Kvantitativní analýza internetového provozu (4)

Ladislav Lhotka
⟨*lhotka@cesnet.cz*⟩

Osnova přednášky

- Práce se vzorky datových toků (spojení)
- Analýza mnohorozměrných dat, data mining
- Principal Component Analysis
- Cluster Analysis
- Interaktivní průzkum dat, vizualizace

Vzorky datových toků

Toky představují střední úroveň agregace, na vysokorychlostních linkách ale i tak generují velký objem dat.

Časové parametry sběru toků:

- *inactive timeout* – ukončení toku
- *active timeout* – průběžná data (vhodné pro dlouhotrvající toky)

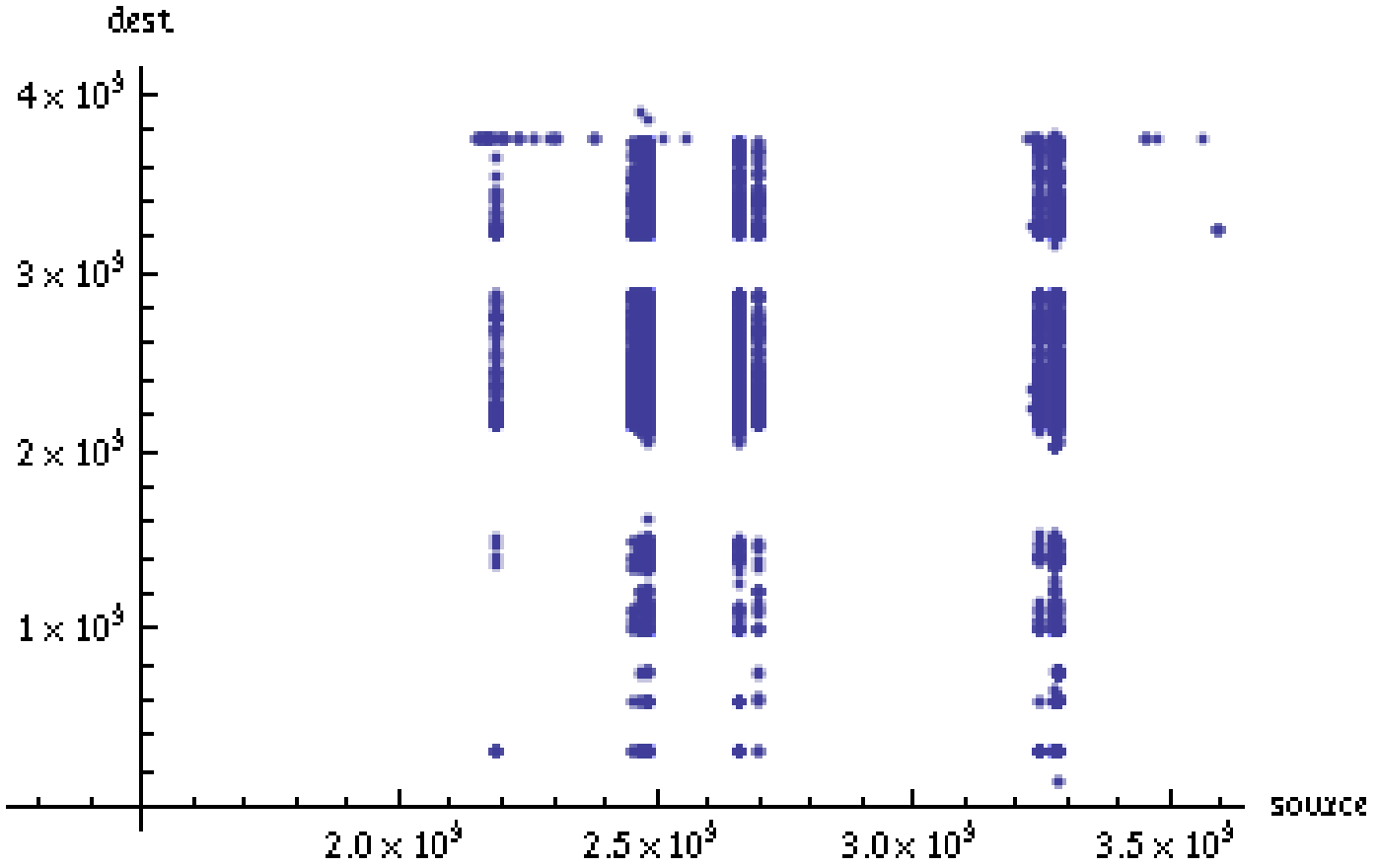
Pětiminutový vzorek dat z mezinárodní linky do GÉANT2 – záznamy o paketech získané pomoci 10 GE karty DAG (Endace).

