ENSEMBLE DECOMPOSITION LEARNING FOR OPTIMAL UTILIZATION OF IMPLICITLY ENCODED KNOWLEDGE IN BIOMEDICAL APPLICATIONS

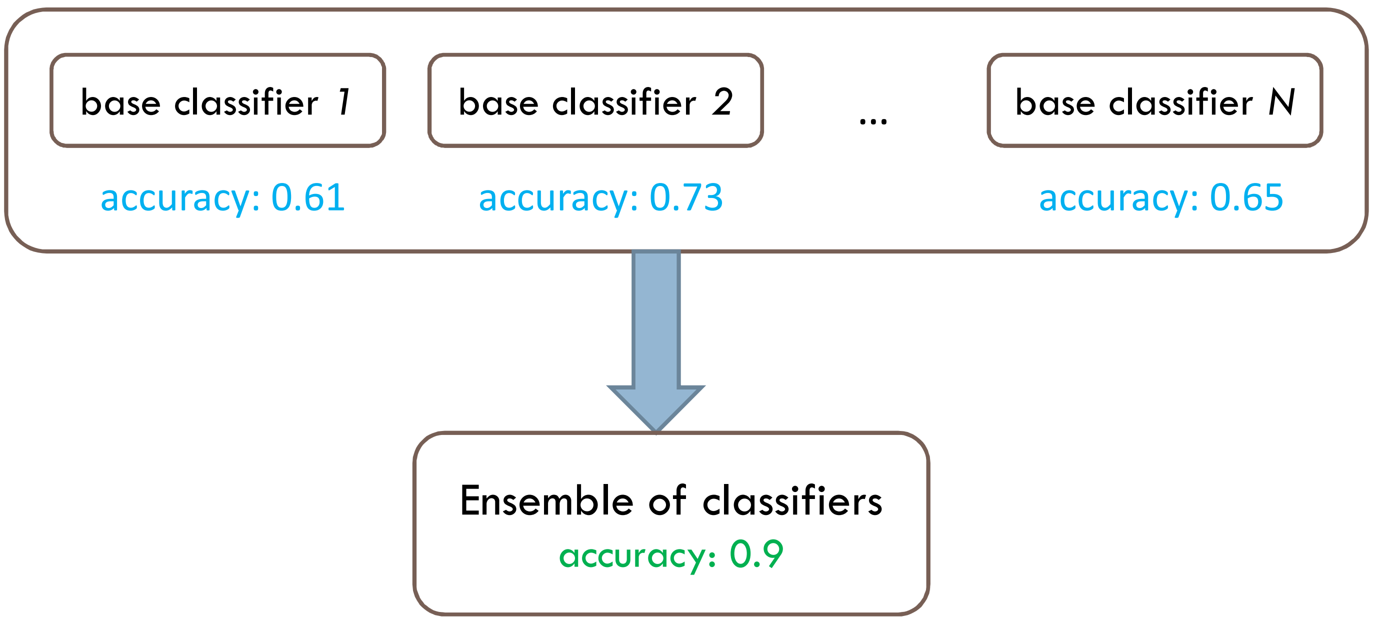
Olga V Senyukova[1], Valeriy V Gavrishchaka[2]

[1] Department of Computational Mathematics and Cybernetics, Lomonosov Moscow State University,

