Smerom k symbolickej reprezentácii potenciálne nekonečných časových radov

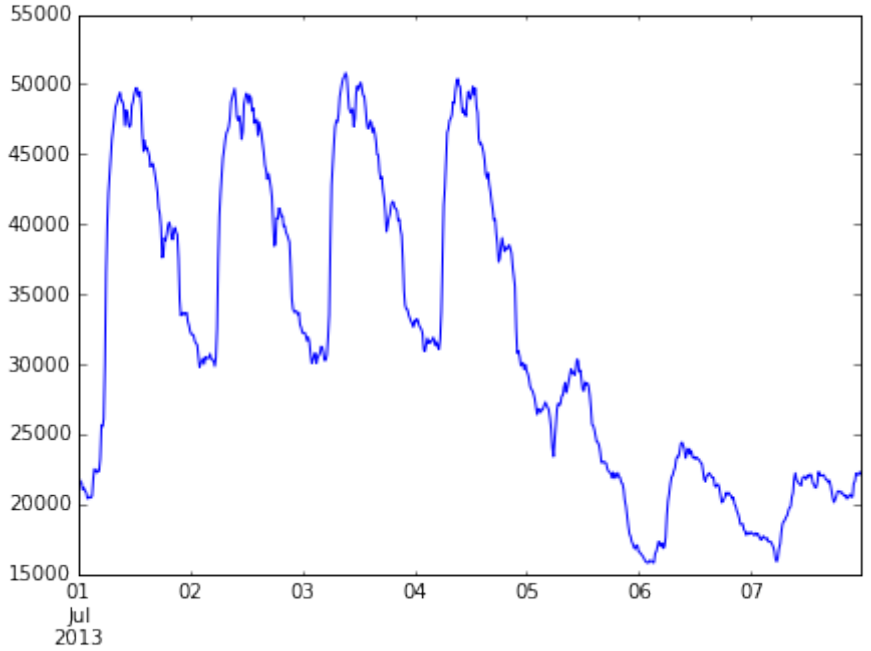
Ing. Jakub Ševcech

Školiteľ: prof. Ing. Mária Bieliková, PhD.

10. novembra 2016
Slovenská technická univerzita v Bratislave

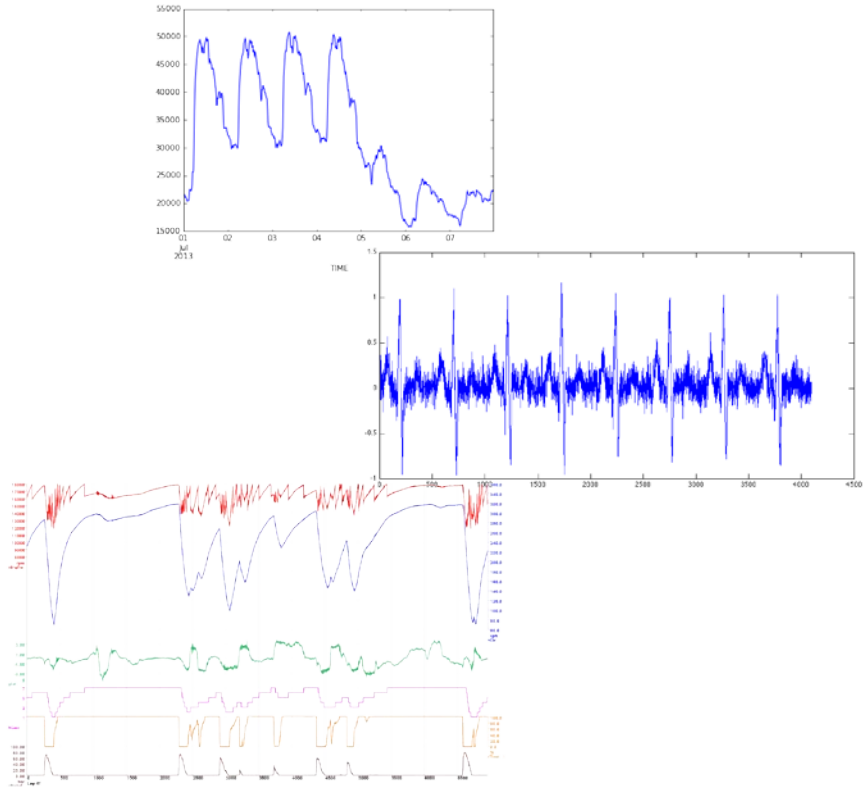
Ciele

1. Spracovávanie veľmi dlhých časových radov (potenciálne nekonečných), kde sa opakuje veľa vzorov
   - Rôzne sezónne a periodické údaje
2. Redukovať dimenzionalitu pomocou opakujúcich sa tvarov
3. Použitie množstva metód z oblasti analýzy textu na časové rady



