Optimalizace vzdálenosti pro multi-instanční shlukovací problémy

Marek Dědič[12]
Školitel: Ing. Tomáš Pevný, Ph.D.[34]

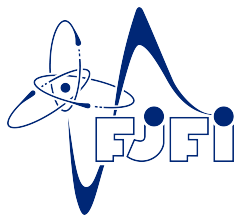
[1]ČVUT v Praze, Fakulta jaderná a fyzikálně inženýrská, Matematická informatika

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

[3]ČVUT v Praze, Fakulta elektrotechnická

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

4. listopadu 2019



Obsah

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

Úloha & Motivace

Úloha

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

Úloha

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

Úloha

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

Motivace

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

MIL

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

End–to–end učení

$$\text{URL} \xrightarrow[\text{příznaků}]{\text{extrakce}} x = (x_1, \ldots, x_d) \xrightarrow{\text{klasifikace}} f(x) = \{+1 - 1$$

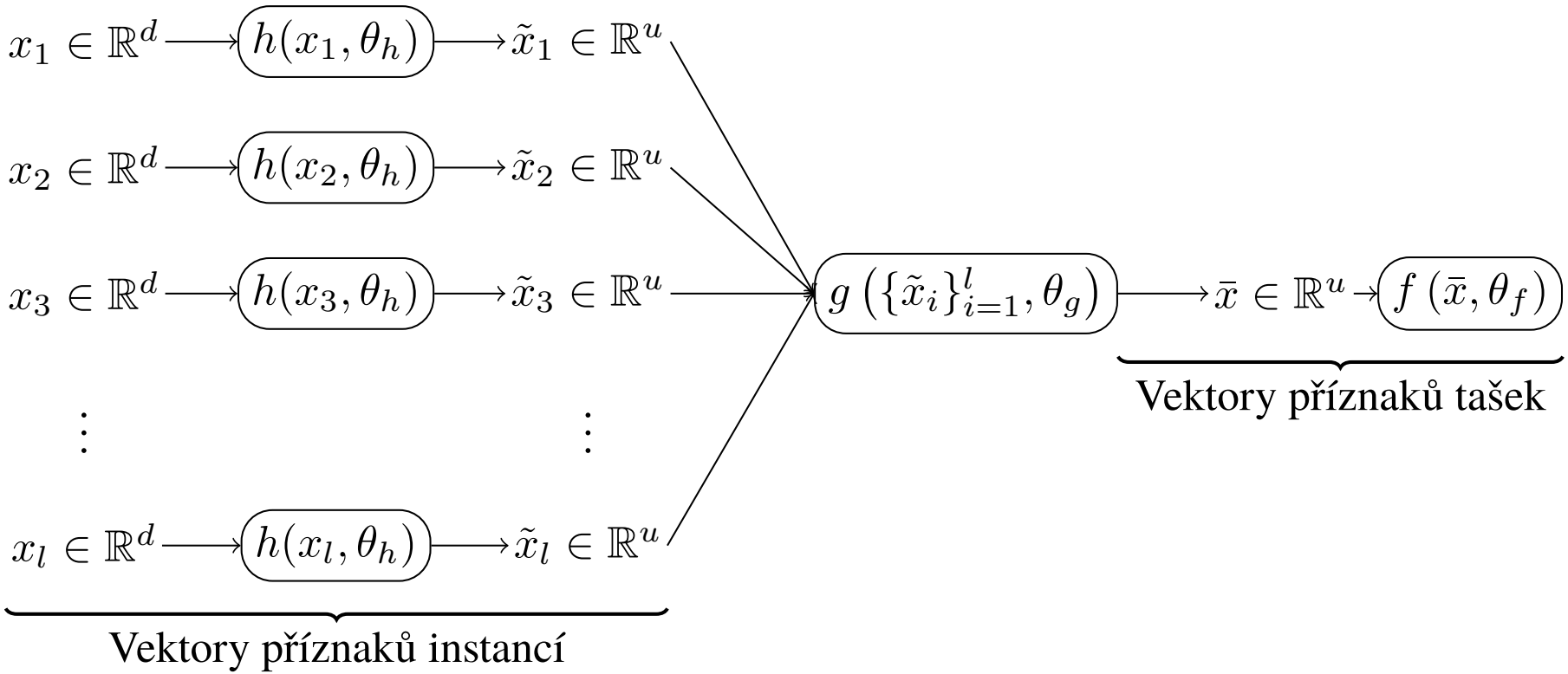
Multi instanční učení

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

Paradigma vloženého prostoru

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

Vkládající funkce ϕ



Contrastive predictive coding

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

Contrastive predictive coding

Původní vyjádření přístupu pomocí constractive predictive coding jako ztrátové funkce