[2] Department of Physics, West Virginia University

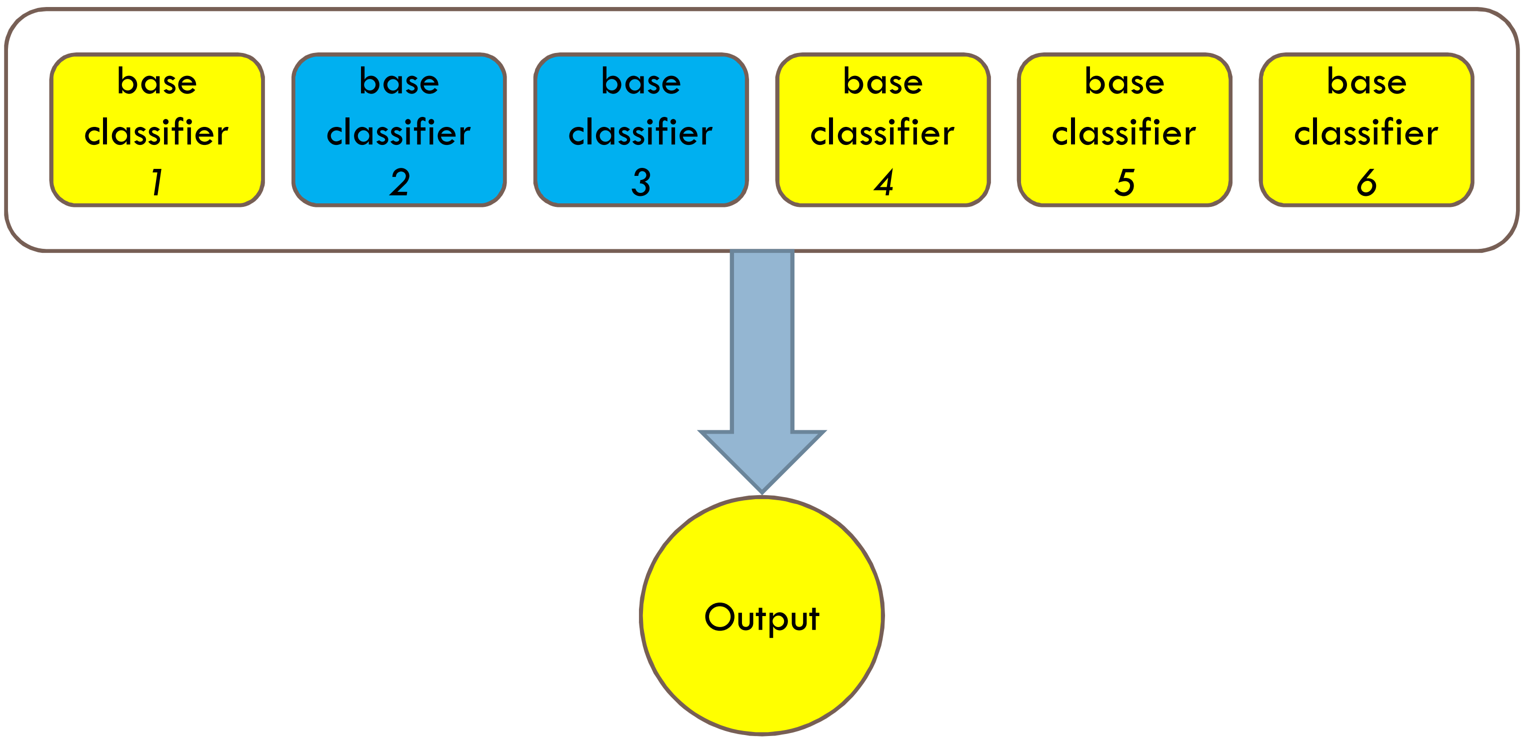
Ensemble Learning Techniques

Ensemble learning techniques can improve performance of other domain-specific base classifier models



