

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Matematički odsjek

Aproksimativno-izometrička difuzijska preslikavanja

Tim 4M

Mihaela Zima
Mate Poljak
Mateja Pejić
Mateo Martinjak

Profesor:
Zlatko Drmač

Zagreb, svibanj 2022.

Sadržaj

1	Uvod	3
2	Problem formulacije	4
2.1	Difuzijsko preslikavanje	5
3	Difuzijska preslikavanja parcijalnog skupa	6
4	Out-of-sample proširenje koje čuva PDM geometriju	7
5	μ-izometrična konstrukcija	8
6	Podatci	9

Spektralnom analizom difuzijske jezgre dobivamo preslikavanje koje podatke šalje u nižedimenzionalan prostor, gdje euklidske udaljenosti između preslikanih podataka predstavljaju difuzijske udaljenosti između odgovarajućih visokodimenzionalnih podataka.

Značajan nedostatak difuzijskog preslikavanja (DM) je potreba za primjenom spektralne dekompozicije na jezgrenu matricu, što postaje neizvedivo za velike skupove podataka.

Predstaviti ćemo učinkovitu aproksimaciju DM ulaganja. Predstavljeni aproksimacijski algoritam proizvodi rječnik podataka identificiranjem malog skupa predstavnika. Zatim se, na temelju ovog rječnika, cijeli skup podataka učinkovito ulaže u niskodimenzionalni prostor. Euklidske udaljenosti u rezultirajućem uloženom prostoru približne su difuzijskim udaljenostima. [1]

1 Uvod

Najnovije metode obrađuju goleme količine visokodimenzionalnih podataka korištenjem mnogostrukosti na kojoj se pretpostavlja da leže točke.

Spektralna analiza jezgre u ovim metodama otkriva unutarnju geometrijsku strukturu podataka. Ova analiza rastavlja naznačenu jezgru i generira svojstvene vektore koji preslikavaju podatke iz ambijentalnog prostora u uloženi prostor koji je obično niske dimenzije.

Postoji rastuća potreba za računalno učinkovitijim metodama koje su praktične za obradu velikih skupova podataka. Glavno računsko opterećenje povezano s metodama jezgre generira se primjenom spektralne dekompozicije na matricu jezgre.

Istaknuti pristup za smanjenje diskutiranog računalnog opterećenja temelji se na Nyströmovoj proširenoj metodi.

Ovaj pristup se temelji na tri faze:

1. Uniformno, nasumično i bez ponavljanje se konstruira podskup podataka.
2. Taj podskup definira jezgru koja je manja od samog skupa podataka. SVD se primjenjuje na manju jezgru.
3. Spektralna dekompozicija manje jezgre proširena je primjenom Nyströmove proširene metode na cijeli skup podataka.

Izlazna vrijednost aproksimirane spektralne dekompozicije pati od nekoliko velikih problema. Smanjenje skupa podataka utječe na kvalitetu spektralne aproksimacije.

Uniformno biranje podskupa podataka obuhvaća većinu distribucije vjerojatnosti podataka. Međutim, rijetki slučajevi, u usporedbi s veličinom podskupa, mogu se izgubiti.

Nyströмова proširena metoda temelji se na invertiranju jezgrene matrice koja je izvedena iz uniformnog biranja podskupa podataka. Ova jezgra ne mora nužno imati puni rang.

Kombiniranje Nyströmovog proširenja sa slučajnim biranjem podskupa podataka može rezultirati netočnim aproksimacijama spektralne dekompozicije.

Usredotočit ćemo se na ublažavanje računske složenosti metode difuzijskih preslikavanja (DM) i omogućavanje njezine primjene za velike skupove podataka.

DM jezgra predstavlja graf u kojem svaka točka odgovara vrhu. Težina svakog brida između bilo kojeg para vrhova odražava sličnost između odgovarajućih točaka na mnogostrukosti i u procesu difuzije. Svojstvene vrijednosti i odgovarajući svojstveni vektori ove jezgrene matrice otkrivaju mnoga svojstva i veze u grafu. Ove svojstvene vrijednosti i svojstveni vektori koriste se za dobivanje DM ulaganja podataka. Difuzijske udaljenosti su sačuvane ovim ulaganjem i izražene su kao euklidske udaljenosti u DM uloženom prostoru, čija je dimenzionalnost obično znatno niža od dimenzionalnosti izvornog ambijentnog prostora podataka.

