Unsupervised Algorithm for Retrieving Characteristic Patterns from Time-Warped Data Collections

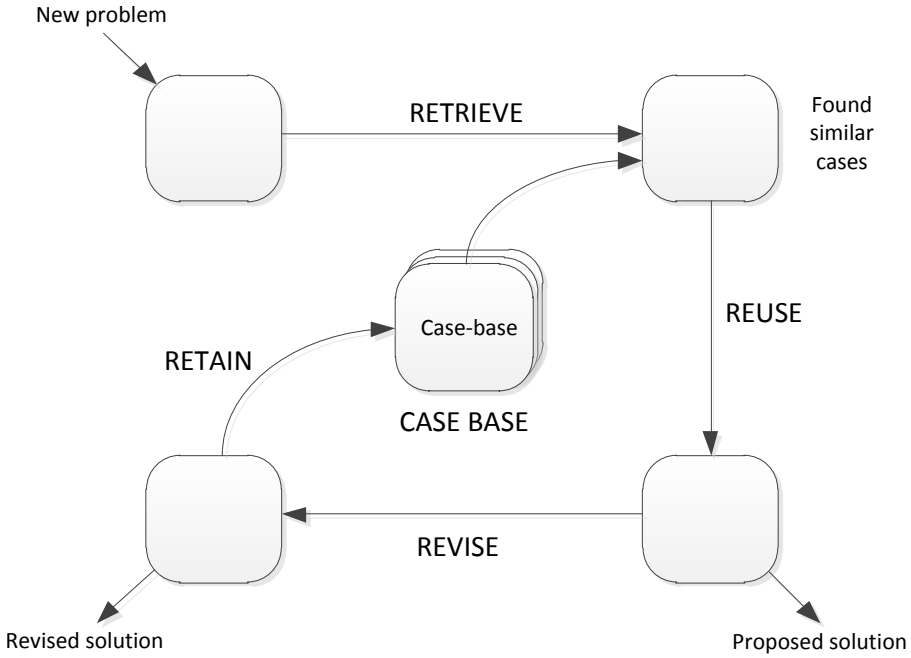
Tomáš Kocyan, Jan Martinovič, Michal Podhorányi, Ivo Vondrák

Searching in Data Collection

- Many types of existing collections often contain repeating sequences which could be called as patterns
  - They can be used for instance in data compression or for prediction
- Extraction of these patterns from data collections with components generated in equidistant time and in finite number of levels is now a trivial task
  - The problem arises for distorted collections
- Focus on processing of measured river discharge volume
  - Looking for typical patterns in this data collection
  - Future research – patterns will be used for simulation of the rainfall runoff process and for prediction of discharge volumes in basin's outlet cross section via Case-Based Reasoning

Case-Based Reasoning

- CBR belongs to a group of artificial intelligence methods
- Process of solving new problems based on the solutions of similar past problems



Case-Based Reasoning

- For achieving the best results it is necessary to Retrieve the most similar cases
- Many supervised and unsupervised methods for looking for patterns and similar situations
- Often cannot handle searching for patterns of different lengths and they are not resistant to distortion

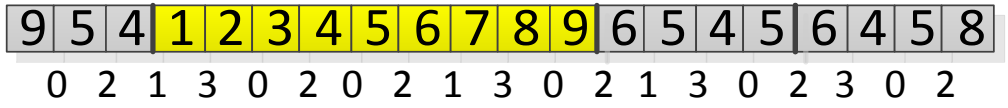
Voting Experts

- Domain-independent algorithm for segmenting categorical time series
  - Into meaningful episodes
  - Unsupervised learning

- Basic VE idea based on simple hypotesis:
  - Typical patterns found in data collection are followed by two statistical indicators:
    - Low internal entropy inside these patterns
    - High boundary entropy on pattern boundaries