Combination of Multiple Base Classifiers

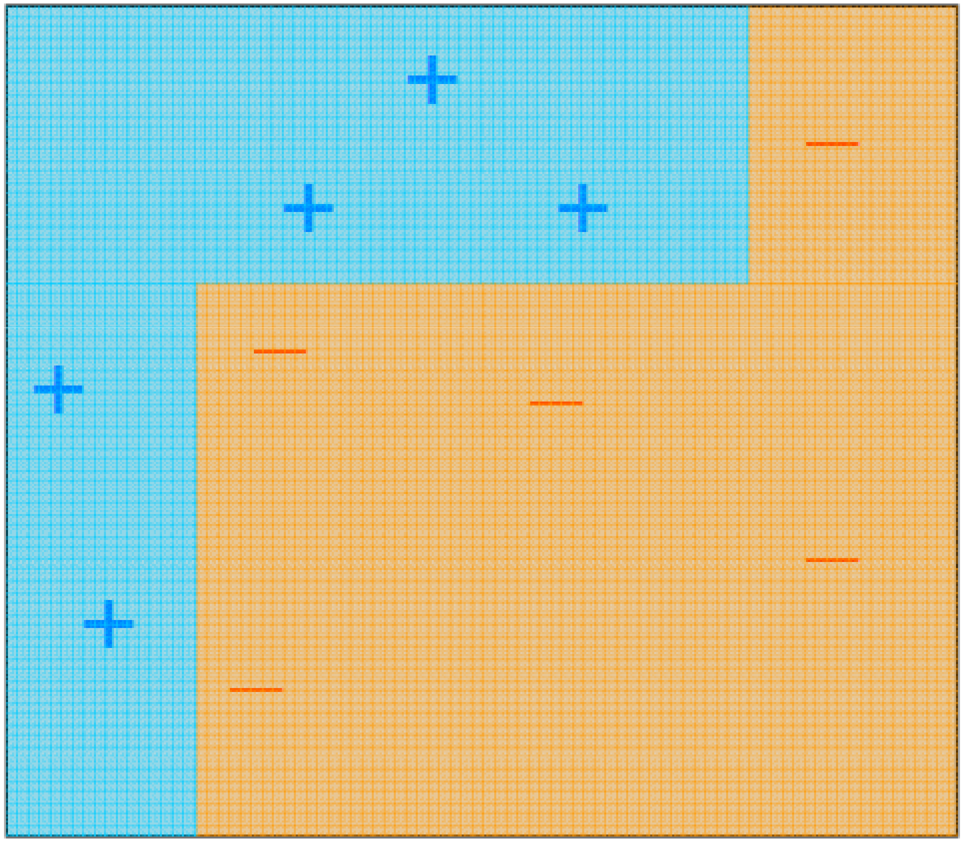
In most of the current ensemble methods the final decision is based on aggregated output of the ensemble – majority voting



