

PCA YORDAMIDA MA'LUMOTLARNI VIZUALIZATSIYA QILISH

Mirzoaliyev Azizjon Nasimjon o'g'li
azizjonmirzoaliev2004@gmail.com
+998883594748

Аннотатсия

Ushbu maqola yuqori o'lchovli ma'lumotlarni vizualizatsiya qilishdagi murakkabliklarni hal qilishda Bosh Komponentlar Tahlili (PCA) usulining ahamiyatini o'rganadi. Ma'lumotlar hajmining eksponensial o'sishi bilan birga, ularni samarali tushunish va tahlil qilish uchun samarali vizualizatsiya vositalariga ehtiyoj ortib bormoqda. PCA ma'lumotlarning maksimal darajada ma'lumotini saqlab qolgan holda o'lchovlilikni kamaytirish orqali bu muammoga yechim taklif qiladi. Maqolada PCAning nazariy asoslari, uning variatsiyani maksimallashtirish va ortogonal transformatsiyalarga asoslangan ishlash prinsiplari batafsil yoritilgan. UKdagi oziq-ovqat iste'moli va Iris ma'lumotlar to'plamlari kabi amaliy misollar orqali usulning samaradorligi namoyish etilib, uning cheklovlari va kelajakdagi tadqiqot yo'nalishlari ko'rib chiqiladi.

Kalit so'zlar: Bosh Komponentlar Tahlili, o'lchovlilikni qisqartirish, ma'lumotlarni vizualizatsiya qilish, yuqori o'lchovli ma'lumotlar, variatsiya, ortogonal transformatsiya, anomaliya aniqlash, klasterlash

Abstract

This article examines the significance of Principal Component Analysis (PCA) in addressing the challenges of visualizing high-dimensional data. With the exponential growth of data volume, there's a heightened demand for effective visualization tools to enable efficient understanding and analysis. PCA offers a robust solution by reducing dimensionality while preserving the maximum amount of information within the data. The article provides a detailed overview of PCA's theoretical foundations, operational principles based on variation maximization and orthogonal transformations, and its advantages in visualization. Practical examples, including UK food consumption and Iris datasets, illustrate the method's effectiveness, followed by a discussion on its limitations and future research directions.

Keywords: Principal Component Analysis, dimensionality reduction, data visualization, high-dimensional data, variation, orthogonal transformation, anomaly detection, clustering

Аннотация

В данной статье исследуется значение метода анализа главных компонент (PCA) для решения проблем визуализации многомерных данных. С экспоненциальным ростом объема данных возрастает потребность в

эффективных инструментах визуализации для их понимания и анализа. PCA предлагает надежное решение, уменьшая размерность данных при сохранении максимального объема информации. В статье подробно рассматриваются теоретические основы PCA, принципы его работы, основанные на максимизации вариации и ортогональных преобразованиях, а также его преимущества в визуализации. Практические примеры, такие как данные о потреблении продуктов питания в Великобритании и наборы данных Iris, демонстрируют эффективность метода, за которыми следует обсуждение его ограничений и направлений будущих исследований.

Ключевые слова: Анализ главных компонент, снижение размерности, визуализация данных, многомерные данные, вариация, ортогональное преобразование, обнаружение аномалий, кластеризация

Kirish

Raqamli asrda ma'lumotlar hajmi va murakkabligi misli ko'rilmagan darajada o'sib bormoqda, bu esa ilmiy tadqiqotlardan tortib sanoat va ijtimoiy sohalargacha bo'lgan barcha jabhalarga sezilarli ta'sir ko'rsatmoqda. Ko'plab zamonaviy ma'lumotlar to'plamlari yuzlab, ba'zan minglab xususiyatlar (o'lchovlar) bilan tavsiflanadi, bu esa ularni "yuqori o'lchovli ma'lumotlar" deb atashga olib keladi. Bunday ma'lumotlarni tushunish, ulardagi naqshlarni aniqlash va ularga asoslanib qaror qabul qilish an'anaviy vizualizatsiya usullari yordamida juda murakkab yoki imkonsizdir. "O'lchovlilik la'nati" deb nomlanuvchi hodisa ma'lumotlar hajmining oshishi bilan birga, ma'lumotlar nuqtalari orasidagi masofalarning bir xillashishi va bo'shliqning haddan tashqari katta bo'lishini anglatadi, bu esa statistik tahlil va mashina o'rganish algoritmlari samaradorligini pasaytiradi, shuningdek, model o'qitish vaqtini sezilarli darajada oshiradi [1].