Učinkovito aproksimiramo DM metodu modificiranjem Nyströmove ekstenzije. Ova aproksimacija, nazvana μ IDM, jamči da je razlika između difuzijskih udaljenosti u DM ulaganju i euklidskih udaljenosti u μ IDM ulaganju, očuvana izometrijski do na danu kontroliranu pogrešku μ . μ IDM koristi niskodimenzionalnu geometriju iz DM ulaganja za konstruiranje rječnika koji aproksimira geometriju cjelokupnog DM ulaganja.

2 Problem formulacije

Neka je \mathcal{M} niskodimenzionalna mnogostrukost koja leži u visokodimenzionalnom euklidskom ambijentnom prostoru \mathbb{R}^m i neka je $d \ll m$ njegova intrinzična dimenzija. Neka je $M \subseteq \mathcal{M}$ skup podataka od $|M| = n$ točaka koje su uzete iz ove mnogostrukosti.

DM ulaže podatke u prostor gdje euklidske udaljenosti između točaka u uloženom prostoru odgovaraju difuzijskim udaljenostima na mnogostrukosti \mathcal{M} .

DM je jezgrene metoda koja se temelji na spektralnoj analizi $n \times n$ jezgrene matrice koja sadrži sličnosti između svih podataka u M .

Definicija 1 (μ -izometrično preslikavanje). Preslikavanja $\Phi : M \rightarrow \mathbb{R}^d$ i $\hat{\Phi} : M \rightarrow \mathbb{R}^{\hat{d}}$ su μ -izometrična ako za svaki $x, y \in M$

$$\left| \|\hat{\Phi}(x) - \hat{\Phi}(y)\| - \|\Phi(x) - \Phi(y)\| \right| \leq \mu.$$

Oznaka $\|\cdot\|$ označava Euklidsku udaljenost u odgovarajućem prostoru.

Predložena metoda konstruira rječnik podataka iz M koje su dovoljne za opisi-
vanje parova udaljenosti između DM uloženih podataka. Zatim se aproksimirana preslikavanje $\hat{\Phi}$ izračunava ekstenzijom izvan uzorka koja čuva uparene difuzijske udaljenosti u rječniku.

2.1 Difuzijsko preslikavanje

DM metoda se temelji na difuzijskoj jezgri K , čiji su elementi

$$k(x, y) \triangleq e^{-\frac{\|x-y\|^2}{\epsilon}}, \quad x, y \in M.$$

Ova jezgra predstavlja sličnosti između točaka podataka u mnogostrukosti. Jezgra se može promatrati kao težinski graf na skupu podataka M . Točke u M koriste se kao vrhovi, a težine bridova definirane su jezgrom K . Stupanj svake točke, odnosno vrha, $x \in M$ u ovom grafu je

$$q(x) = \sum_{y \in M} k(x, y).$$

Normalizacija jezgre s ovim stupnjem nam daje red-stohastičku prijelaznu matricu P čiji elementi za $x, y \in M$ su $p(x, y) = k(x, y) / q(x)$.

Metoda DM ulaže podatke (točke) iz mnogostrukosti \mathcal{M} u Euklidski prostor čija je dimenzionalnost niža od dimenzionalnosti izvornih podataka. Poželjno je raditi sa simetričnom konjugiranom matricom od P , koja je označena s A , čije su vrijednosti

$$[A]_{(x,y)} = a(x, y) \triangleq \frac{k(x, y)}{\sqrt{q(x)q(y)}} = \sqrt{q(x)}p(x, y)\frac{1}{\sqrt{q(y)}}, \quad x, y \in M.$$

Svojstvene vrijednosti od A , $1 = \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n > 0$, te njima pripadni svojstveni vektori ϕ_1, \dots, ϕ_n , koriste se za konstruiranje difuzijskog preslikavanja $\Phi : M \rightarrow \mathbb{R}^\delta$

$$\Phi(x) = [q^{-1/2}(x)(\sigma_1\phi_1(x)), \dots, q^{-1/2}(x)(\sigma_\delta\phi_\delta(x))]$$

zadovoljno mali δ , što je dimenzija uloženog prostora.