Výzvy pri porovnávaní časových radov

- **Množstvo metrík podobnosti**, ale len veľmi málo použiteľných na dlhé časové rady
- Zaujíma nás vnútorná **štruktúra**, nie podobnosť hodnôt

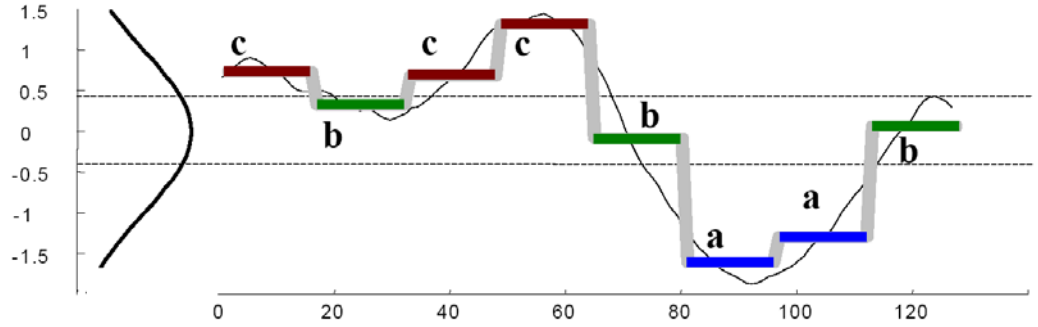


Symbolické reprezentácie časových radov

- Redukcia dimenzionality
- Aproximatívne zachovanie štruktúry
- Umožňujú aplikáciu metód z oblasti spracovania reťazcov a textov

Rôzne spôsoby transformácie časových radov na symboly

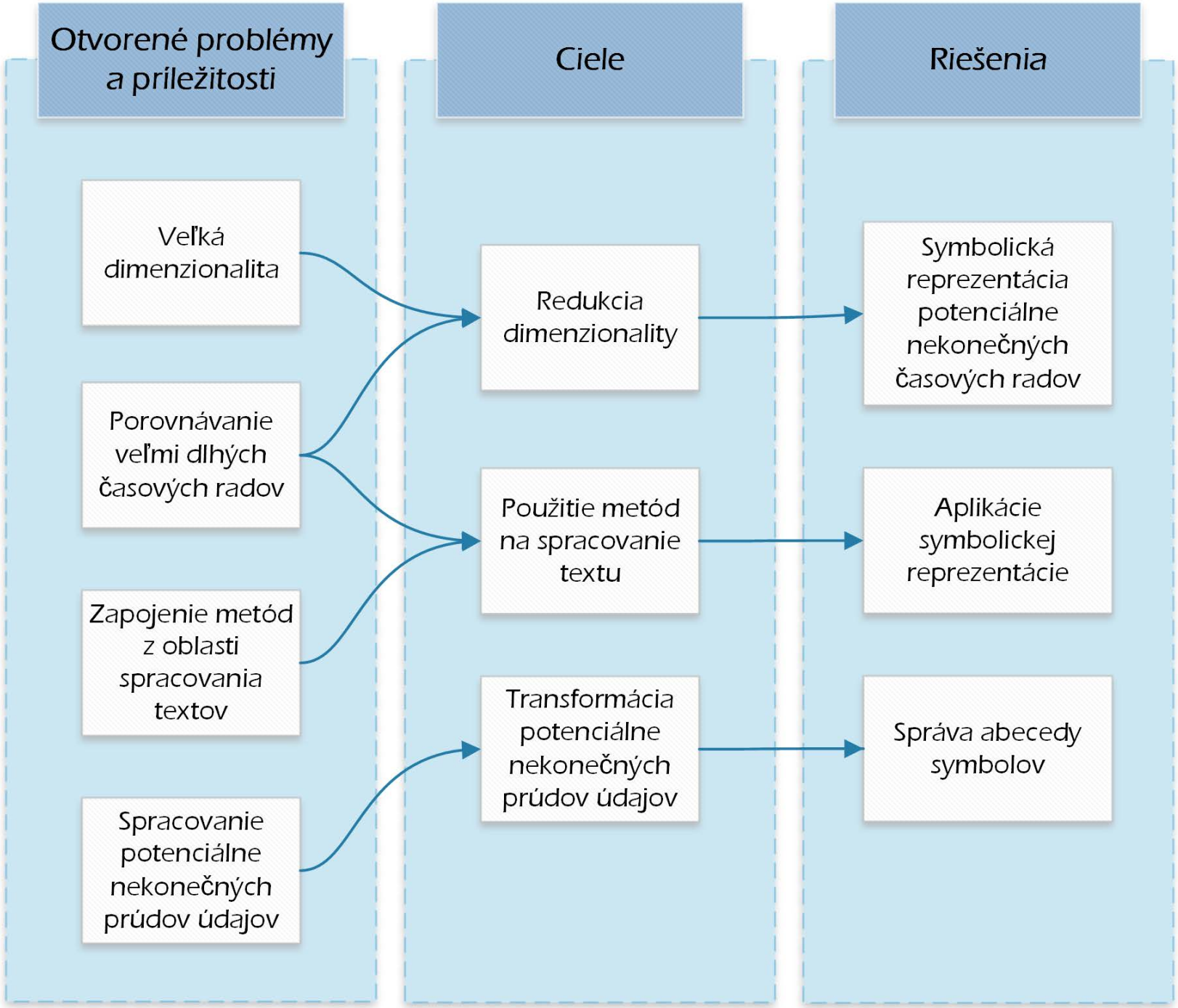
- Zhlukovaním podsekvencií [Das 98]
- Agregovaním hodnôt v oknách [Lin 03]
- Skupina bag-of-words reprezentácií, kde krátke okná sú transformované na slová a ich histogramy reprezentujú celý dokument [Lin 12; Wang 13; Bailly 15]