- 578944960 bajtů
- 6163012 paketů
- 152010 toků (inact. timeout = 10 s)
- 45284 různých párů IP adres

Záznam o paketu

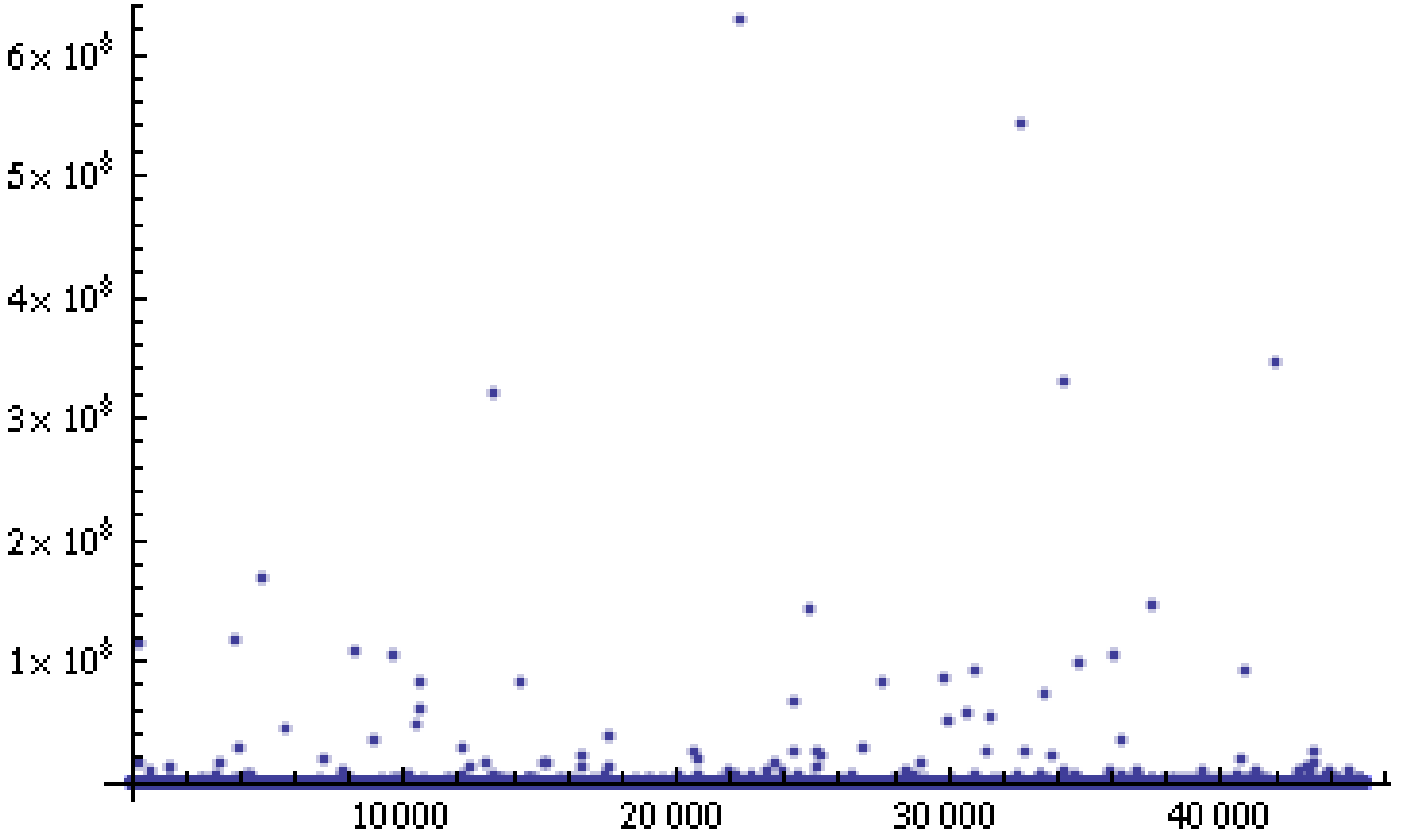
```
print 6163011: file offset 0x2281ff60
ts=0x4603e58b95f3efc0 2007-03-23 14:34:51.5857534 UTC
type: ERF Ethernet
dserror=0 rxerror=0 trunc=0 vlen=1 iface=0 rlen=96 lctr=0 wlen=1438
dst=00:90:69:f5:dd:7a src=00:0e:38:a4:c4:40
etype=0x0800
ip: version=4 headerwords=5 tos=0 length=1420
ip: id=38972 flags=0x2 fragmentoffset=0
ip: ttl=62 protocol=6 checksum=0xa4da
ip: sourceaddress=195.113.232.82
ip: destinaddress=61.14.17.131
tcp: sourceport=80 destinport=41262
tcp: sequence=0xf460818c
tcp: acknowledgement=0x36ac4dd9
tcp: offset=5 control=16 window=8576
tcp: checksum=0x5483 urgent=0
```