Tipično, primjena DM-a na skup podataka M veličine n uključuje sljedeće korake:

1. Koristite jednadžbu $k(x, y) \triangleq e^{-\frac{\|x-y\|^2}{\epsilon}}$, $x, y \in M$, za konstruiranje $n \times n$ jezgre.
2. Izračunajte dijagonalnu matricu Q koja sadrži za točke iz M stupnjeve $q_i \triangleq \sum_{j=1}^n K_{ij}$, za sve $i = 1, \dots, n$.
3. Normalizirajte K pomoću Q da biste dobili $n \times n$ simetričnu jezgru afiniteta $A = Q^{-1/2}KQ^{-1/2}$.
4. Definirajte svojstvene vrijednosti i svojstvene vektore od A primjenom SVD-a na $A = \Phi\Sigma\Phi^T$ da biste dobili matrice

$$\Sigma = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n \end{bmatrix}, \Phi = \begin{bmatrix} | & \cdots & | \\ \phi_1 & \cdots & \phi_n \\ | & \cdots & | \end{bmatrix}$$

koje sadrže svojstvene vrijednosti i svojstvene vektore od A .

5. Koristite matricu $Q^{-1/2}\Phi\Sigma$ kako biste uložili svaku točku $x_i \in M, i = 1, \dots, n$, kao i -ti redak ove matrice. Spektar matrice A opada eksponencijalno i samo mali broj svojstvenih vektora je potreban za dobivanje pouzdanog niskonidimensionalnog uloženog prostora.

3 Difuzijska preslikavanja parcijalnog skupa

Računanje DM ulaganja iz jednadžbe

$$\Phi(x) = [q^{-1/2}(x)(\sigma_1\phi_1(x)), \dots, q^{-1/2}(x)(\sigma_s\phi_s(x))]$$

zahtijeva spektralnu dekompoziciju pune $n \times n$ simetrične difuzijske jezgre. Izvođenje ove dekompozicije na velikim skupovima podataka računalno je skupo. U ovom odjeljku opisujemo učinkovitu metodu za izračunavanje parova difuzijskih udaljenosti između točaka parcijalnog skupa podataka $S \subset M$. Pretpostavljamo da je, bes smanjenja općenitosti, $M = \{x_1, \dots, x_n\}$ i $S = \{x_1, \dots, x_s\}$, $s < n$. Definirajmo parcijalnu jezgru \tilde{K} kao gornju $s \times n$ podmatricu Gaussove jezgre K iz jednadžbe

$$\tilde{k}(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{\epsilon}}$$

Neka je \tilde{Q} $s \times s$ dijagonalna matrica čiji su dijagonalni elementi stupnjevi

$$\tilde{q}(x_i) = \sum_{j=1}^n \tilde{k}(x_i, x_j), i = 1, \dots, s.$$

Sada možemo definirati $s \times n$ jezgru afiniteta \tilde{A} između skupa S i M pomoću

$$\tilde{A} \triangleq \tilde{Q}^{-1/2}\tilde{K}Q^{-1/2}.$$

Definicija 2. Parcijalno difuzijsko preslikavanje (PDM) $\tilde{\Phi} : S \rightarrow \mathbb{R}^s$ parcijalnog skupa S je dano s:

$$\tilde{\Phi}(x) \triangleq [\tilde{q}^{-1/2}(x)(\tilde{\delta}_1\tilde{\phi}_1(x)), \dots, \tilde{q}^{-1/2}(x)(\tilde{\delta}_s\tilde{\phi}_s(x))] \quad (1)$$

Teorem 3. Geometrija od S nakon ulaganja DM je sačuvana pomoću PDM koji je primjenjen na S . To jest, za svaki $x, y \in S$, $\|\tilde{\Phi}(x) - \tilde{\Phi}(y)\| = \|\Phi(x) - \Phi(y)\|$ i $\langle \tilde{\Phi}(x), \tilde{\Phi}(y) \rangle = \langle \Phi(x), \Phi(y) \rangle$.

