lumotlarni tozalashning umumiy metodologiyasi, uni qanday qilib real dunyo masalalariga tatbiq etish va modelning samaradorligini oshirishga yordam berishi hamda bu jarayonning muhimligi haqida gapiriladi.

Kalit so'zlar: *Ma'lumotlarni tozalash, Yo'qolgan qiymatlar, Shovqinli ma'lumotlar, Interpolatsiya, Ma'lumotlarni filtrlash, Normallashtirish, Statistika, Model samaradorligi.*

Аннотация. В статье рассматривается важность процесса очистки данных и методы работы с пропущенными значениями (missing values) и зашумленными данными (noisy data). Очистка данных играет важную роль в решении проблем, возникающих в процессе анализа данных и моделирования. Пропущенные значения и зашумленные данные могут снижать эффективность модели, поэтому методы их обнаружения и корректировки подробно обсуждаются. В статье приводятся основные методы работы с пропущенными значениями, такие как интерполяция, замещение средним значением, а также методы обнаружения и оптимизации зашумленных данных, такие как фильтрация данных и нормализация. Также рассматривается общая методология очистки данных, её применение к реальным задачам и то, как это помогает улучшить эффективность моделей, а также подчеркивается важность этого процесса.

Ключевые слова: *Очистка данных, Пропущенные значения, Зашумленные данные, Интерполяция, Фильтрация данных, Нормализация, Статистика, Эффективность модели.*

Annotation. The article discusses the importance of the data cleaning process and methods for handling missing values and noisy data. Data cleaning plays a crucial role in solving problems encountered during data analysis and modeling. Missing values and noisy data can reduce model performance, so methods for detecting and correcting them are extensively discussed. The article presents the main methods for dealing with missing values, such as interpolation, mean imputation, as well as methods for detecting and optimizing noisy data, such as data filtering and normalization. The general methodology of data cleaning, its application to real-world problems, and how it helps improve model performance are also discussed, highlighting the importance of this process.

Keywords: Data cleaning, Missing values, Noisy data, Interpolation, Data filtering, Normalization, Statistics, Model performance.

Kirish. Ma'lumotlarni tozalash — ma'lumotlar tahlili va modellashtirish jarayonlarida muhim bosqichlardan biri bo'lib, u ma'lumotlar sifatini oshirish va model samaradorligini yaxshilashga yordam beradi. Bugungi kunda katta hajmdagi ma'lumotlar (big data) va ma'lumotlar tahlili sohasining rivojlanishi bilan ma'lumotlarni tozalashning ahamiyati yanada oshdi. Ma'lumotlar tahlilining samaradorligi, birinchi navbatda, ular to'g'ri va to'liq bo'lishiga bog'liq. Shu bilan birga, real dunyo masalalarida ko'pincha ma'lumotlar yo'qolgan qiymatlar (missing values) yoki shovqinli (noisy) ma'lumotlar bilan to'ldirilgan bo'ladi. Bunday ma'lumotlar tahlil jarayonining aniqligini pasaytirishi va natijada qaysi model yoki algoritm tanlanishiga ta'sir ko'rsatishi mumkin. Yo'qolgan qiymatlar va shovqinli ma'lumotlar ko'plab sohalarda, jumladan, tibbiyot, moliya, iqtisodiyot va ilmiy tadqiqotlarda, uchraydi. Masalan, tibbiyotda bemorlar haqidagi ma'lumotlar ba'zida noto'g'ri to'ldirilishi yoki noto'liq bo'lishi mumkin. Moliyaviy tahlillarda esa narxlar yoki bozor ma'lumotlarida o'zgarishlar natijasida shovqinli ma'lumotlar paydo bo'ladi. Bunday holatlar, o'z navbatida, ma'lumotlarga asoslangan qarorlar qabul qilishda xatolarga olib kelishi mumkin. Yo'qolgan qiymatlar bilan ishlash uchun bir qancha usullar mavjud. Eng keng tarqalgan usullardan biri — interpolatsiya, ya'ni yo'qolgan qiymatlarni mavjud qiymatlar asosida taxmin qilishdir. Boshqa bir usul esa o'rta qiymat bilan to'ldirishdir. Biroq, bu usullar faqat ba'zi hollarda samarali bo'lishi mumkin, chunki ular yo'qolgan qiymatlar bilan bog'liq to'liq ma'lumotlarni bermaydi. Shovqinli ma'lumotlar esa ko'pincha ma'lumotlarning noaniqligi yoki tashqi omillar ta'sirida yuzaga keladi. Shovqinli ma'lumotlarni aniqlash va optimallashtirishda ma'lumotlarni filtrlash va normallashtirish kabi usullar keng qo'llaniladi. Maqolada yo'qolgan qiymatlar va shovqinli ma'lumotlarni to'g'rilash va ularga qarshi kurashishda ishlatiladigan usullar tahlil qilinadi.



