

Optimalizace vzdálenosti pro multi-instanční shlukovací problémy

Marek Dědič¹²

Školitel: Ing. Tomáš Pevný, Ph.D.³⁴

¹ČVUT v Praze, Fakulta jaderná a fyzikálně inženýrská, Matematická informatika

²Cisco Systems Inc., Karlovo náměstí 10, Praha 2

³ČVUT v Praze, Fakulta elektrotechnická

⁴Avast Software s.r.o., Pikrtova 1737/1a, Praha 4

4. listopadu 2019



- Úloha & Motivace
- MIL
- Contrastive predictive coding
- Triplet loss
- Magnet loss
- Závěr

Úloha & Motivace

- Úloha & Motivace
- MIL
- Contrastive predictive coding
- Triplet loss
- Magnet loss
- Závěr

Hledání zobrazení úloh do prostorů, ve kterých je lze shlukovat.

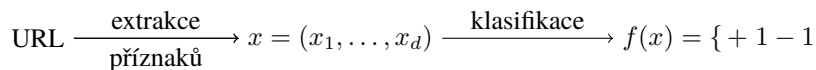
Hledání zobrazení multi-instančních úloh do prostorů, ve kterých je lze shlukovat.

Hledání zobrazení multi-instančních úloh do prostorů, ve kterých je lze shlukovat pomocí metod strojového učení.

- Umožnit shlukování a následnou klasifikaci IP adres
- Využít struktury dat pomocí multi-instančního učení
- Vyzkoušet unsupervised metody

- Úloha & Motivace
- MIL
- Contrastive predictive coding
- Triplet loss
- Magnet loss
- Závěr

End-to-end učení



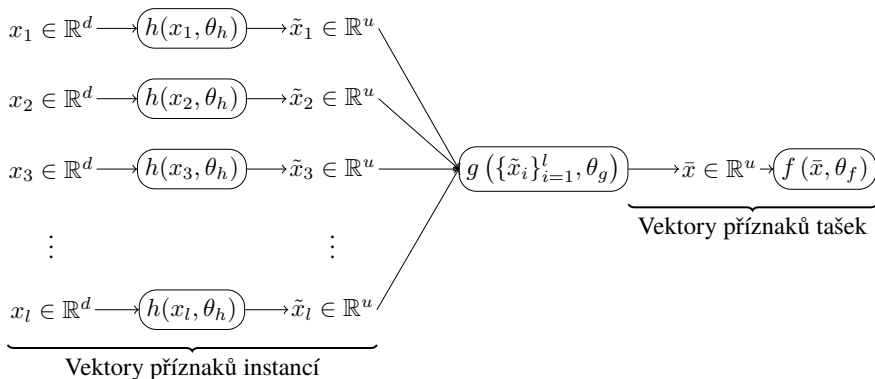
Multi instanční učení

$$\text{URL} \xrightarrow[\text{příznaků}]{\text{extrakce}} x = \left\{ \begin{array}{c} (x_{1,1}, \dots, x_{1,d}) \\ (x_{2,1}, \dots, x_{2,d}) \\ \vdots \\ (x_{b,1}, \dots, x_{b,d}) \end{array} \right\} \xrightarrow{\text{klasifikace}} f(x) = \{ +1 -1$$

Paradigma vloženého prostoru

$$\left\{ \begin{array}{c} (x_{1,1}, \dots, x_{1,d}) \\ (x_{2,1}, \dots, x_{2,d}) \\ \vdots \\ (x_{b,1}, \dots, x_{b,d}) \end{array} \right\} \xrightarrow[\text{do vektoru}]{\phi(b) \text{ vloží tašku}} \bar{x} \in \mathbb{R}^m \xrightarrow{\text{klasifikace}} f(\bar{x}) = \{+1, -1\}$$

Vkládající funkce ϕ



Contrastive predictive coding

- Úloha & Motivace
- MIL
- Contrastive predictive coding
- Triplet loss
- Magnet loss
- Závěr

Contrastive predictive coding

Původní vyjádření přístupu pomocí contrastive predictive coding jako ztrátové funkce