Dakle ulaganja koja čuvaju difuzijske udaljenosti parcijalnog skupa veličine s možemo dobiti dekompozicijom $s \times s$ matrice umjesto puno veće $n \times n$ matrice.

4 Out-of-sample proširenje koje čuva PDM geometriju

PDM pruža ulaganje $\tilde{\phi} : S \rightarrow \mathbb{R}^s$ parcijalnog skupa S gdje je $s = |S|$. Kako bi proširili to ulaganje na cijeli skup M koristi se out-of-sample metoda proširenja tako da je $\tilde{\phi}$ sačuvan za S . To zovemo prošireno preslikavanje.

Mi ćemo ograničiti naša proširena preslikavanja tako da imaju iste udaljenosti parova kao i PDM. Zbog toga proširena preslikavanja čuvaju difuzijske udaljenosti u S . Za dani parcijalni skup $S \subset M$, veličine s i njegov komplement $\bar{S} = M \setminus S$ veličine $n - s$ možemo opisati jezgru A sa strukturom

$$A = \begin{bmatrix} A_{(S,S)} & A_{(S,\bar{S})} \\ A_{(S,\bar{S})}^T & A_{(\bar{S},\bar{S})} \end{bmatrix} \quad (2)$$

gdje blok $A_{(S,S)} \in \mathbb{R}^{s \times s}$, čuva difuzijske afinitete između točaka skupa S , blok $A_{(\bar{S},\bar{S})} \in \mathbb{R}^{(n-s) \times (n-s)}$ čuva afinitete točaka skupa \bar{S} , a blok $A_{(S,\bar{S})} \in \mathbb{R}^{s \times (n-s)}$ čuva afinitete između točaka u S i \bar{S} . Iz te formulacije dobivamo:

$$\tilde{A} = \begin{bmatrix} A_{(S,S)} & A_{(S,\bar{S})} \end{bmatrix} \quad (3)$$

Aproksimirana $n \times n$ afinitetna matrica:

$$\hat{A} = \hat{\psi} \tilde{A} \hat{\psi}^T = \begin{bmatrix} A_{(S,S)} & A_{(S,\bar{S})} \\ A_{(S,\bar{S})}^T & A_{(S,\bar{S})}^T (A_{(S,S)})^{-1} A_{(S,\bar{S})} \end{bmatrix} \quad (4)$$

Difuzijska afinitetna matrica se može aproksimirati proširenjem (4). DM ulaganje se temelji na spektralnoj dekompoziciji difuzijske afinitetne matrice A , koja je aproksimirana sa \hat{A} . Stoga, kako bi aproksimirali DM ulaganje koristeći spomenuto proširenje, matricu \hat{A} dekompoziramo kao

$$\hat{A} = \hat{\phi} \Lambda \hat{\phi}^T, \quad (5)$$

gdje je $\hat{\phi}$ $n \times s$ matrica s ortonormiranim stupcima, a Λ je $s \times s$ dijagonalna matrica.

Definicija 4. Neka su $\hat{\phi}$ i Λ matrice iz (5). Ortogonalno Nyström preslikavanje (ONM) je preslikavanje $\hat{\phi} : M \rightarrow \mathbb{R}^s$, dano s:

$$\hat{\phi}(x) \triangleq [q^{-1/2}(x)(\lambda_1 \hat{\phi}_1(x)), \dots, q^{-1/2}(x)(\lambda_s \hat{\phi}_s(x))], \quad (6)$$

gdje su $\phi_1, \dots, \phi_s \in \mathbb{R}^n$ stupci matrice $\hat{\phi}$, a λ_i je i -ti dijagonalni element od Λ . Drugim riječima, ONM ulaže svaku točku iz M u \mathbb{R}^s sa odgovarajućim redom matrice $Q^{-1/2} \hat{\phi} \Lambda$.

ONM iz gornje definicije ulaže čitav skup M u \mathbb{R}^s te koristi spektralnu dekompoziciju $s \times s$ matrice umjesto spektralne dekompozicije $n \times n$ matrice.