Voting Experts – the algorithm

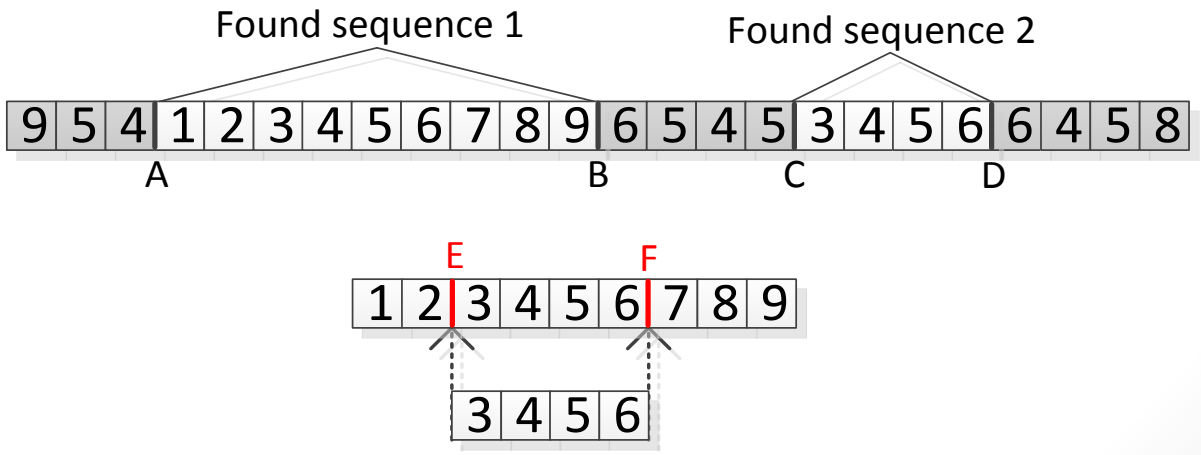
1. Build an nGram tree from the input
   - Calculate statistics for each node of this tree (internal and boundary entropy)
   - Standardize these values in nodes at the same depth
2. Pass a sliding window of length *n* over the input and let experts vote
   - Each of the experts has its own point of view on current context
   - Experts vote for the best location for the split
   - Usualy two experts:
     1. Expert votes for locations with the highest boundary entropy
     2. Expert votes for locations with a minimal sum of internal spit entropy
3. Look for local maximums which overcome selected threshold.
   - These points are adepts for a split of sequence.



Voting Experts

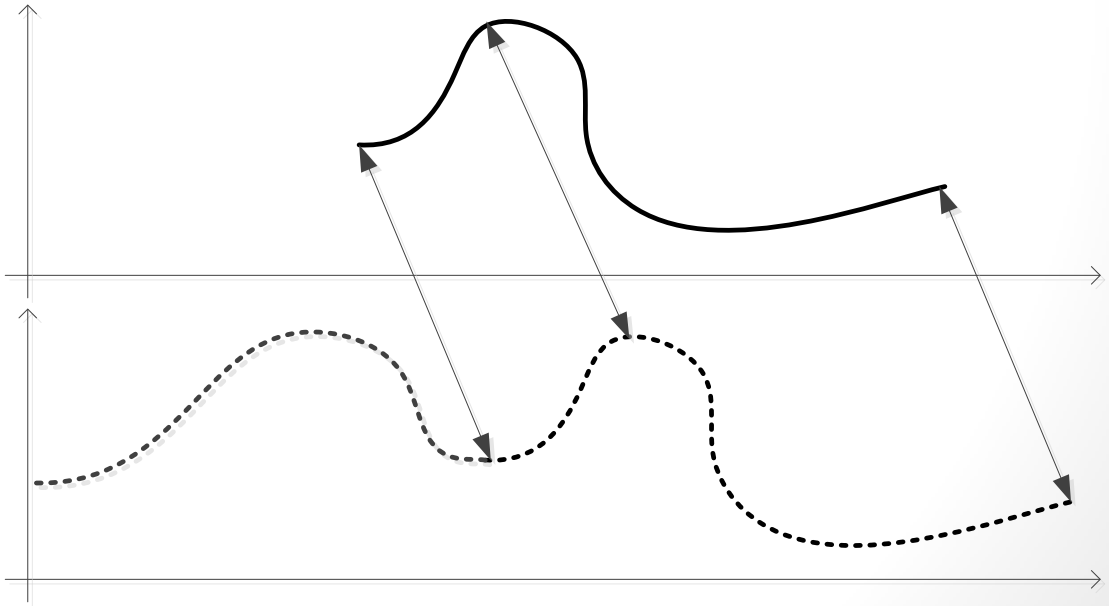
- Several ways how to improve the basic Voting Experts algorithm
  - Custom expert can be added to voting process
  - Methods based on repeated or hierarchical segmenting of the input