Zastoupené IP adresy

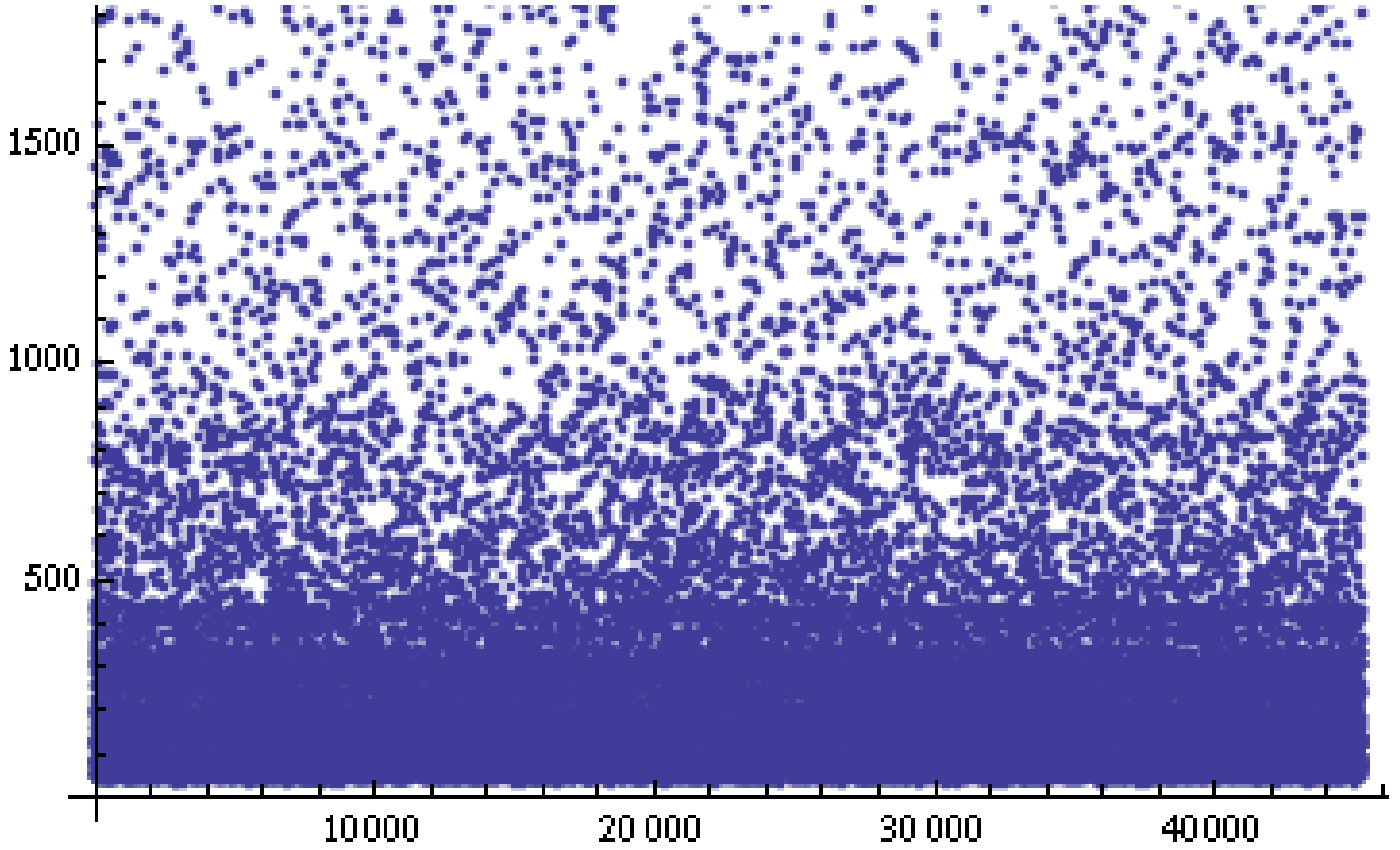


Počty bajtů

Celkový počet bajtů pro konkrétní pár IP adres:



“Čára” poblíž nuly



Vícerozměrná experimentální data

Měříme p parametrů:

$$X = (X_1, X_2, \ldots, X_p)$$

Střední hodnota X_i je μ_i pro $i = 1, 2, \ldots, p$

Vzorek n pozorování:

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{pmatrix}$$

Principal Component Analysis

Cíl: Nálezt vhodnou ortogonální transformaci souřadnic, v níž bude charakter dat co nejzřetelnější, vyniknou "ulétlá" měření (outliers) atd.

První hlavní komponenta: lineární kombinace původních proměnných

$$Y_1 = \alpha_{11}X_1 + \alpha_{21}X_2 + \cdots + \alpha_{p1}X_p$$

Chceme nalézt takové koeficienty, aby Y_1 měla maximální rozptyl.

$$a_1 = \begin{pmatrix} \alpha_{11} \\ \alpha_{21} \\ \vdots \\ \alpha_{p1} \end{pmatrix}$$

Rozptyl lineární kombinace