Obmedzenia spojené so spracovaním prúdu údajov

Obmedzenia spracovania prúdu údajov:

- Prvky prichádzajú postupne premenlivou rýchlosťou
- Nemáme kontrolu nad poradím
- Prúdy sú **potenciálne nekonečné**
- Obmedzené prostriedky na spracovanie a uchovávanie historických údajov.
  - **Obmedzená pamäť** pre vytváraný model
  - **Jediný prechod** údajmi. Konštantný čas spracovania každého prvku.

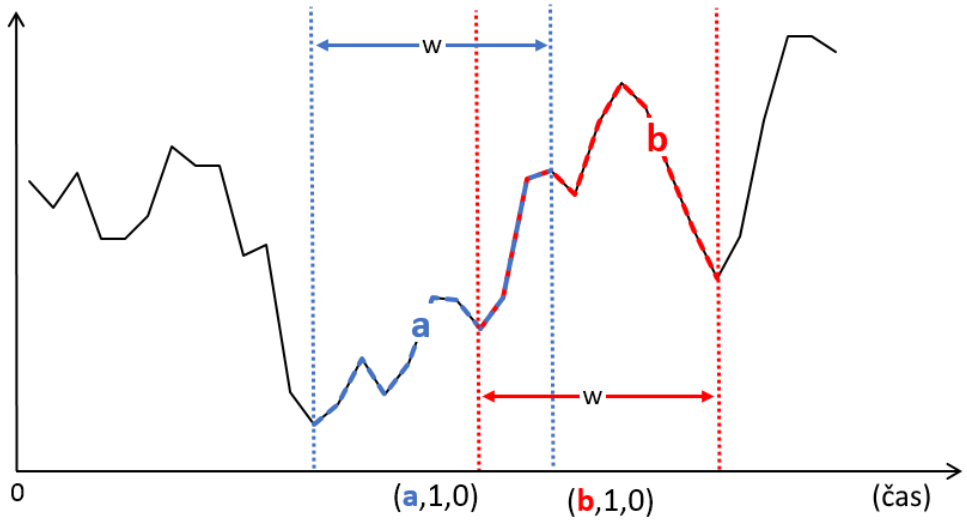

Otvorené problémy a príležitosti
Ciele
Riešenia
Veľká dimenzionalita
Porovnávanie veľmi dlhých časových radov
Zapojenie metód z oblasti spracovania textov
Spracovanie potenciálne nekonečných prúdov údajov
Redukcia dimenzionality
Použitie metód na spracovanie textu
Transformácia potenciálne nekonečných prúdov údajov
Symbolická reprezentácia potenciálne nekonečných časových radov
Aplikácie symbolickej reprezentácie
Správa abecedy symbolov

