EKSTRAKCIJA ZNAČAJKI IZ 1D PODATAKA KORIŠTENJEM DEKOMPOZICIJE TENZORA*

A. Jukić

Zavod za laserska i atomska istraživanja i razvoj

RAČUNARSKA ZNANOST I INŽENJERSTVO

SAŽETAK

Predstavljena je metoda za ekstrakciju značajki iz jednodimenzionalnih podataka. Linearna transformacija izvornih značajki u reducirani skup konstruirana je dekompozicijom podatkovnog tenzora prema Tucker-2 modelu. Jednodimenzionalni podaci složeni su u tenzor trećeg reda, čime je omogućena primjena tenzorske analize za 1D podatke. Metoda je primijenjena na problemu detekcije raka prostate i raka jajnika iz masenih spektara dobivenih analizom proteina u uzorcima krvnog seruma pacijenata.

UVOD

- Ekstrakcija i odabir značajki su ključni problemi u analizi podataka sa velikim brojem varijabli.
- Korištenjem malog broja značajki izbjegava se prilagođavanje klasifikatora podacima za učenje (eng. *overfitting*), te se smanjuju potrebni računalni resursi i trajanje analize.
- Smanjivanje dimenzionalnosti, odnosno broja značajki, može se raditi tako da: (*i*) biramo podskup izvornog skupa značajki; (*ii*) radimo transformaciju izvornih značajki na reducirani skup.
- Tenzori su prirodan prikaz višedimenzionalnih podataka, pa su dekompozicije tenzora prikladne za njihovu analizu.
- Ideja je iskoristiti tenzorski prikaz u analizi jednodimenzionalnih podataka.

DEKOMPOZICIJA TENZORA

- Tenzor 3. reda $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ sa elementima $\left\{x_{i_1 i_2 i_3}\right\}_{i_1,i_2,i_3=1}^{I_1,I_2,I_3}$

- Osnovni model za dekompoziciju je Tucker-3 model

$$\underline{\mathbf{X}} \approx \underline{\widehat{\mathbf{X}}} = \underline{\mathbf{G}} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \times_3 \mathbf{A}^{(3)}$$

koji se za pojedini element može zapisati kao

$$x_{i_1 i_2 i_3} \approx \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \sum_{j_3=1}^{J_3} g_{j_1 j_2 j_3} a_{i_1 j_1}^{(1)} a_{i_2 j_2}^{(2)} a_{i_3 j_3}^{(3)}$$

- $\underline{\mathbf{G}} \in \mathbb{R}^{J_1 \times J_2 \times J_3}$ je jezgreni tenzor, matrice $\mathbf{A}^{(k)} \in \mathbb{R}^{I_k \times J_k}$ su faktori
- Uključivanjem jednog faktora u jezgru dobiva se Tucker-2 model

$$\underline{\mathbf{X}} \approx \underline{\widehat{\mathbf{X}}} = \underline{\mathbf{F}} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \quad (1)$$

- Dekompozicija općenito nije jedinstvena: nužno je uvesti ograničenja za jezgreni tenzor i faktore za jedinstvenu faktorizaciju.

- HOOI algoritam: minimizacija Frobeniusove norme između izvornog tenzora $\underline{\mathbf{X}}$ i modela $\underline{\widehat{\mathbf{X}}}$: $D[\underline{\mathbf{X}}||\underline{\widehat{\mathbf{X}}}] = \left\|\underline{\mathbf{X}} - \underline{\widehat{\mathbf{X}}}\right\|_F^2$, uz ortogonalnost faktora i potpunu ortogonalnost i uređenost jezgrenog tenzora. Zbog ortogonalnosti faktora slijedi

$$\underline{\mathbf{F}} = \underline{\mathbf{X}} \times_1 \mathbf{A}^{(1)T} \times_2 \mathbf{A}^{(2)T} \quad (2)$$

EKSTRAKCIJA ZNAČAJKI

- Neka je sa $\mathbf{X}^{(k)} \in \mathbb{R}^{I_1 \times I_2}$ predstavljen *k*-ti 2D uzorak (matrica) iz skupa za učenje, koji pripada klasi c_k. Za ekstrakciju reduciranog skupa značajki koristimo zajedničku aproksimativnu dijagonalizaciju matrica

$$\mathbf{X}^{(k)} \approx \mathbf{A}^{(1)} \mathbf{F}^{(k)} \mathbf{A}^{(2)T} \quad (3)$$