Propozicija 5. *Neka su $\tilde{\phi}$ i $\hat{\phi}$ redom, PDM i ONM funkcije ulaganja. Tada za svaki $x, y \in S$, $\|\tilde{\phi}(x) - \tilde{\phi}(y)\| = \|\hat{\phi}(x) - \hat{\phi}(y)\|$ i $\langle \tilde{\phi}(x), \tilde{\phi}(y) \rangle = \langle \hat{\phi}(x), \hat{\phi}(y) \rangle$.*

Korolar 6. *Geometrija od S pod DM ulaganjem je sačuvana ONM ulaganjem. To jest, za svaki $x, y \in S$, $\|\hat{\phi}(x) - \hat{\phi}(y)\| = \|\phi(x) - \phi(y)\|$ i $\langle \hat{\phi}(x), \hat{\phi}(y) \rangle = \langle \phi(x), \phi(y) \rangle$.*

Definicija $\hat{\phi}$ i Λ kao SVD:

$$C = \Psi \Lambda \Psi^T, \quad (7)$$

gdje je C definiran kao

$$C \triangleq A_{(S,S)} + A_{(S,S)}^{-1/2} A_{(S,\bar{S})} A_{(S,\bar{S})}^T A_{(S,S)}^{-1/2}. \quad (8)$$

Pomoću Ψ i Λ dobijemo $\hat{\phi}$ kao:

$$\hat{\phi} = \begin{bmatrix} A_{(S,S)} \\ A_{(S,\bar{S})} \end{bmatrix} A_{(S,S)}^{-1/2} \Psi \Lambda^{-1/2} \quad (9)$$

5 μ -izometrična konstrukcija

Opisujemo konstruktivnu metodu za određivanje skupa $S \in M$ tako da rezultirajući ONM iz definicije (3) bude μ -izometričan sa DM, što koristi cijelu difuzijsku jezgru. Predložena metoda koristi jedan sken cijele baze M i optimizira riječnik odabranog skupa S za svaku procesuiranu točku.

Dani algoritam je iterativni i postepeno konstruira riječnik podskupa S te odgovarajući ONM. Za opis algoritma, koristili smo sljedeće oznake za skup podataka $M = \{x_1, x_2, \dots, x_n\}$. Kako algoritam skenira M samo jednom, gdje svaka iteracija promatra jedinstvenu točku skupa, indeksiranje točaka skupa daje trenutni korak iteracije. To jest u j -tom ($j \in \{1, 2, \dots, n\}$) koraku iteracije promatramo j -tu točku skupa M . Za j -tu iteraciju algoritam daje podriječnik $S_j = \{y_1, y_2, \dots, y_{n_j}\}$, gdje je S_j podskup od $M_j = \{x_1, x_2, \dots, x_{n_j}\}$, $n_j \leq j$. Algoritam konstruira monotono rastući niz podriječnika, to jest, $S_{j-1} \subset S_j$ za svaki $j \in \{2, \dots, n\}$. Konačni riječnik S_n označavamo sa S , a sa $\hat{\phi}$ označavamo ONM $\hat{\phi} : M \rightarrow \mathbb{R}^{n_j}$ primijenjeno na S_j .

Definicija 7. Neka matrice

$$[\hat{\phi}_k(S_k)] = \begin{bmatrix} - & \hat{\phi}_k(y_1) & - \\ & \vdots & \\ - & \hat{\phi}_k(y_{n_k}) & - \end{bmatrix}, [\hat{\phi}_l(S_k)] = \begin{bmatrix} - & \hat{\phi}_l(y_1) & - \\ & \vdots & \\ - & \hat{\phi}_l(y_{n_k}) & - \end{bmatrix} \quad (10)$$

sadržavaju koordinate točaka iz riječnika S_k prema $\hat{\phi}_k$ i $\hat{\phi}_l$, redom. Linearnu Map-to-Map (MTM) transformaciju $T_{k,l} : \mathbb{R}^{n_k} \rightarrow \mathbb{R}^{n_l}$ definiramo kao primjenu matrice $[T_{k,l}] \triangleq [\hat{\phi}_k(S_k)]^{-1}[\hat{\phi}_l(S_k)]$ na vektore $u \in \mathbb{R}^{n_k}$ tako da $T_{k,l}(u) = u[T_{k,l}] \in \mathbb{R}^{n_l}$.