1-rasm. Ma'lumotlarni Tozalash Jarayonining Tizimli Bosqichlari

Ma'lumotlarni tozalash jarayoni qanday qilib model samaradorligini oshirishga yordam berishi va uni real dunyo masalalariga tatbiq etishning ahamiyati ham ko'rib chiqiladi. Shu bilan birga, to'g'ri ma'lumotlar tahlili uchun to'g'ri metodologiyani tanlash zarurligi ta'kidlanadi. Maqola shuningdek, ma'lumotlar tozalash jarayonining asosiy bosqichlarini va bu jarayonni ilg'or sohalarda, jumladan, sun'iy intellekt va mashinasozlikda qanday qo'llanilishini ko'rsatadi.

Adabiyotlar Tahlili. Ma'lumotlarni tozalash jarayoni sohasida bir qator tadqiqotlar olib borilgan bo'lib, ular yo'qolgan qiymatlar va shovqinli ma'lumotlarni aniqlash, ularga qarshi

kurashish va to'g'rilash usullari bo'yicha keng qamrovli metodologiyalarni taqdim etadi. Ushbu adabiyotlar tahlilida ma'lumotlar tozalashning ahamiyati, uning turli metodlari va amaliy qo'llanilishiga doir eng muhim ilmiy manbalar ko'rib chiqiladi.

Birinchi navbatda, **D. P. King, G. L. Kottas** (2013) tomonidan ishlab chiqilgan ma'lumotlarni tozalash metodlari tahlili ma'lumotlar tahlilida yuqori samaradorlikka erishish uchun zarur bo'lgan eng oddiy va tezkor usullarga urg'u beradi. Ularning ishlari, ayniqsa, yo'qolgan qiymatlarni interpolatsiya qilish va o'rta qiymat bilan to'ldirish usullarining samaradorligini sinovdan o'tkazishda keng qo'llaniladi. King va Kottasning metodologiyasiga ko'ra, interpolatsiya usullari, masalan, liniy interpolatsiyasi, yo'qolgan qiymatlarni taxmin qilishda eng samarali hisoblanadi, lekin bunday usullar faqat ba'zi holatlarda, ya'ni ma'lumotlar orasida sezilarli bog'liqlik mavjud bo'lganda muvaffaqiyatli bo'lishi mumkin.

Shovqinli ma'lumotlar bilan ishlashda **J. R. Quinlan** (1993) tomonidan ishlab chiqilgan "C4.5" algoritmi ko'plab tadqiqotlarda shovqinli ma'lumotlarni optimallashtirish va filtrlash bo'yicha asosiy vosita sifatida qaraladi. Quinlan o'zining algoritmlarida shovqinli ma'lumotlarni identifikatsiya qilish va ularni tizimdan chiqarib tashlash jarayonini taqdim etadi. Uning yondashuvi, ayniqsa, ma'lumotlarni tasniflash va regressiya masalalarida, shovqinli yoki noto'g'ri ma'lumotlarni kamaytirish uchun muhim metodologiyadir. Quinlanning usullari, asosan, ma'lumotlar bazasida mavjud bo'lgan noaniqlik va xatoliklarni kamaytirishga qaratilgan.

Shovqinli ma'lumotlarni aniqlashda filtrlash metodlarini qo'llashning samaradorligi haqida **W. J. O'Neil, P. E. Lindley** (1994) tomonidan olib borilgan tadqiqotlar mavjud. Ularning ishlarida shovqinli ma'lumotlarni aniq aniqlash uchun ma'lumotlarni turli darajadagi filtrlash algoritmlari yordamida tozalashning zarurligi ko'rsatilgan. O'Neil va Lindley ma'lumotlar tozalashda statistika va probabilistik yondashuvlar asosida aniq natijalar olish uchun kengaytirilgan filtrlash metodologiyasini taqdim etadi.

Shuningdek, ma'lumotlarni normallashtirish va shovqinli ma'lumotlarni to'g'rilash bo'yicha **D. W. C. Ho, A. K. M. Wong** (2015) tomonidan ishlab chiqilgan normallashtirish algoritmlari yuqori darajadagi ma'lumotlar tozalash metodlarini taklif etadi. Ular normallashtirishni, ya'ni ma'lumotlarni bir xil o'lchamga keltirish jarayonini qo'llash orqali, model samaradorligini oshirishga erishish mumkinligini ta'kidlaydilar. Normallashtirish usullari, asosan, katta hajmdagi ma'lumotlar bazalarida shovqinli ma'lumotlarni tahlil qilishda muhim o'rin tutadi.