- Own improvement – postprocessing of high precision cuts
  - Two-way voting with high threshold
  - DTW post-processing the output



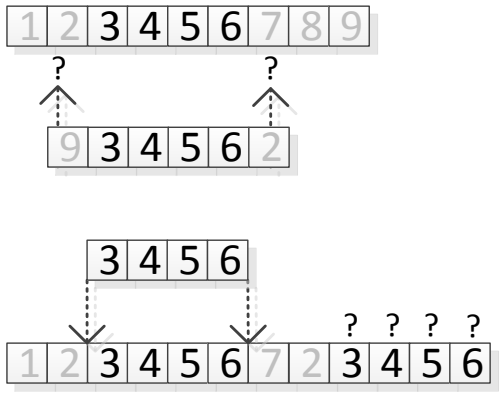
Dynamic Time Warping

- Technique to find an optimal alignment between two given sequences under certain restrictions
- Sequences are warped in a nonlinear fashion to match each other
- Searching subsequences
- Mapping cost can be quantified



Dynamic Time Warping

- Works perfectly in a case of searching exact pattern
- In real situations - exact patterns are not available
  - Surrounded by additional values
  - Repeated several times in the sequence



- Own DTW modification was created

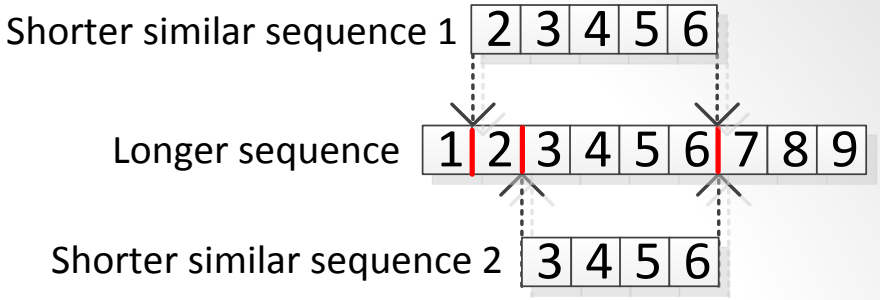
Post-Process Algorithm

1. First of all, the high precision (but not complete) cuts are created by splitting the input with high level of threshold by the Two-Way Voting Experts method.

2. Let's suppose that there are *m* unique sequences which have been created according to cuts from step 1.

3. A $m \times m$ distance matrix is build.

Voting Experts



4. For each pair in this matrix, where the length of sequence s_1 is bigger than length of sequence s_2:
   a) The optimal mapping of shorter sequence s_2 to longer sequence s_1 is found by using DTW modified for searching subsequences.
   b) If the mapping cost does not overcome selected threshold, the longest sequence s_1 stores the shorter sequence s_2 into its own list of similar sequences.
   c) Each of the shorter sequences points to positions in the longer sequence, where it should be split. Because there are usually more than one similar shorter sequences, it is pointed to several locations whereas many of these locations are duplicated. For this reason, the votes are collected into internal vote storage.
   d) After these votes are collected, the local maximums are detected. These places are suggested as new cuts in original input.