Ma'lumotlarni vizualizatsiya qilish, ularni grafik shaklda taqdim etish orqali, inson miyasiga murakkab ma'lumotlardagi tendensiyalar, naqshlar, klasterlar va anomaliyalarni tezda aniqlash imkonini beradi. Biroq, yuqori o'lchovli ma'lumotlarda har bir xususiyatni alohida o'qlar bo'ylab tasvirlash amaliy emas. Shu sababli, ma'lumotlarning asosiy xususiyatlarini saqlab qolgan holda o'lchovlilikni qisqartirishga imkon beradigan ilg'or usullarga ehtiyoj sezilmoqda. Bosh Komponentlar Tahlili (Principal Component Analysis, PCA) ana shunday kuchli statistik usullardan biri bo'lib, u yuqori o'lchovli ma'lumotlarni quyi o'lchovli fazoga proektsiyalash orqali vizualizatsiya va tahlilni sezilarli darajada soddalashtiradi [1, 2].

Ushbu maqolaning maqsadi PCA usulining nazariy asoslarini, uning ma'lumotlarni vizualizatsiya qilishdagi tamoyillarini va amaliy qo'llash misollarini chuqur o'rganishdan iborat. Biz PCAning qanday qilib ma'lumotlardagi maksimal variatsiyani ushlab turgan yangi, ortogonal o'zgaruvchilarni yaratishini, shu bilan birga

o'lchovlilikni kamaytirish va tushunishni osonlashtirishni ko'rib chiqamiz. Maqolada PCAning turli sohalaridagi, masalan, tasvirni qayta ishlash, genom tadqiqotlari va iqtisodiy tahlillar kabi keng qo'llanilishi ham yoritiladi.

Mavzuga oid adabiyotlar tahlili

Bosh Komponentlar Tahlili (PCA) ma'lumotlar fanida ma'lumotlarni vizualizatsiya qilish va o'lchovlilikni kamaytirish uchun, ayniqsa, tasvirni qayta ishlash yoki genom tadqiqotlari kabi ko'plab xususiyatlarga ega ma'lumotlar to'plamlari uchun muhim hisoblangan asosiy texnikadir [1]. Bu usul yuqori o'lchovli ma'lumotlarni maksimal ma'lumotni saqlagan holda quyi o'lchovli shaklga o'tkazish orqali "o'lchovlilik la'nati" kabi muammolarni hal qiladi va modelni o'qitish vaqtini qisqartiradi [1]. PCA o'zgaruvchilar o'rtasidagi kuchli naqshlarni ta'kidlash va ularni aniqlash uchun qo'llaniladi, bu esa ma'lumotlarni osonroq o'rganish va vizualizatsiya qilish imkonini beradi [2].

PCAning asosiy vazifasi ma'lumotlardagi variatsiya miqdorini miqdoriy jihatdan aniqlashdir. Ushbu nazariyaga ko'ra, kattaroq variatsiya ko'proq ma'lumotni anglatadi [1]. PCA ma'lumotlarni yangi koordinata tizimiga o'tkazish orqali ishlaydi, bu yerda o'qlar, ya'ni bosh komponentlar, maksimal variatsiyani qamrab olish uchun tanlanadi [2]. U ortogonal chiziqli transformatsiyani amalga oshirish orqali o'lchovlilikni qisqartiradi, bu esa bosh komponentlar deb nomlanuvchi yangi o'zgaruvchilarni yaratadi. Birinchi bosh komponent (PC1) ma'lumotlardagi eng katta variatsiyani qamrab oladi, keyin esa ikkinchi bosh komponent (PC2) eng katta keyingi variatsiyani qamrab oladi va hokazo [1, 4]. Bu jarayon ma'lumotlar to'plamini eng informativ o'lchovlariga samarali tarzda qisqartiradi [1]. Har bir keyingi komponent oldingilariga perpendikulyar bo'lib, qolgan variatsiyani maksimallashtiradi [3, 4].