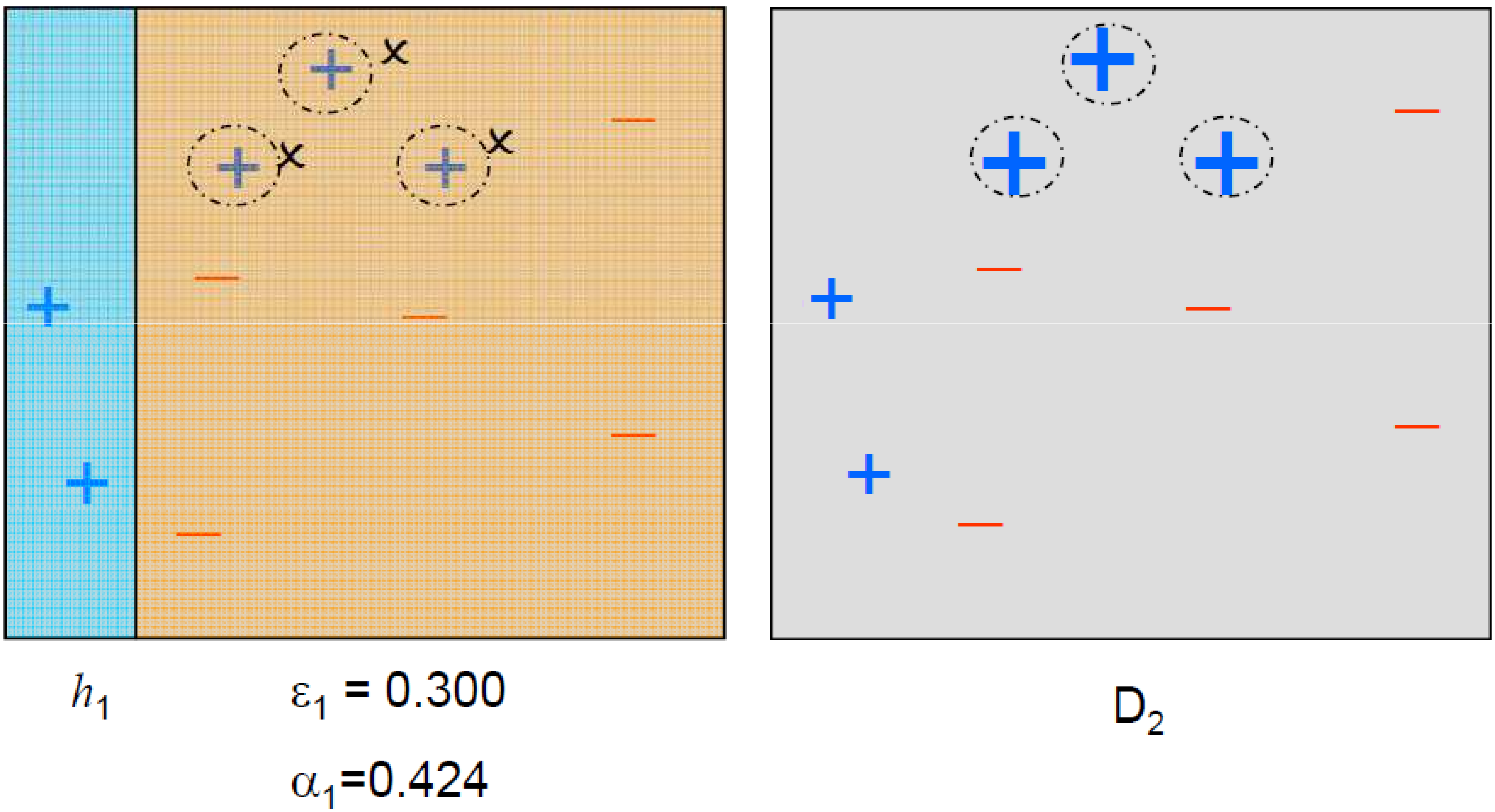
Boosting

- Boosting-like techniques are capable of reducing both variance and bias parts of the model error
- Base classifiers (i.e. *weak learners*) can be just slightly better then random, while the accuracy of the final ensemble could be much higher
- Boosting provides robust and practical means for significant performance improvement of existing domain-specific models
- There are several learning stages (*boosting rounds)*
  - on every stage we choose a base classifier that demonstrated the best performance on samples, which are the most difficult for previous base classifiers
  - measure of difficulty is fixed by weights assigned to samples from the training set
  - the resulting classifier is a linear combination of base classifiers