Voting Experts

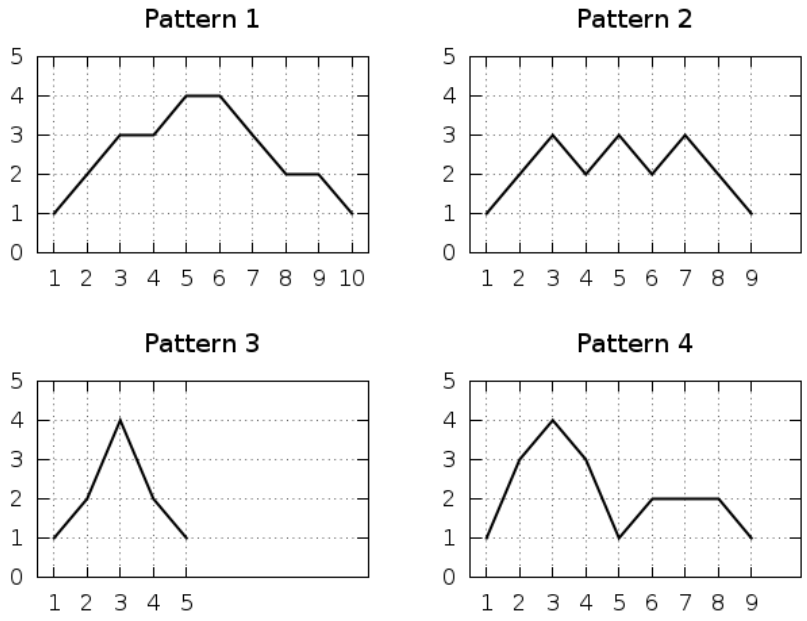
5. The granted votes from step 4d are summed with votes of frequency and entropy experts in the input.
   - The local maximums of votes are searched again.
   - The cuts are made in locations where the number of granted votes is higher than the specified threshold.

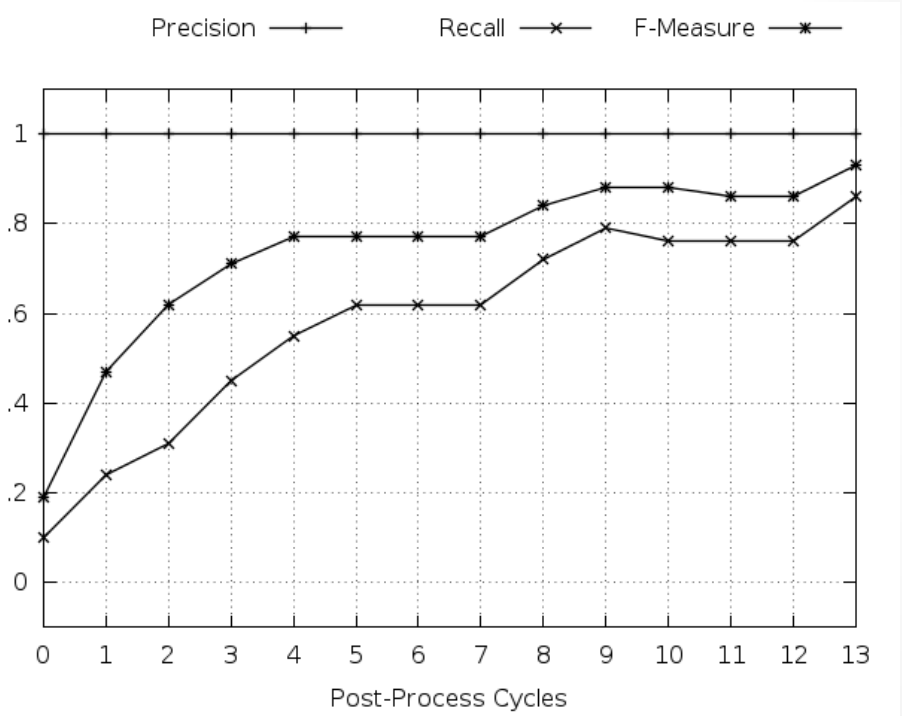
6. Algorithm ends or it can continue with step 2 for further refinement.