Reprezentácia časových radov pomocou opakujúcich sa symbolov

Transformácia opakujúcich sa tvarov na symboly

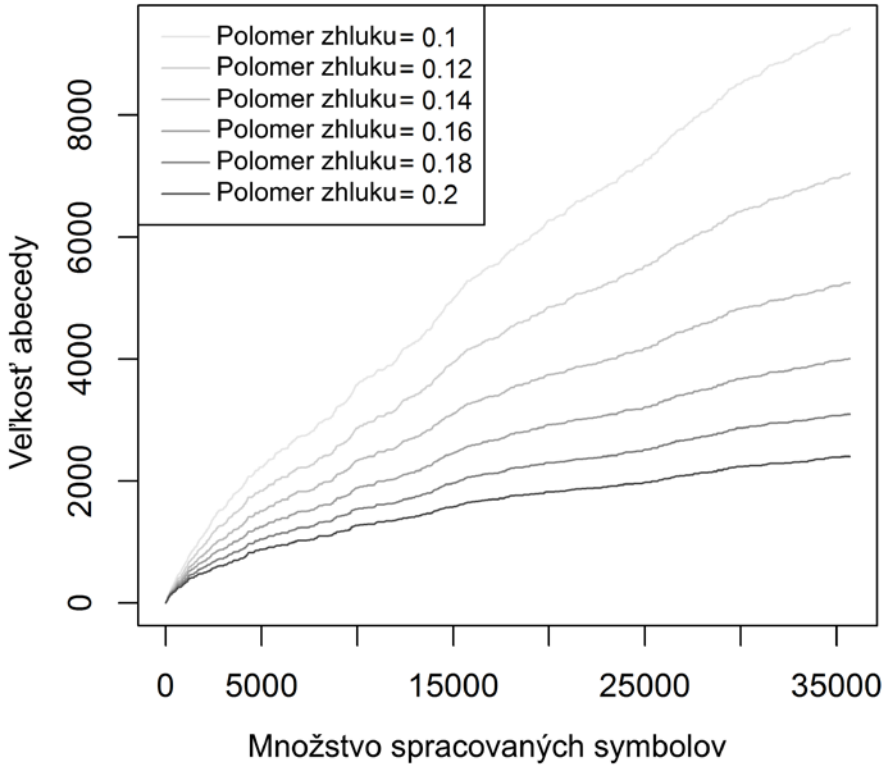
Incremental Subsequence Clustering (ISC)

1. Časový rad rozdelený na prekrývajúce sa sekvencie
2. Normalizácia sekvencií a spojenie do zhlukov
3. Identifikátor zhluku a normalizačné koeficienty použité ako reprezentácia symbolu



Transformácia určená na sezónne údaje, kde sa opakuje viacero symbolov

- Na rozdiel od [Das 98] používame algoritmus Leader pre inkrementálnu transformáciu
  - Počet symbolov nie je vopred obmedzený
  - Keď sa objaví nový tvar, vznikne nový symbol
  - Rýchlosť vzniku nových symbolov s množstvom spracovaných údajov klesá



Metrika podobnosti nad symbolickou reprezentáciou

$$SymD(\hat{Q}, \hat{C}) = \sqrt{\frac{\sum_{i=1}^{\lceil \frac{n-w}{s} \rceil} max(0; ED(\hat{q}_i, \hat{c}_i) - 2r)^2}{\lceil \frac{w}{s} \rceil}}$$

SymD je vlastne Euklidova vzdialenosť symbolov kde:

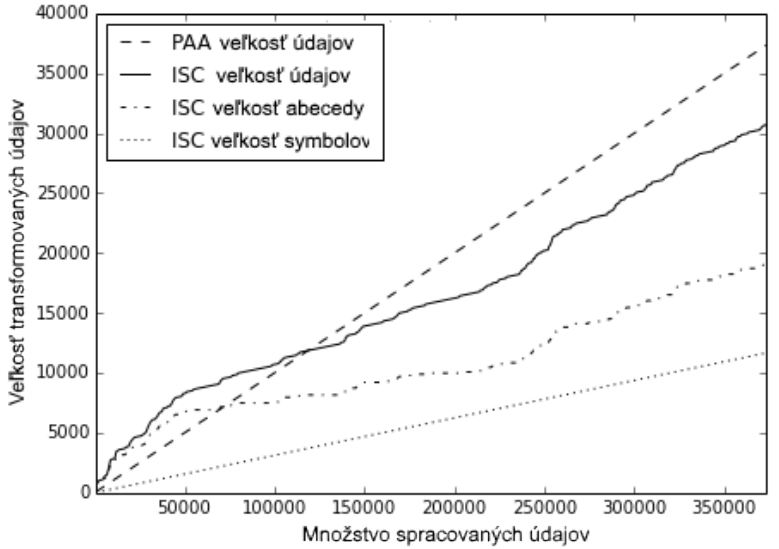
$$max(0; ED(\hat{q}_i, \hat{c}_i) - 2r)$$

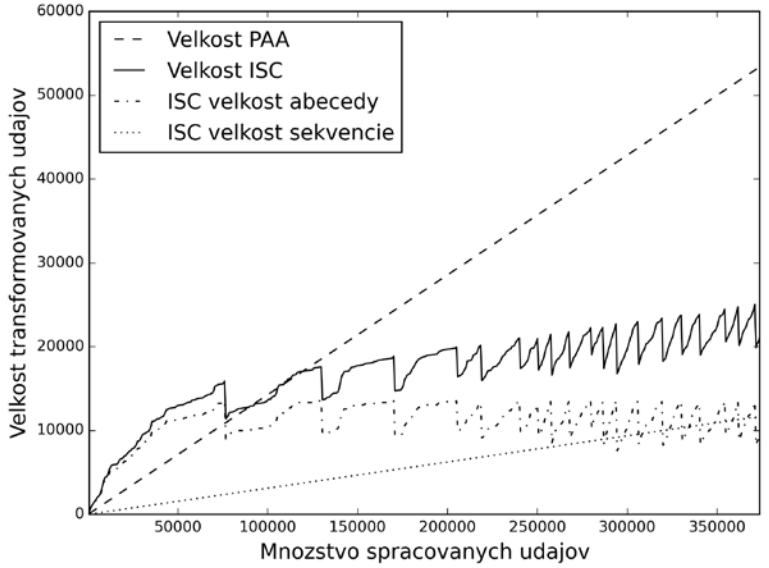
je maximálna vzdialenosť dvoch sekvencií zaradených do zhlukov s korekciou o priemer tohto zhluku

- SymD zdola ohraničuje Euklidovu vzdialenosť
  - Dôkaz uvedený v práci
- Dá sa použiť v indexovacej schéme GEMINI
- Nespĺňa trojuholníkovú nerovnosť a preto sa nedá použiť pri priestorových indexovacích štruktúrach ako napríklad KD-strom
- Dajú sa použiť aj iné metriky podobnosti, pre ktoré však nie sú definované garancie
  - Levenshteinova vzdialenosť
  - Podobnosti histogramov pri použití na tvorbu vektorovej reprezentácie

Problém rastúcej abecedy symbolov pri spracovaní prúdu údajov

- Potreba zabezpečenia konštantného času spracovania prvku v prúde
- 4 prístupy pre zabúdanie málo frekventovaných zhlukov
- Vymeniteľný algoritmus na hľadanie frekventovaných prvkov
- Overenie na veľmi dlhých sezónnych údajoch (10 rokov merania elektrickej spotreby, >300 000 hodnôt)

