Speeding up the Classification of Biomedical Signals via Instance Selection

Krisztian Buza[1], Julia Koller[2]

[1] University of Warsaw, Poland, Budapest University of Technology and Economics, Hungary, chrisbuza@yahoo.com
[2] University of Debrecen, Hungary, jkoller4@gmail.com

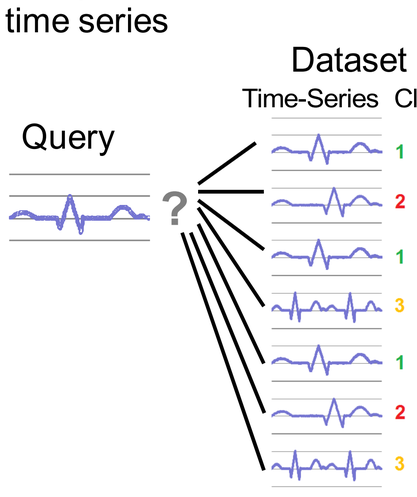
1. Background

- Time-series classification is the common theoretical background of various recognition and prediction problems associated with biomedical signals, such as ECG and EEG, e.g. reduction of brake distance of cars, detection of heart diseases

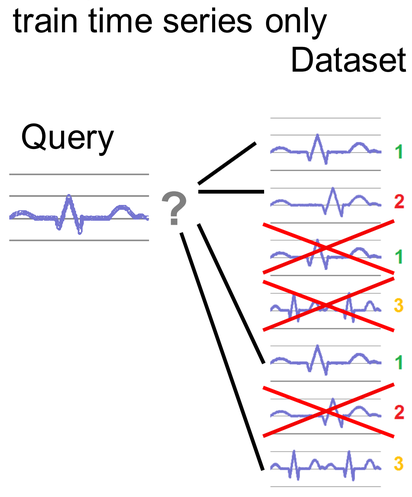
- We aim at solving such problems automatically. Approach: nearest neighbor models with dynamic time warping (DTW)

2. Speeding up nearest neighbor classification by instance selection

Standard nearest neighbor: Comparison to ***all*** train time series



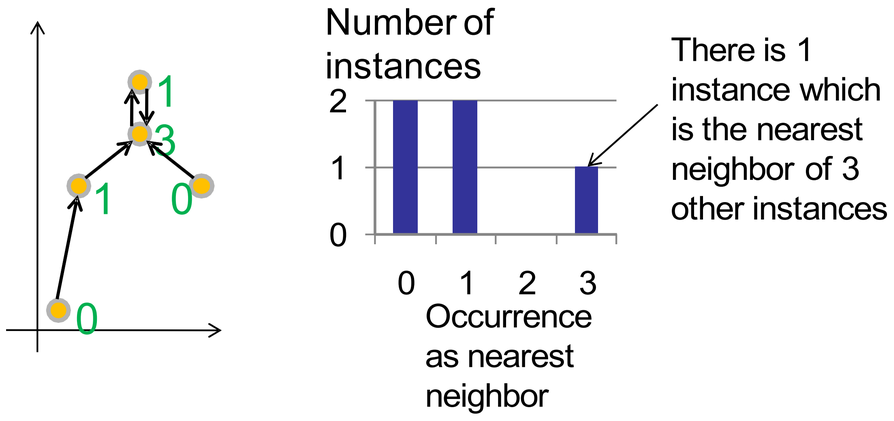
With instance selection: Comparison to the ***selected*** train time series only



3. Good and bad neighbors, presence of good and bad hubs

Instance *y* is a good (bad) *k*-nearest neighbor of the instance *x* if
(i) *y* is one of the *k*-nearest neighbors of *x*, and
(ii) both have the same (different) class labels.

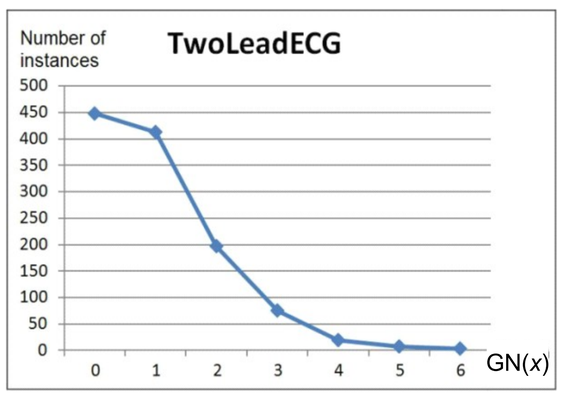
Simplified example (in vector space)