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

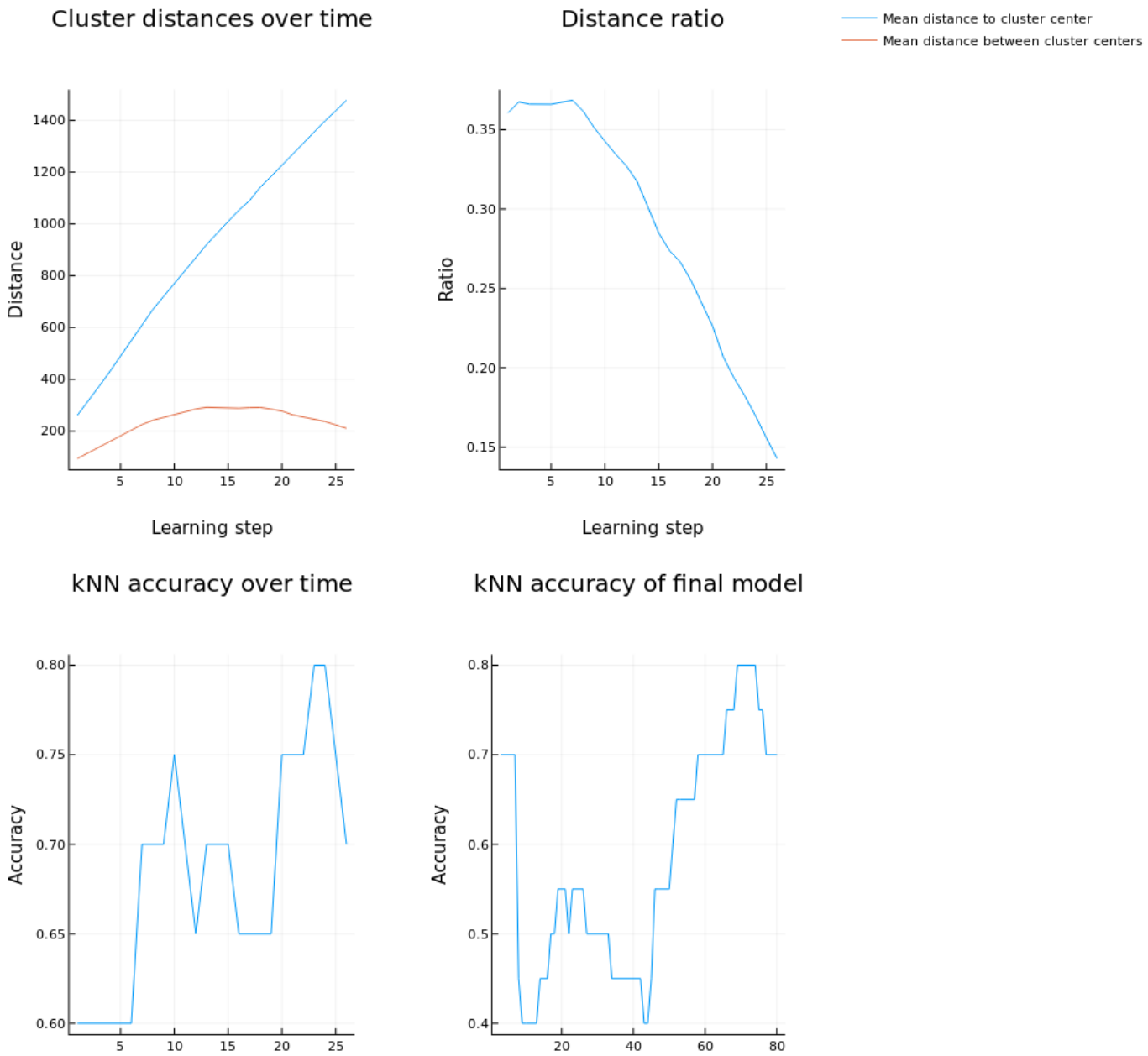
Contrastive predictive coding

Později zjednodušeno na

$$D_{ij} = \left\| f\left(B_i^{(1)}\right) - f\left(B_j^{(2)}\right) \right\|_2^2$$

$$\frac{1}{n} \sum_{i=1}^{n} \left(\log\left(D_{ii}\right) - \log\left(\sum_{i \neq j} D_{ij} \right) \right)$$