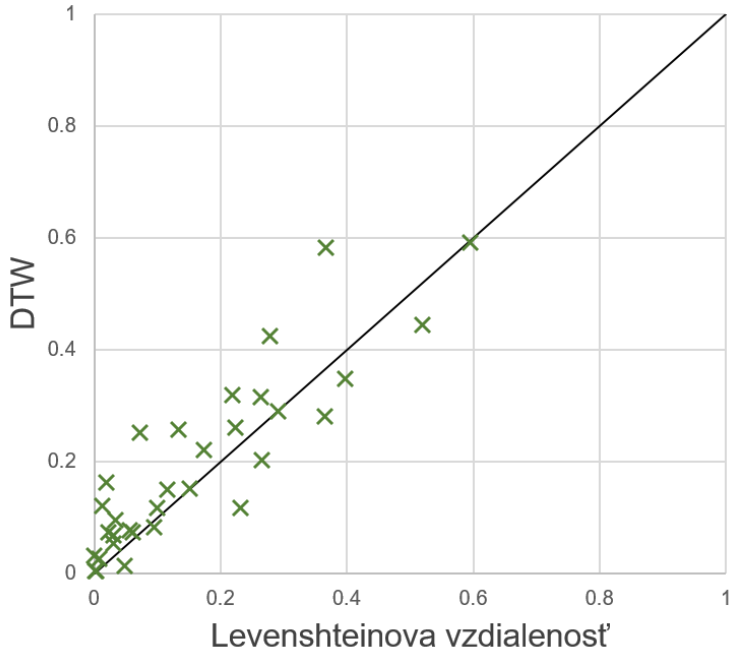


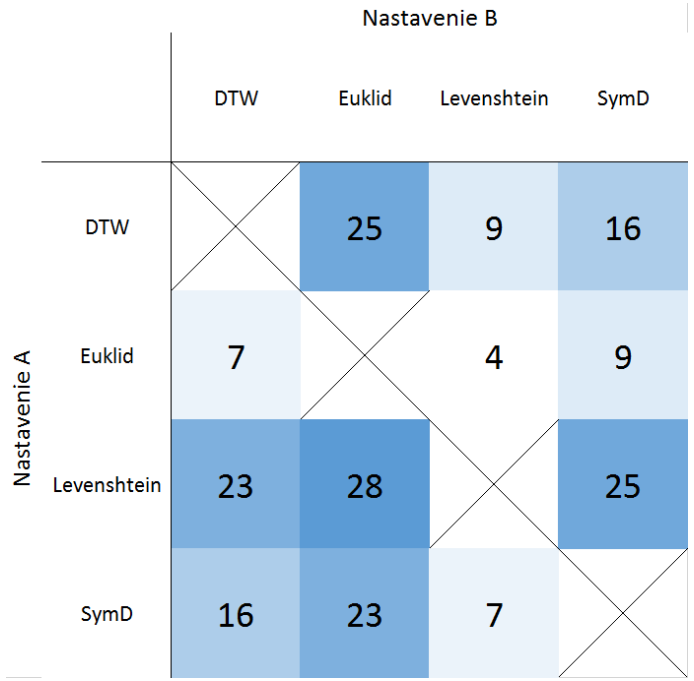


Aplikácie reprezentácie založenej na opakujúcich sa symboloch

Klasifikácia rôznych typov časových radov

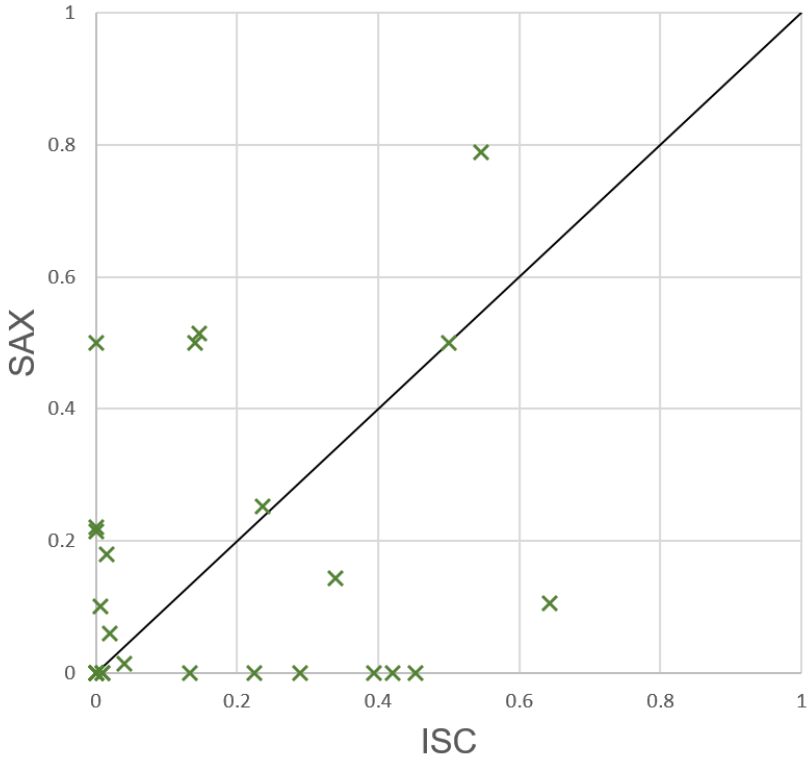
- UCR kolekcia datasetov [Keogh 11]
  - 32 sád s rôznymi charakteristikami
- Overenie vlastností na rôznych typoch časových radov
- Porovnanie klasifikačnej chyby
  - „Surové“ údaje a rôzne metriky podobnosti
  - ISC reprezentácia a SymD alebo Levenshteinova vzdialenosť
- Na viacerých datasetoch bola chybovosť pomocou ISC menšia ako na surových údajoch