$$\nu_1 = E(Y_1) = \alpha_{11}\mu_1 + \alpha_{21}\mu_2 + \cdots + \alpha_{p1}\mu_p$$

$$\begin{aligned} D(Y_1) &= E((Y_1 - \nu_1)^2) = E((a_1'(X - \mu))^2) \\ &= E(a_1'(X - \mu)(X - \mu)'a_1) = a_1'\mathbf{C}a_1, \end{aligned}$$

kde $\mathbf{C}$ je kovarianční matice:

$$\mathbf{C}_{ij} = C(X_i, X_j).$$

V praxi používáme odhad kovarianční matice ze vzorku a $a_1'\mathbf{C}a_1$ je pak odhad rozptylu Y_1.

Optimalizační úloha

$$\mathrm{D}(Y_1) = \alpha_1' \mathbf{C} \alpha_1 \longrightarrow \max$$

s normalizační podmínkou $\alpha_1' \alpha_1 = 1$.

Jde tedy o nalezení vázaného extrému, který se řeší metodou Lagrangeových multiplikátorů. Výsledek:

$$\mathbf{C} \alpha_1 = \lambda_1 \alpha_1. \tag{1}$$

Lagrangeův multiplikátor λ_1 je tedy vlastním číslem matice $\mathbf{C}$ a α_1 je příslušný vlastní vektor.

Z rovnice (1) plyne $\mathrm{D}(Y_1) = \alpha_1' \mathbf{C} \alpha_1 = \lambda_1$, takže za λ_1 volíme *největší* vlastní číslo.