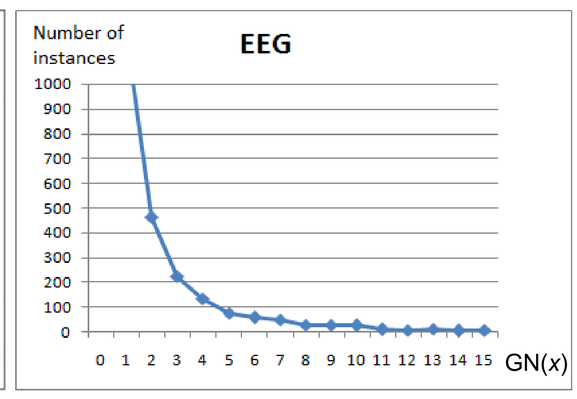


Good (bad) hub: an instance which appears frequently as good (bad) nearest neighbor of the other instances.

GN(*x*), BN(*x*) = how many times instance *x* appears as good/bad nearest neighbors of other instances

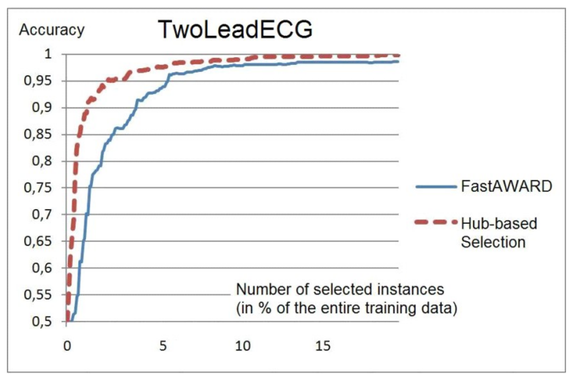
4. Hubs in databases of real biomedical signals

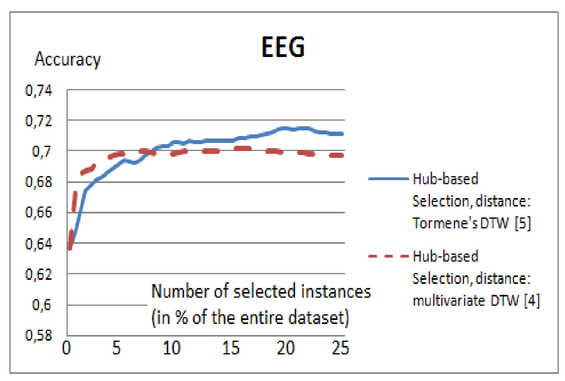




5. Our approach

Rank instances using a hubness-based score, such as GN(*x*) or GN(*x*) – 2BN(*x*), and select the top-ranked instances





- **TwoLeadECG**: from the UCR time series repository, hub-based selection according to GN(*x*), for FastAward see [6]
- **EEG data**: from UCI machine learning repository, hub-based selection according to GN(*x*) – 2 BN(*x*), for Tormene's DTW [5] we normalized the time-series

Acknowledgement – We acknowledge the DAAD-MÖB researcher exchange program (grant no. 39859).

References

[1] Buza, K., Nanopoulos, A., Schmidt-Thieme L., Koller J. (2011): *Fast Classification of Electrocardiograph Signals via Instance Selection*, First IEEE conference on Healthcare Informatics, Imaging and Systems Biology (HISB)
[2] Buza, K., Nanopoulos, A., Schmidt-Thieme, L. (2011): *INSIGHT: Efficient and Effective Instance Selection for Time-Series Classification*, PAKDD, LNCS, Vol. 6635, Springer
[3] Radovanović, M., Nanopoulos, A., Ivanović, M. (2009): *Nearest neighbors in high-dimensional data: The emergence and influence of hubs*, 26th International Conference on Machine Learning (ICML'09), pp. 865-872
[4] K. Buza (2011): *Fusion Methods for Time-Series Classification*, Peter Lang Verlag
[5] Tormene, P., Giorgino, T., Quaglini, S., Stefanelli, M. (2009): *Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation*, Artificial Intelligence in Medicine, Vol. 45, Issue 1, pages 11-34
[6] Xi, X., Keogh, E., Shelton, C., Wei, L., Ratanamahatana, C.A. (2007): *Fast Time Series Classification Using Numerosity Reduction,* ICML 2006. LNCS, Vol. 4503, Springer