Klasifikácia veľmi dlhých časových radov pomocou vektorovej reprezentácie

- ISC sa dá použiť na vytvorenie tzv. bag-of-words reprezentácie.
- Sekvencia symbolov použitá na vytvorenie histogramu
- Porovnanie so symbolmi vytvorenými pomocou SAX
- Syntetické, veľmi dlhé časové rady vytvorené spájaním časových radov z rovnakej triedy v UCR kolekcií datasetov.
- Euklidova vzdialenosť na porovnanie histogramov
- Na viacerých datasetoch sme dokázali zmenšiť klasifikačnú chybu

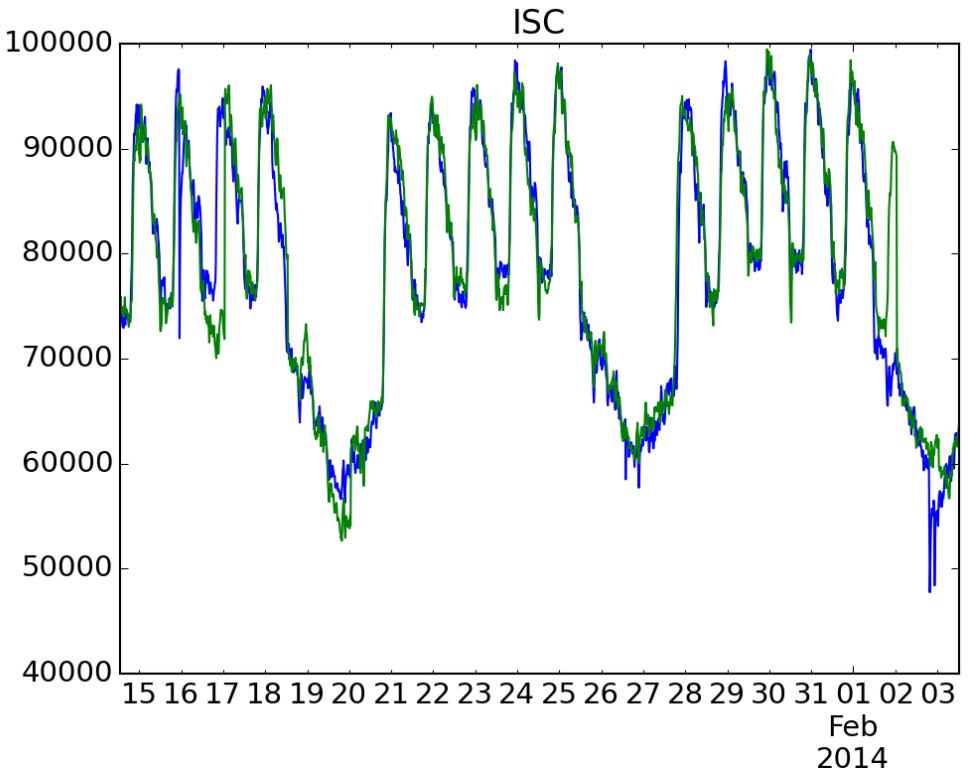


Klasifikácia ďalšieho symbolu vo veľmi dlhých časových radoch

- Ukážka použitia metód, ktoré nie sú priamo použiteľné na numerické hodnoty
  - Klasifikácia rukou písaných znakov
  - Kniha transformovaná do veľmi dlhého časového radu
  - Viac-hodnotové časové rady pre každý znak
  - Symboly rôznych dĺžok
  - Vytváranie Markovovského modelu počas spracovania
- Výsledok
  - Použitie prechodových pravdepodobností pri klasifikácií signifikantne zvýšilo presnosť klasifikácie
  - ISC je použiteľné aj pri viac-hodnotových časových radoch a pri podsekvenciách rôznej dĺžky pomocou extrapolácie alebo elastickej metriky podobnosti