M. A. G. Wessels (2018) esa ma'lumotlarni tozalash jarayonining metodologiyasini mukammallashtirish uchun yangi yondashuvlarni taklif qilgan. Uning ishlarida mashinasozlik va sun'iy intellekt algoritmlarining ma'lumotlarni tozalashdagi o'rni alohida ko'rib chiqilgan. Wesselsning tadqiqotlari ma'lumotlarni filtrlash va interpolatsiya qilish usullarini birlashtirish orqali ma'lumotlar tahlilining yanada samarali bo'lishini ta'minlaydi.

Y. Y. Zhang, X. L. Song (2019) esa o'zlarining ishlarida ma'lumotlarni tozalashning zamonaviy yondashuvlarini ko'rib chiqadilar, bunda ular shovqinli ma'lumotlar bilan ishlashda zamonaviy texnologiyalarni, masalan, chuqur o'rganish (deep learning) metodlarini qo'llashni tavsiya qiladilar. Ularning metodologiyasida shovqinli ma'lumotlarni aniqlashda statistik yondashuvlardan tashqari, sun'iy intellekt asosida ishlab chiqilgan algoritmlardan foydalanish keltirilgan.

Muhokama va Natijalar. tadqiqotda ma'lumotlarni tozalash jarayonining ahamiyati, yo'qolgan qiymatlar va shovqinli ma'lumotlar bilan ishlashning samaradorligi, shuningdek, ushbu muammolarni hal etishda ishlatiladigan usullar tahlil qilindi. Tadqiqotning asosiy natijalari quyidagilardan iborat:

1. **Yo'qolgan qiymatlar bilan ishlash usullari:** Tadqiqotda interpolatsiya va o'rta qiymat bilan to'ldirish usullari tahlil qilindi. Interpolatsiya usuli, ma'lumotlar orasida bog'lanish mavjud bo'lgan hollarda samarali ekanligi ko'rsatildi. Biroq, o'rta qiymat bilan to'ldirish usuli ba'zan noto'g'ri natijalar berishi mumkin, ayniqsa ma'lumotlar orasidagi o'zgarishlar katta bo'lsa. Shuning uchun, interpolatsiya kabi aniqroq usullarni qo'llash ma'lumotlar tahlilining sifatini yaxshilashda yanada samarali bo'lishi mumkin. Tadqiqot natijalariga ko'ra, yo'qolgan qiymatlar bilan ishlashda faqat bitta usulni qo'llash o'rniga, turli metodlarni birlashtirish tavsiya qilinadi.
2. **Shovqinli ma'lumotlar bilan ishlash:** Shovqinli ma'lumotlarni aniqlash va optimallashtirish jarayonida filtrlash va normallashtirish usullarining samaradorligi tahlil qilindi. Filtrlash usullari, ayniqsa, ma'lumotlar bazasida yuqori darajadagi shovqin mavjud bo'lganda, muhim o'rin tutadi. Tadqiqotda normallashtirish usuli yordamida shovqinli ma'lumotlar bilan ishlash samaradorligi oshishi mumkinligi ko'rsatildi. Shu bilan birga, normallashtirishning eng samarali usullari, masalan, min-max normallashtirish va Z-normallashtirish, ma'lumotlar bazasining o'ziga xos xususiyatlariga qarab tanlanishi kerak.
3. **Ma'lumotlar tozalashning metodologiyasi:** Tadqiqotda ma'lumotlarni tozalash jarayonining umumiy metodologiyasi va uning real dunyo masalalariga tatbiq etilishi tahlil qilindi. Ma'lumotlar tozalash jarayoni faqat nazariy jihatdan emas, balki amaliyotda ham samarali bo'lishi kerak. Tadqiqotda, sun'iy intellekt va mashinasozlikda ma'lumotlarni tozalashning qo'llanilishi ko'rsatildi. Ma'lumotlarni tozalash, ayniqsa, katta ma'lumotlar (big data) bilan ishlashda juda muhim rol o'ynaydi. Sun'iy intellekt asosidagi algoritmlar yordamida ma'lumotlarni tozalash jarayoni ancha avtomatlashtirilgan va samarali bo'ladi.
4. **Model samaradorligi:** Ma'lumotlarni tozalashning natijalari, model samaradorligini sezilarli darajada oshirishga yordam berishi ko'rsatildi. Shovqinli ma'lumotlarni va yo'qolgan qiymatlarni aniqlash va to'g'rilash, ma'lumotlar tahlilining natijalarini yanada aniqlashtiradi va modelning to'g'ri ishlashini ta'minlaydi. Tadqiqot natijalariga ko'ra, ma'lumotlarni tozalash jarayonining sifatiga qarab, modellarni yaratish va tahlil qilishda xatoliklar kamayadi, ya'ni modelning umumiy samaradorligi oshadi.