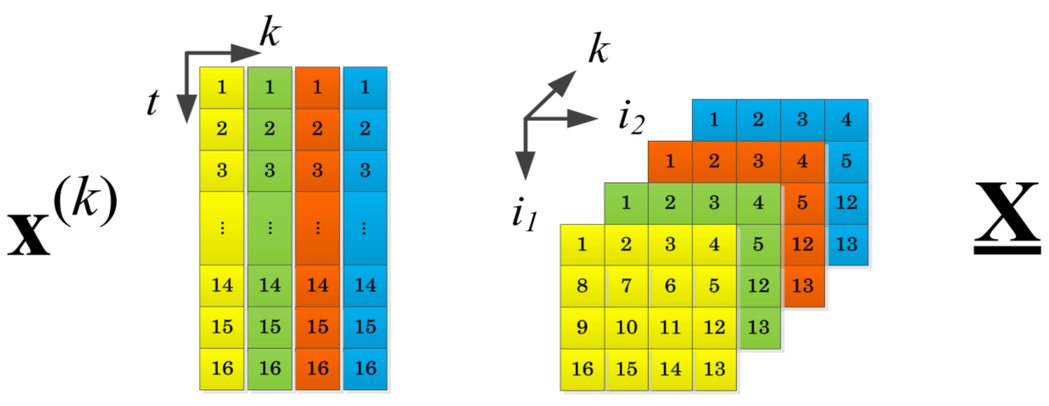
za sve *k*=1,..,*K*.

- Matrice $\mathbf{A}^{(1)} \in \mathbb{R}^{I_1 \times J}$ i $\mathbf{A}^{(2)} \in \mathbb{R}^{I_2 \times J}$, $J \ll I_1, I_2$, koriste se za linearnu transformaciju i dobivanje reduciranog skupa značajki za *k*-ti uzorak $\mathbf{F}^{(k)} \in \mathbb{R}^{J \times J}$. Time smo uzorak sa $I_1 \cdot I_2$ elemenata predstavili sa J^2 značajki.
- Zajednička dijagonalizacija (3) ekvivalentna je dekompoziciji (1) ako svaki uzorak za učenje $\mathbf{X}^{(k)}$ odgovara *k*-tom frontalnom odsječku tenzora $\underline{\mathbf{X}}$.
- Dekompozicija sa HOOI algoritmom daje ortogonalne faktore, pa iz (2) dobivamo tenzor $\underline{\mathbf{F}} \in \mathbb{R}^{J \times J \times K}$ čiji je *k*-ti frontalni odsječak jednak $\mathbf{F}^{(k)}$.

- Od uzoraka za testiranje formiramo tenzor $\underline{\mathbf{X}}_{test}$, a pripadne značajke dobivamo kao

$$\underline{\mathbf{F}}_{test} = \underline{\mathbf{X}}_{test} \times_1 \mathbf{A}^{(1)T} \times_2 \mathbf{A}^{(2)T}$$

- 2D uzorke dobivamo iz 1D podataka $\mathbf{x}^{(k)} \in \mathbb{R}^T, T = I_1 \cdot I_2$ preslagivanjem u matricu $\mathbf{X}^{(k)} \in \mathbb{R}^{I_1 \times I_2}$ kako je prikazano na slici.



EKSPERIMENTI

Demonstracija na problemu detekcije raka prostate i raka jajnika. Korišteni su javno dostupni podaci iz programa kliničke proteomike, NCI, SAD (http://home.ccr.cancer.gov/ncifdaproteomics/)

Podaci i parametri

- Niskorezolucijski maseni spektri, dobiveni SELDI-TOF postupkom iz bioloških uzoraka, svrstani u pripadne (poznate) klase. Svaki uzorak sastoji se od T_0=15154 elemenata koji odgovaraju omjerima masa-naboj *m*/*z*
- Rak prostate: 69 pozitivnih i 63 kontrolna uzorka
- Rak jajnika: 100 pozitivnih i 100 kontrolnih uzoraka
- *T*=15129, I_1=I_2=123, dimenziju jezgre *J* mijenjamo
- Klasifikator učimo na značajkama $\mathbf{f}^{(k)} = vect\left(\mathbf{F}^{(k)}\right)$, a testiramo na značajkama $\mathbf{f}_{\text{test}}^{(m)} = vect\left(\mathbf{F}_{test}^{(m)}\right)$

Rezultati

Broj značajki, J^2	Rak prostate	Rak jajnika
36	91.9±5.6 / 91.4±4.8	84.3±6.1 / 81.4±5.9
100	98.2±2.8 / 95.6±3.6	91.1±4.6 / 87.7±5.0
256	98.8±2.2 / 96.5±3.4	93.6±3.7 / 91.9±4.1
400	99.3±1.6 / 97.8±3.3	95.5±3.2 / 93.4±3.6
625	99.6±1.2 / 98.7±2.9	96.8±2.9 / 95.4±3.5

Osjetljivost i specifičnost u postotcima (srednja vrijednost ± standardna devijacija) dobivene korištenjem linearnog SVM klasifikatora, estimirane 2-fold krosvalidacijom sa 200 slučajnih particija.

* I. Kopriva, A. Jukić, A. Cichocki. Feature extraction for cancer prediction by tensor decomposition of 1D protein expression levels. To be presented at the IASTED Conference on Computational Bioscience, Cambridge, UK, July 11-13, 2011.

Mentor: Ivica Kopriva

Znanstveni susreti 3. vrste, Institut Ruđer Bošković, 7.-8. srpnja 2011.