Experiments

- Typical test of VE – searching words in continuous text
  - Georgre Orwell – 1984
  - Spaces and punctuations (dots, dashes, new lines etc.) are removed
  - Goal of the algorithm is to put spaces back into correct places
  - Correct placement is known – very easy to quantify accuracy
- Applied on various texts
- Solution overcame almost all monitored quality indicators
  - Recall improved by up to 18%
  - F-measure about 5.5% in average

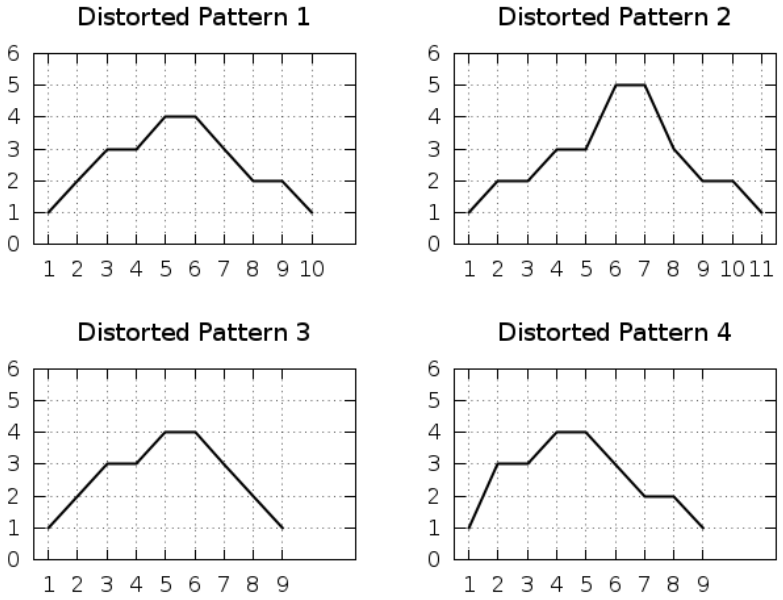
Experiments – Artificial collection





Algorithm	Precision	Recall	F-Measure
Basic Voting Experts	0.71	0.68	0.70
VE with Post-Process	1	0.86	0.93

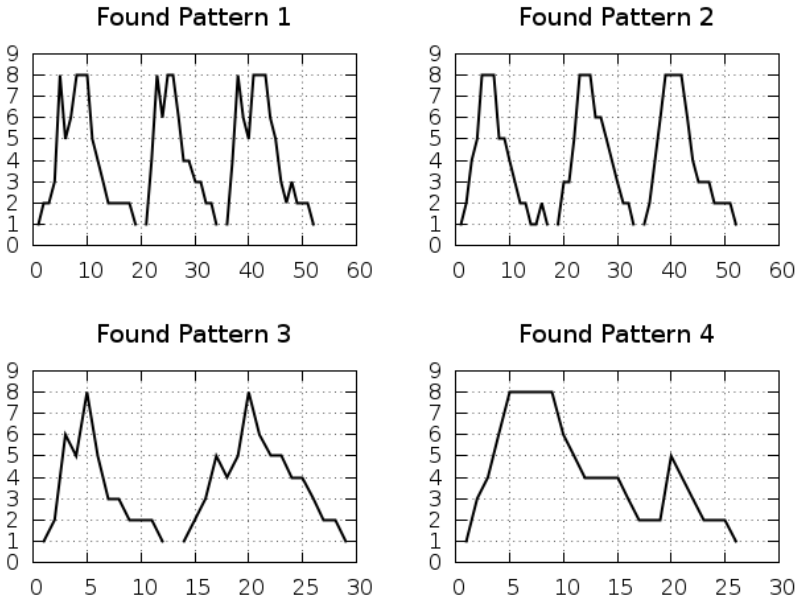
Experiments – Distorted collection



- Applied distortion randomly manipulated with:
  - Length of patterns
  - Amplitude of patterns

Algorithm	Precision	Recall	F-Measure
Basic Voting Experts	0.48	0.58	0.53
VE with Post-Process	0.91	0.72	0.81

Experiments – Real collection



- River database
  - USGS (U.S. Geological Survey) Water Data for the Nation
  - Discharge data from stations located on the main rivers in the U.S.A.
  - Years from 1986 to 2007
  - 30 minute step
- Data was encoded by SAX
  - (Symbolic Aggregate approXimation)

Conclusion and future work

- Proposed solution overcomes qualitative indicators of original VE
- Offers different point of view to the searching patterns
- Future research will be focused on:
  - Optimizing and improving proposed algorithm's performance
  - Automatic settings of configuration's parameters
  - Searching the universal encoding algorithms for transforming general time series

Thank you for your attention