Další hlavní komponenty

Druhá hlavní komponenta

$$Y_2 = \alpha_{12}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{p2}X_p$$

$$D(Y_2) = a_2'\mathbf{C}a_2 \longrightarrow \max$$

za podmínek $a_2'a_2 = 1$ a $a_2'a_1 = 0$ (ortogonalita).

Vyjde $\mathbf{C}a_2 = \lambda_2 a_2$, kde λ_2 volíme jako druhé největší vlastní číslo a a_2 je pak příslušný vlastní vektor.

Souřadná soustava hlavních komponent

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1p} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{p1} & \alpha_{p2} & \dots & \alpha_{pp} \end{pmatrix}$$

Transformace souřadnic:

$$Y = \mathbf{A}'X$$

Navíc platí

$$\mathbf{A}'\mathbf{C}\mathbf{A} = \Lambda,$$

kde $\Lambda = \mathsf{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$.

Podíl i-té hlavní komponenty na rozptylu: $\lambda_i / \mathrm{tr}(\mathbf{C})$.

Redukce dimenze

Zvolíme jen několik hlavních komponent, což odpovídá *projekci* do vektorového podprostoru generovaného příslušnou podmnožinou vlastních vektorů.

$$\hat{\mathbf{A}} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1q} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{p1} & \alpha_{p2} & \dots & \alpha_{pq} \end{pmatrix},$$

kde $q < p$. X pak transformujeme na kratší vektor:

$$\hat{Y} = \hat{\mathbf{A}}' X.$$

Cluster analysis

Cluster (shluk) je založen na pojmu blízkosti nebo podobnosti určité skupiny objektů (měření), který je kvantifikován koeficientem nepodobnosti (dissimilarity).

Vlastnosti:

1. $\forall X : d(X, X) = 0$
2. $\forall X, Y : d(X, Y) = d(Y, X) \geq 0$

Speciálním případy je *metrika*, která musí navíc splňovat

$$\forall X, Y, Z : d(X, Y) \leq d(X, Z) + d(Z, Y)$$

nebo *ultrametrika*, pro niž navíc platí:

$$\forall X, Y, Z : d(X, Y) \leq \max(d(X, Z), d(Z, Y)).$$

Vzdálenost mezi clustery

1. *single linkage*

$$\Delta_s(X, Y) = \min\{d(x, y) | x \in X, y \in Y\}$$

2. *complete linkage*

$$\Delta_c(X, Y) = \max\{d(x, y) | x \in X, y \in Y\}$$

3. *average linkage*

$$\Delta_a(X, Y) = \frac{1}{|X||Y|} \sum_{\substack{x \in X \\ y \in Y}} d(x, y)$$

Vlastnosti shlukovacích algoritmů

Podle způsobu vytváření shluků se dělí na *aglomerativní* a *divizivní*.

Podle uspořádání shluků různé velikosti se dělí na *hierarchické* a *nehierarchické*.

Shluky jsou obvykle disjunktní, jinak jde o tzv. *fuzzy clustering*.

Metody shlukování se dělí na *sekvenční* a *simultánní*.

Nejběžnějšími metodami jsou aglomerativní hierarchické – výstuoem je tzv. *dendrogram*.

Algoritmus (naivní)

1. Na začátku tvoří každý objekt svůj vlastní cluster. Vypočítá se matice vzdáleností mezi těmito jednoprvkovými clustery. Položíme $L = 0$ (level).
2. V matici vzdáleností se najde minimální hodnota, odpovídající clusterům C_i a C_j. Položíme $L = \Delta(C_i, C_j)$.
3. Z matice se vyjmou i-tý a j-tý řádek i sloupec a doplní se nový řádek a sloupec pro nový cluster $C_{ij} = C_i \cup C_j$.
4. Je-li matice vzdáleností jednoprvková, pak skonči, jinak jdi na 2.

Interaktivní průzkum dat

Program *GGobi:* `http://www.ggobi.org`

Funguje samostatně nebo ve spojení s R.