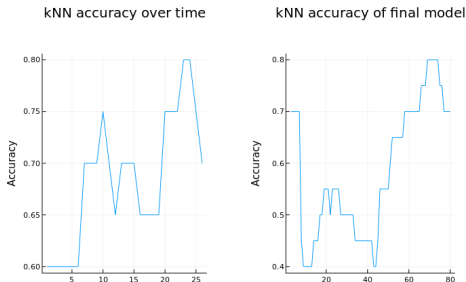
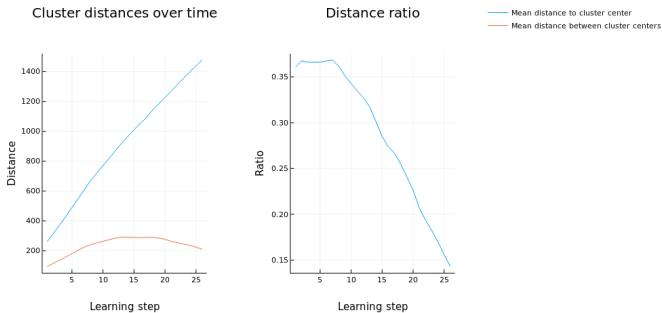
$$\log \left\| f \left(B_n^{(1)} \right) - f \left(B_n^{(2)} \right) \right\|^2 - \log \sum_{j=1}^K \left\| f \left(B_n^{(1)} \right) - f \left(B_j' \right) \right\|^2$$

Contrastive predictive coding

Později zjednodušeno na

$$D_{ij} = \left\| f \left(B_i^{(1)} \right) - f \left(B_j^{(2)} \right) \right\|_2^2$$
$$\frac{1}{n} \sum_{i=1}^n \left(\log (D_{ii}) - \log \left(\sum_{i \neq j} D_{ij} \right) \right)$$

Contrastive predictive coding - předběžné výsledky (Musk2)



Triplet loss

- Úloha & Motivace
- MIL
- Contrastive predictive coding
- **Triplet loss**
- Magnet loss
- Závěr

Triplet loss

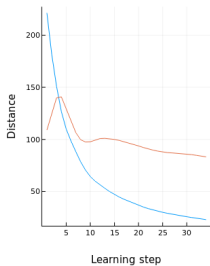
Triplet loss je alternativou vyžadující supervised přístup

$$y_{ij} = \begin{cases} 1 & \text{for } y_i = y_j \\ 0 & \text{otherwise} \end{cases}$$

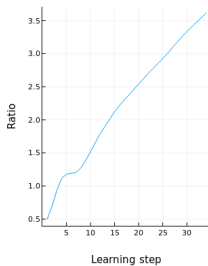
$$\sum_{ij} y_{ij} D_{ij} + c \sum_{ijl} y_{ij} (1 - y_{il}) \max(0, 1 + D_{ij} - D_{il})$$

Triplet loss - předběžné výsledky (Musk2)

Cluster distances over time

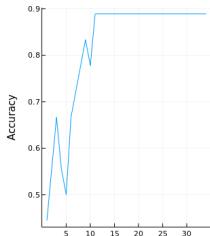


Distance ratio

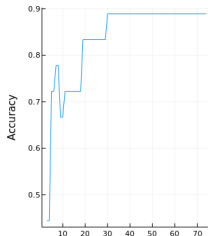


— Mean distance to cluster center
— Mean distance between cluster centers

kNN accuracy over time



kNN accuracy of final model

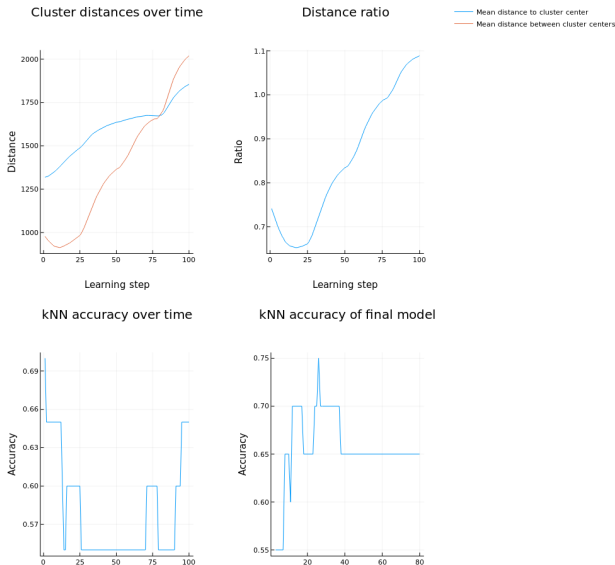


Magnet loss

- Úloha & Motivace
- MIL
- Contrastive predictive coding
- Triplet loss
- **Magnet loss**
- Závěr

Magnet loss - předběžné výsledky (Musk2)

Magnet loss je vylepšenou verzí triplet loss, využívající více sousedních vzorků.



- Úloha & Motivace
- MIL
- Contrastive predictive coding
- Triplet loss
- Magnet loss
- Závěr

- Vyzkoušeno několik přístupů ke shlukování dat
- Aplikováno multi-instanční učení na problém shlukování dat
- Porovnání přístupů na veřejně dostupných standardních datasetech.