Využitie symbolov pri predpovedaní ďalších hodnôt časového radu

- Použitie ISC na krátkodobú predikciu ďalších hodnôt
  - Porovnávanie poslednej časti spracovávaného časového radu s abecedou vytvorených symbolov
  - Najpodobnejší symbol z abecedy použitý ako predikcia
- Zníženie chyby predikcie a rýchle naučenie nových vzorov
  - Porovnanie s Holt-Winters a Exponenciálnym vyrovnávaním
- Možnosť rozšírenia o učenie zložitejších vzorov, pozeranie sa ďalej do minulosti alebo o zohľadnenie frekventovanosti vzorov



Sumarizácia možných aplikácií symbolickej reprezentácie

- Príklady použitia na klasifikáciu a predikciu ďalších hodnôt
- Krátke aj dlhé časové rady
- Diskretizácia časového radu
- Pomocou vektorovej reprezentácie je možné použiť metódy na spracovanie dokumentov
- Jedno aj viac-hodnotové časové rady
- Rozšíriteľné na použitie pre symboly rôznej dĺžky

Pokračovanie v ďalšom výskume a zhodnotenie

Sumarizácia hlavných prínosov práce

- **Transformácia** časových radov do symbolickej reprezentácie
- **Metrika podobnosti** zdola ohraničujúca Euklidovu vzdialenosť a teda použiteľná pri indexovaní
- Príklady použitia reprezentácie pri **klasifikácii a predikcii** ďalších hodnôt na rôznych typoch údajov
- Reprezentácia sa dá použiť na spracovanie časových radov metódami na **spracovanie reťazcov** znakov a **textov**
- Adaptácia pre použitie pri spracovaní **prúdov údajov**
- Výsledky jadra dizertačnej práce publikované na viacerých zahraničných konferenciách a v **CC časopise**.

Ševcech, Jakub, and Mária Bieliková. "Repeating Patterns as Symbols for Long Time Series Representation." *Journal of Systems and Software,* Elsevier (2016), Indexed in Current Contents, Impact Factor 1.352.

Pokračovanie v ďalšom výskume

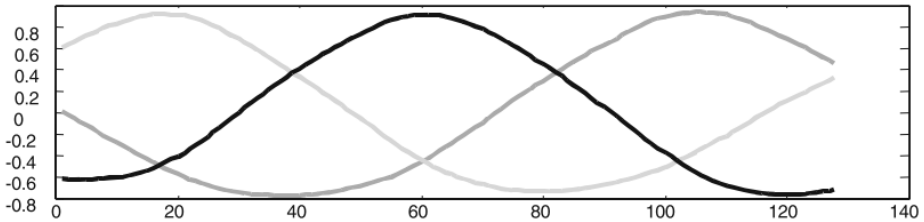
- Overili sme použiteľnosť pri spracovávaní jedného časového radu, na jednom počítači
  - Jedna abeceda na viacerých prúdoch údajov
  - Udržiavanie abecedy v distribuovanom prostredí
- Špecifiká prepojenia symbolickej reprezentácie s metódami na spracovanie textu
  - Rýchle vyhľadávanie vo veľkých kolekciách dlhých časových radov
- Nastavenie jednej veľkosti symbolu obmedzuje doménu
  - Použitie symbolov rôznej dĺžky
  - Časové rady, kde dĺžka periódy nie je definovaná len časom, ale nejakým externým faktorom
  - Napríklad ECG signály kde sa opakujúce symboly naťahujú s tepom

Inkrementálnym zhlukovaním sa vytvárajú zmysluplné zhluky

Zmena zhlukovacieho algoritmu oproti [Das 98] umožnila:

- Inkrementálnu transformáciu
- Tvorbu zmysluplných zhlukov [Keogh 04]

$$Zmysluplnosť(X,Y) = \frac{Stredná\ vzdialenoť(X)}{Stredná\ vzdialenosť\ medzi(X,Y)}$$



- Veľké hodnoty znamenajú, že tvar centra zhluku nezávisí od spracovaných údajov.
- Pri veľmi malých dĺžkach podsekvencií, reprezentujú zhluky len zopár základných tvarov