Contrastive predictive coding - předběžné výsledky (Musk2)



Triplet loss

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

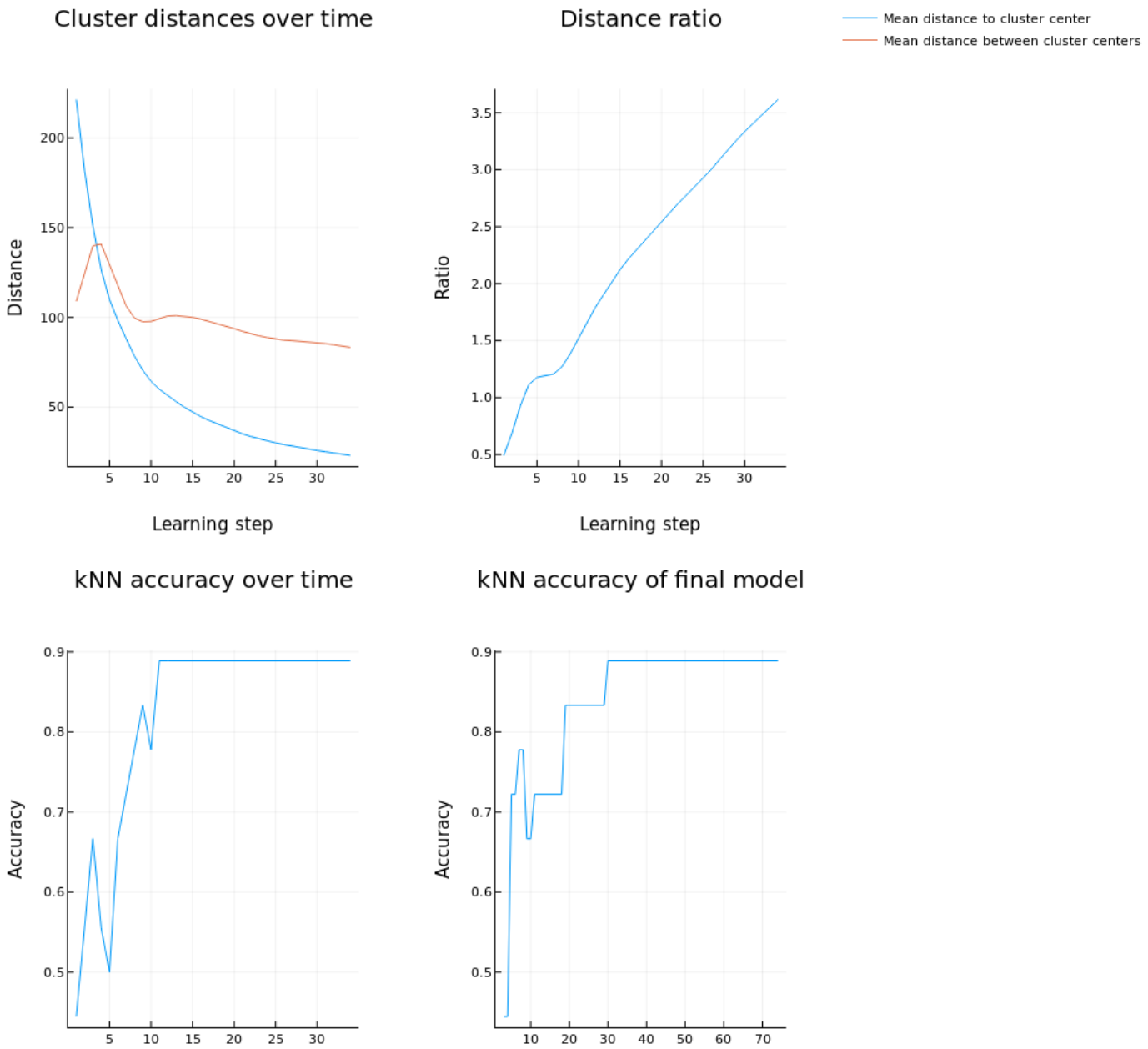
Triplet loss

Triplet loss je alternativou vyžadující supervised přístup

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

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

Triplet loss - předběžné výsledky (Musk2)



Závěr

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

Závěr

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