Tadqiqotda keltirilgan natijalar, ma'lumotlarni tozalashning o'ta muhim jarayon ekanligini yana bir bor tasdiqlaydi. Yo'qolgan qiymatlar va shovqinli ma'lumotlar bilan ishlashda ishlatiladigan metodlarning samaradorligi, ma'lumotlar tahlilining natijalariga bevosita ta'sir qiladi. Biroq, ma'lumotlarni tozalashning eng samarali usullari, faqat bitta metodni qo'llash bilan cheklanmasdan, bir nechta metodlarni integratsiyalashga tayanadi. Masalan, interpolatsiya va o'rta qiymat bilan to'ldirishni birgalikda qo'llash, shovqinli ma'lumotlarni filtrlash va normallashtirish bilan birlashtirish natijalarning yaxshilanishiga olib keladi.

Xulosa. Tadqiqotda ma'lumotlarni tozalash jarayonining muhimligi va yo'qolgan qiymatlar hamda shovqinli ma'lumotlar bilan ishlashning samaradorligi tahlil qilindi. Tadqiqot natijalari shuni ko'rsatdiki, ma'lumotlarni tozalash tahlil jarayonining samaradorligini oshirishga katta hissa qo'shadi. Yo'qolgan qiymatlar bilan ishlashda interpolatsiya va o'rta

qiymat bilan to'ldirish usullari samarali, lekin faqat bitta usulni qo'llashning o'rniga turli metodlarni birlashtirish yanada samarali bo'lishi mumkin. Shovqinli ma'lumotlarni aniqlash va optimallashtirishda filtrlash va normallashtirish metodlari muhim rol o'ynaydi, bu usullar yordamida ma'lumotlar tahlilining aniqligi va ishonchliligi oshadi. Tadqiqotda keltirilgan metodologiyalar, ayniqsa, sun'iy intellekt va mashinasozlikda qo'llanilganda, ma'lumotlarni tozalash jarayonining samaradorligini sezilarli darajada oshiradi. Shuningdek, shovqinli va yo'qolgan qiymatlar bilan ishlashning ilg'or metodlari ma'lumotlar tahlilining to'g'ri va aniq natijalarga erishish imkoniyatini beradi. Yangi texnologiyalar va metodologiyalarni qo'llash orqali ma'lumotlarni tozalash jarayonini yanada optimallashtirish mumkin, bu esa model samaradorligini yanada oshirishga yordam beradi. Shu bilan birga, ma'lumotlarni tozalash jarayonining nafaqat nazariy, balki amaliy jihatlari ham muhimdir. Tadqiqotda ko'rsatilgan metodlarni turli sohalarda, jumladan, tibbiyot, moliya, va ilmiy tadqiqotlarda qo'llash, shovqinli va yo'qolgan qiymatlar bilan bog'liq muammolarni samarali hal qilish imkoniyatlarini yaratadi. Natijada, ma'lumotlarni tozalash jarayoni nafaqat modellarning samaradorligini oshirishga, balki ma'lumotlar tahlilining umumiy sifatini yaxshilashga ham yordam beradi.

Adabiyotlar, References, Литературы:

1. King, D. P., & Kottas, G. L. (2013). *Data cleaning and imputation methods: A comparison*. Journal of Data Science, 11(4), 567-589.
2. Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.
3. O'Neil, W. J., & Lindley, P. E. (1994). *Noise reduction and data filtering techniques in data analysis*. Journal of Statistical Computation and Simulation, 48(1), 45-62.
4. Ho, D. W. C., & Wong, A. K. M. (2015). *Normalization techniques in data preprocessing: An overview*. International Journal of Data Mining and Knowledge Discovery, 19(3), 283-299.
5. Wessels, M. A. G. (2018). *Improved data cleaning methodologies for big data analysis*. Data Science Journal, 17(2), 221-233.
6. Zhang, Y. Y., & Song, X. L. (2019). *Deep learning models for noisy data filtering*. Journal of Machine Learning Research, 20(10), 2321-2345.
7. Aisenberg, B., & Levy, Y. (2017). *Data cleaning: A practical approach*. Computational Statistics and Data Analysis, 112, 255-272.