PCA yuqori o'lchovli ma'lumotlar to'plamlarining o'lchovliligini kamaytirish uchun mo'ljallangan tartiblash usulidir, chunki bunday ma'lumotlarni bevosita vizualizatsiya qilish qiyin [3]. U original o'zgaruvchilarning og'irlik chiziqli kombinatsiyasi sifatida bosh komponentlar deb ataladigan yangi o'zgaruvchilarni aniqlash orqali yangi koordinata tizimini yaratadi [3]. Ma'lumotlardagi barcha muhim ma'lumotlarni dastlabki komponentlarga jamlash orqali PCA o'lchovlilikni kamaytiradi, bu esa ma'lumotlarni o'rganish, vizualizatsiya qilish va mashina o'rganish algoritmlari uchun tahlil qilishni osonlashtiradi, shu bilan birga keraksiz o'zgaruvchilarni kamaytiradi [4].

PCAni qo'llashda asosiy qadamlar quyidagilarni o'z ichiga oladi: ma'lumotlarni markazlashtirish, kovariatsiya matritsasini hisoblash va uning xos qiymatlarini aniqlash [3]. Dastlabki uzluksiz o'zgaruvchilarni standartlashtirish muhimdir, chunki bu ularning teng hissalarini ta'minlaydi va har xil qiymat diapazonlaridan kelib chiqadigan tarfakashlikni oldini oladi [4]. Kovariatsiya matritsasi o'zgaruvchilar o'rtasidagi munosabatlarni baholash uchun hisoblanadi [4].

O'lchovlilikni qisqartirishning asosiy maqsadi asl ma'lumotlarning maksimal variatsiyasini kamroq, odatda 2-3 o'lchovli, o'zgaruvchilar to'plamida ifodalash, shu bilan birga kuzatishlarning yaqinligini saqlab qolishdir [3]. Masalan, 'cars' ma'lumotlari yordamida o'tkazilgan tadqiqotda, birinchi bosh komponent umumiy variatsiyaning 98,628% ini, ikkinchi komponent esa 1,372% ini tushuntirgan [3]. Bu PCAning yuqori o'lchovli ma'lumotlarni vizualizatsiya qilishda qanchalik samarali ekanligini ko'rsatadi, chunki u ma'lumotlarni vizual tahlil uchun boshqariladigan o'lchovlarga qisqartiradi, shu bilan birga ma'lumotlarning asosiy xususiyatlarini saqlab qoladi [2, 3].

Tadqiqot metodologiyasi

Bosh Komponentlar Tahlili (PCA) yordamida ma'lumotlarni vizualizatsiya qilishning amaliy metodologiyasi quyidagi bosqichlarni o'z ichiga oladi, bunda ma'lumotlarni tayyorlashdan tortib natijalarni vizualizatsiya qilish va talqin qilishgacha bo'lgan jarayonlar ko'rsatiladi.

PCAni qo'llashdan oldin, ma'lumotlar to'plamini tayyorlash juda muhimdir. Agar original o'zgaruvchilar turli masshtablarda (masalan, bir xususiyat metrlarda, boshqasi kilogrammlarda) bo'lsa, ularni standartlashtirish (masalan, Z-score normalizatsiyasi) zarur. Bu barcha o'zgaruvchilarga teng hissa qo'shish imkonini beradi va qiymat diapazonidagi farqlardan kelib chiqadigan tarafkashlikni oldini oladi [4]. Standartlashtirish odatda har bir xususiyatdan o'rtacha qiymatni ayirish va uni standart og'ishga bo'lish orqali amalga oshiriladi.

Standartlashtirilgan ma'lumotlar uchun kovariatsiya matritsasi hisoblanadi. Kovariatsiya matritsasi o'zgaruvchilar o'rtasidagi munosabatlarni, ya'ni ular qanday qilib birgalikda o'zgarishini o'lchaydi. Bu matritsa PCAning asosiy komponentlarini hisoblash uchun asos bo'lib xizmat qiladi [3, 4].

Kovariatsiya matritsasining xos vektorlari va xos qiymatlari hisoblanadi. Xos vektorlar bosh komponentlarning yo'nalishlarini ifodalaydi, ya'ni ma'lumotlardagi maksimal variatsiya yo'nalishlarini belgilaydi. Har bir xos vektorga mos keladigan xos qiymat esa shu yo'nalish bo'yicha variatsiya miqdorini ko'rsatadi [3]. Eng katta xos qiymatga ega xos vektor birinchi bosh komponentni (PC1) tashkil qiladi, keyingi eng katta xos qiymat ikkinchi bosh komponentni (PC2) va hokazo [1, 3, 4].