Inicijalno, riječnik je skup koji sadrži samo jednu točku x_1 . Zatim, za svaku iteraciju k , točke iz $M_k = \{x_1, x_2, \dots, x_k\}$ koje su prethodno skenirane, aproksimiramo pomoću konstruiranog riječnika $S_k \subseteq M_k$. Algoritam prelazi na sljedeću točku x_{k+1} i provjerava je li aproksimacija njenog ulaganja pomoću riječnika S_k dovoljno precizna. Ako da, algoritam prelazi na sljedeću iteraciju i riječnik se ne mijenja, inače, dodaje točku x_{k+1} u riječnik S_k . U sljedećoj iteraciji imamo $S_{k+1} = S_k \cup \{x_{k+1}\}$.

Theorem 8. *Neka je ϕ Dm ulaganje od M . Neka je $S \subseteq M$ riječnik konstruiran gore opisanim algoritmom i neka je $\hat{\phi}$ ONM bazirano na riječniku. Tada za svaki $x, y \in M$, $\|\hat{\phi}(x) - \hat{\phi}(y)\| \approx \|\phi(x) - \phi(y)\|$, gdje je μ maksimalna greška aproksimacije.*

Dakle, prethodni teorem pokazuje da parametar μ iz ranije opisanog algoritma diktira najgori moguću grešku aproksimirane parovne udaljenosti od μIDM .

Algorithm 5.1: The μ -isometric DM (μIDM).

Input: data points: $x_1, \dots, x_n \in \mathbb{R}^m$.
Parameters: Distance error bound μ , Gaussian width ε
Output: The approximated DM coordinates $\hat{\phi}(x_i)$, $i = 1, \dots, n$

- 1: Initialize the dictionary: $S_1 \leftarrow \{x_1\}$
- 2: Initialize Q (Eq. (2.2)) and \tilde{A} given S_1 (Eq. (4.2))
- 3: Initialize the embedding: $\hat{\phi} \leftarrow \text{ONM}$ (Definition 4.1) of S_1 .
- 4: for $\kappa = 1$ to $n - 1$ do
 - Set $S' \leftarrow S_\kappa \cup \{x_{\kappa+1}\}$
 - Compute \tilde{Q}' (Eq. (2.2)) and \tilde{A}' given S'
 - Compute $\hat{\phi}' \leftarrow \text{ONM}$ (Definition 4.1) of S'
 - Membership Test:
 - Compute $T \leftarrow \text{MTM}$ (Definition 5.1) from $\hat{\phi}(\cdot)$ to $\hat{\phi}'(\cdot)$
 - Compute $\beta \leftarrow \|T(\hat{\phi}(x_{\kappa+1})) - \hat{\phi}'(x_{\kappa+1})\|$
 - If $\beta > \frac{\mu}{2}$
 - Set $S_{\kappa+1} \leftarrow S'$
 - Set $\tilde{Q} \leftarrow \tilde{Q}'$ and $\tilde{A} \leftarrow \tilde{A}'$
 - Set $\hat{\phi} \leftarrow \hat{\phi}'$
 - Else
 - Set $S_{\kappa+1} \leftarrow S_\kappa$
 - End if
- 5: Output the approximated diffusion coordinates $\hat{\phi}(x_1), \dots, \hat{\phi}(x_n)$

6 Podatci

Svaki skup je ulagan u visoko dimenzionalan prostor, zatim je uniformno uzet uzorak od 500 točaka. Ti skupovi su ulagani u \mathbb{R}^{17} koristeći nasumični puni-rang lineranu transformaciju, čija je reprezentativna matrica 17×3 matrica kojoj su ulazi uniformni iz $[0,1]$.

Literatura

- [1] URL: <https://www.sciencedirect.com/science/article/pii/S1063520314000803>.