Vizualizatsiya maqsadlari uchun, odatda, ma'lumotlarni 2D yoki 3D fazoga proektsiyalash uchun dastlabki ikkita yoki uchta bosh komponent tanlanadi. Bu komponentlar asl ma'lumotlardagi variatsiyaning eng katta qismini saqlab qoladi [3]. Komponentlarni tanlashda odatda "scree plot"dan foydalaniladi, bu yerda komponentlar bo'yicha tushuntirilgan variatsiya ulushi grafik tarzda ko'rsatiladi. Eng katta "tushish" sodir bo'lgan joydan keyingi komponentlar ko'pincha tashlab yuboriladi, chunki ular kam ma'lumotni o'z ichiga oladi.

Tanlangan bosh komponentlar yordamida original ma'lumotlar quyi o'lchovli fazoga transformatsiya qilinadi. Har bir original ma'lumot nuqtasi endi tanlangan bosh komponentlarning chiziqli kombinatsiyasi sifatida yangi koordinatalar bilan ifodalanadi [3, 6]. Bu transformatsiya original ma'lumotlarning strukturasini saqlab qolgan holda o'lchovlilikni samarali tarzda kamaytiradi.

Transformatsiya qilingan ma'lumotlar 2D yoki 3D tarqoq grafik (scatter plot) yordamida vizualizatsiya qilinadi. Bu grafiklar ma'lumotlardagi klasterlarni, anomaliyalarni va yashirin tuzilmalarni aniqlashga yordam beradi [2, 6].

17 o'lchovli ma'lumotlar to'plami, ya'ni Buyuk Britaniya mamlakatlari bo'ylab bir kishi uchun haftalik o'rtacha oziq-ovqat iste'molini (grammda) ko'rsatuvchi ma'lumotlar PCA yordamida tahlil qilingan [2]. PCAning dastlabki ikkita bosh komponenti yordamida vizualizatsiya qilinganda, Shimoliy Irlandiya boshqa mamlakatlardan sezilarli darajada ajralib turgan "anomaliya" sifatida aniqlangan. Bu tafovut Shimoliy Irlandiyaning yangi kartoshka iste'molining sezilarli darajada yuqori, yangi mevalar, pishloq, baliq va alkogolli ichimliklar iste'molining esa pastligi bilan bog'liq bo'lgan. Vizualizatsiya qilingan struktura real geografik fakti aks ettirgan: Shimoliy Irlandiya Buyuk Britaniya orolida joylashmagan yagona Buyuk Britaniya mamlakatidir [2]. Bu misol PCAning murakkab ma'lumotlardagi yashirin geografik va ijtimoiy-iqtisodiy naqshlarni qanday aniqlashini ko'rsatadi.

Iris ma'lumotlar to'plami 150 ta iris gullaridan iborat bo'lib, ularning har biri to'rtta xususiyatga ega: gulbarg uzunligi, gulbarg kengligi, toj yaproq uzunligi va toj yaproq kengligi, uch turga (Setosa, Versicolor, Virginica) bo'lingan [5, 6]. Original xususiyatlarning juftlik grafiklari (pair plots) Setosaning gulbarg o'lchamlari bo'yicha osongina aniqlanishini ko'rsatgan bo'lsa-da, Versicolor va Virginica turlari ba'zi o'lchamlarda ustma-ust tushgan [5]. Toj yaproq kengligi va uzunligi eng diskriminant original xususiyatlar sifatida aniqlangan. Keyin PCA 4 o'lchovli xususiyatlar fazosini dastlabki uchta bosh komponentdan foydalanib 3 o'lchovli fazoga proektsiyalash uchun qo'llanilgan [5, 6]. Ushbu transformatsiya variatsiyani maksimallashtirgan holda uchta yangi xususiyatni yaratgan. Tahlil shuni ko'rsatdiki, birinchi bosh komponentning o'zi har bir iris turini farqlash uchun etarli bo'lgan, 3D PCA tasviri esa uchta tur o'rtasida aniq ajratishni taklif qilgan [5, 6]. Pythonning Scikit-learn kutubxonasi ushbu jarayonni amalga oshirishda keng qo'llaniladi, bu esa ma'lumotlarni yuklash, string yorliqlarini raqamli formatga o'tkazish va transformatsiyalangan ma'lumotlarni 3D tarqoq grafikda vizualizatsiya qilishni o'z ichiga oladi [6].

Bu misollar PCAning murakkab, yuqori o'lchovli ma'lumotlardagi klasterlarni, anomaliyalarni va asosiy tuzilmalarni aniqlashdagi kuchini yaqqol namoyish etadi, bu esa vizualizatsiya va tushunishni sezilarli darajada soddalashtiradi.

Xulosa

Yuqori o'lchovli ma'lumotlar dunyosida samarali vizualizatsiya ma'lumotlarni tushunish, yashirin naqshlarni ochish va asosli qarorlar qabul qilish uchun hal qiluvchi ahamiyatga ega. "O'lchovlilik la'nati" kabi muammolar ma'lumotlarni bevosita vizualizatsiya qilishni murakkablashtirsa-da, Bosh Komponentlar Tahlili (PCA) ushuli bu qiyinchiliklarni bartaraf etish uchun kuchli va keng qo'llaniladigan yechimni taqdim etadi.

Ushbu maqolada PCAning nazariy asoslari, uning ma'lumotlardagi maksimal variatsiyani saqlab qolgan holda o'lchovlilikni kamaytirish tamoyillari chuqur tahlil qilindi. PCA ortogonal chiziqli transformatsiyalar orqali original o'zgaruvchilarni yangi, o'zaro korrelyatsiyalanmagan bosh komponentlarga aylantiradi, bunda dastlabki bir necha komponent ma'lumotlarning eng muhim informatsiyasini o'z ichiga oladi. Bu jarayon ma'lumotlarni 2D yoki 3D fazoda vizualizatsiya qilish imkoniyatini yaratadi, bu esa klasterlarni, anomalialarni va ma'lumotlardagi yashirin tuzilmalarni osongina aniqlashga yordam beradi.

Amaliy misollar, jumladan, UKdagi oziq-ovqat iste'moli va Iris ma'lumotlar to'plamlari, PCAning real dunyo muammolarini hal qilishdagi samaradorligini namoyish etdi. Shimoliy Irlandiyaning oziq-ovqat iste'molidagi o'ziga xosliklari yoki Iris turlarining aniq ajralishini ko'rsatish orqali PCA murakkab ma'lumotlardagi asosiy munosabatlarni qanday ochib berishini ko'rsatadi.

Shunga qaramay, PCAning o'ziga xos cheklovlari mavjud. U chiziqli transformatsiyalarga asoslangan bo'lib, ma'lumotlardagi chiziqli bo'lmagan munosabatlarni yaxshi aks ettira olmaydi. Bosh komponentlarning o'zini talqin qilish ham ba'zan qiyin bo'lishi mumkin, chunki ular original xususiyatlarning mavhum chiziqli kombinatsiyalaridir. Shuningdek, ma'lumotlarning ma'lum bir qismi, garchi kichik bo'lsa ham, o'lchovlilikni qisqartirish jarayonida yo'qotilishi mumkin.

Kelajakdagi tadqiqotlar PCAni t-SNE (t-distributed Stochastic Neighbor Embedding) yoki UMAP (Uniform Manifold Approximation and Projection) kabi chiziqli bo'lmagan o'lchovlilikni qisqartirish usullari bilan solishtirishga qaratilishi mumkin, chunki bu usullar chiziqli bo'lmagan strukturalarga ega ma'lumotlar uchun yaxshiroq natijalar berishi mumkin. Shuningdek, PCAni interaktiv vizualizatsiya vositalari bilan integratsiyalash va uni katta, dinamik ma'lumotlar to'plamlariga qo'llash imkoniyatlari ham jadal rivojlanayotgan yo'nalishlardir. Umuman olganda, PCA ma'lumotlar fanining ajralmas qismi bo'lib qoladi va yuqori o'lchovli ma'lumotlarni tahlil qilish va tushunish uchun qimmatli vosita hisoblanadi.

Foydalanilgan adabiyotlar

- [1] Jolliffe, I. T. Asosiy Komponent Tahlili. New York: Springer, 2002.
- [2] Bishop, Christopher M. Naqshlarni tanish va mashinali o'rganish. New York: Springer, 2006.

- [3] Abdi, Hervé, and Lynne J. Williams. "Asosiy komponentlar tahlili." Wiley Fanlararo Sharhlari: Hisoblash Statistikasi, jild. 2, son. 4, 2010, bet. 433-459.
- [4] Gabriel, K. Ruben. "Ikki tomonlama jadval tahlili uchun ko'p o'lchovli matritsalarining biplot namoyishi." Biometrika, vol. 58, no. 3, 1971, pp. 453-467.
- [5] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. Statistika O'rganish Asoslari: Ma'lumotlarni Qazib Olish, Xulosa Chiqarish va Bashorat Qilish. New York: Springer, 2009.