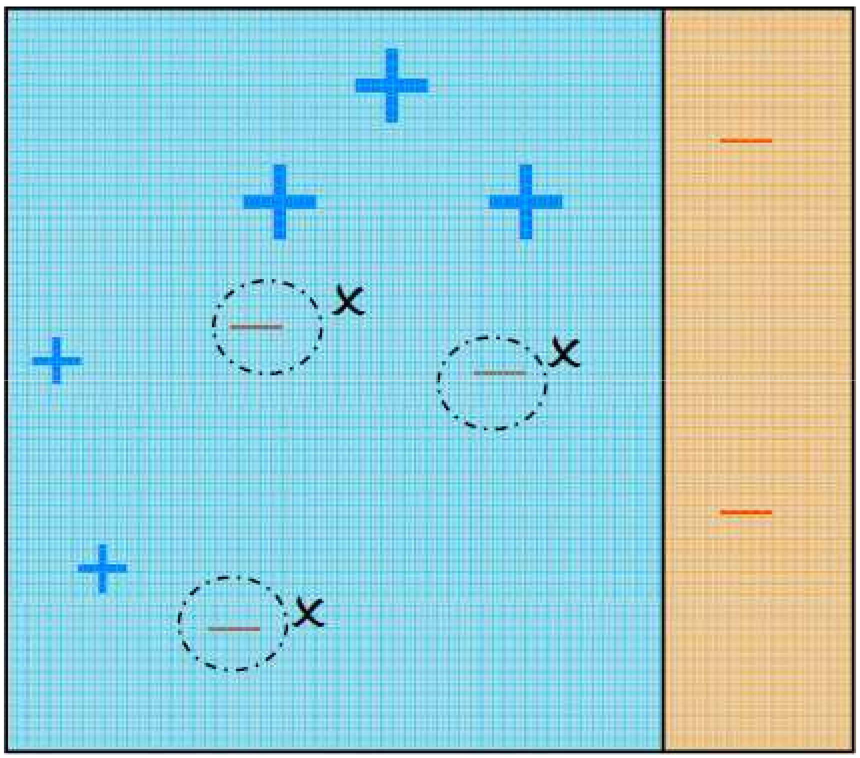
A Good Classifier Example



Iteration 1 of 3

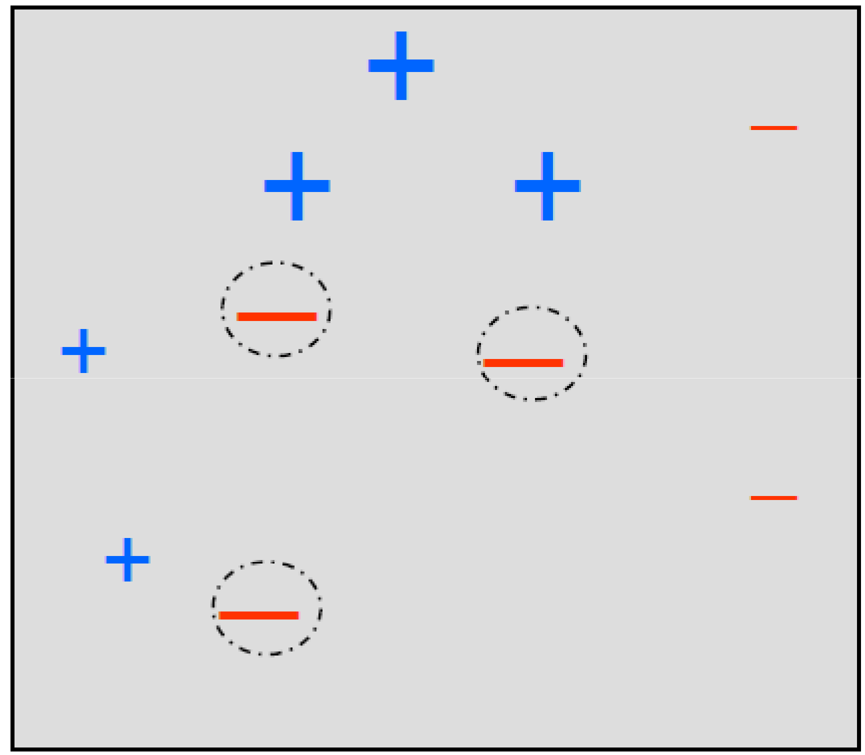


Iteration 2 of 3



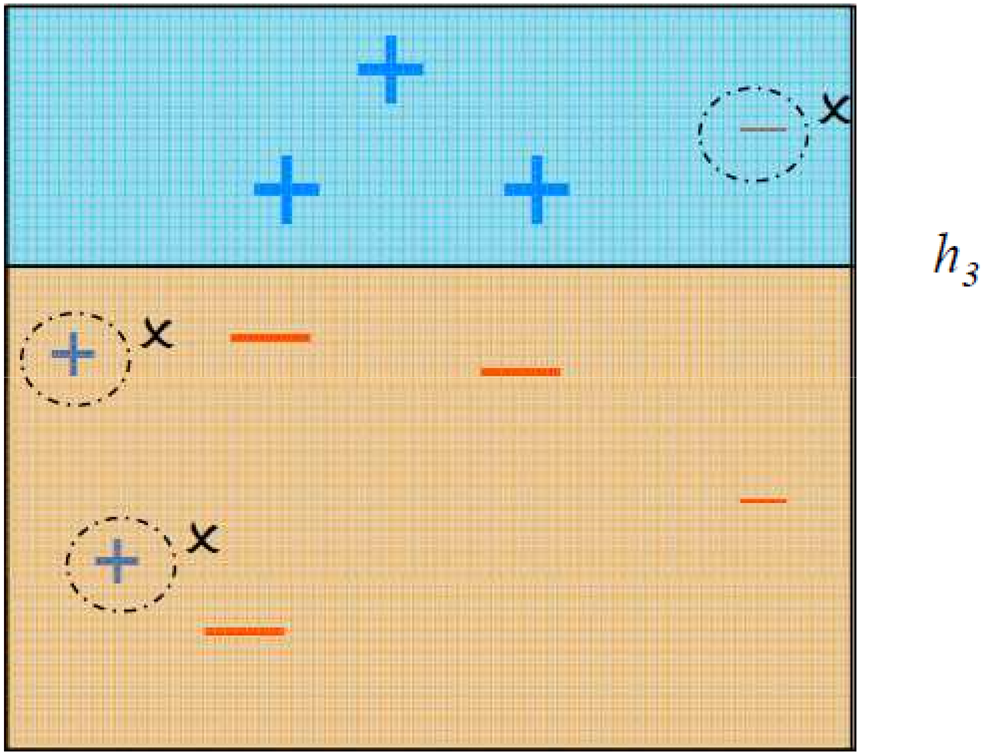
$\varepsilon_2 = 0.196$ $\quad h_2$

$\alpha_2 = 0.704$



D_2

Iteration 3 of 3

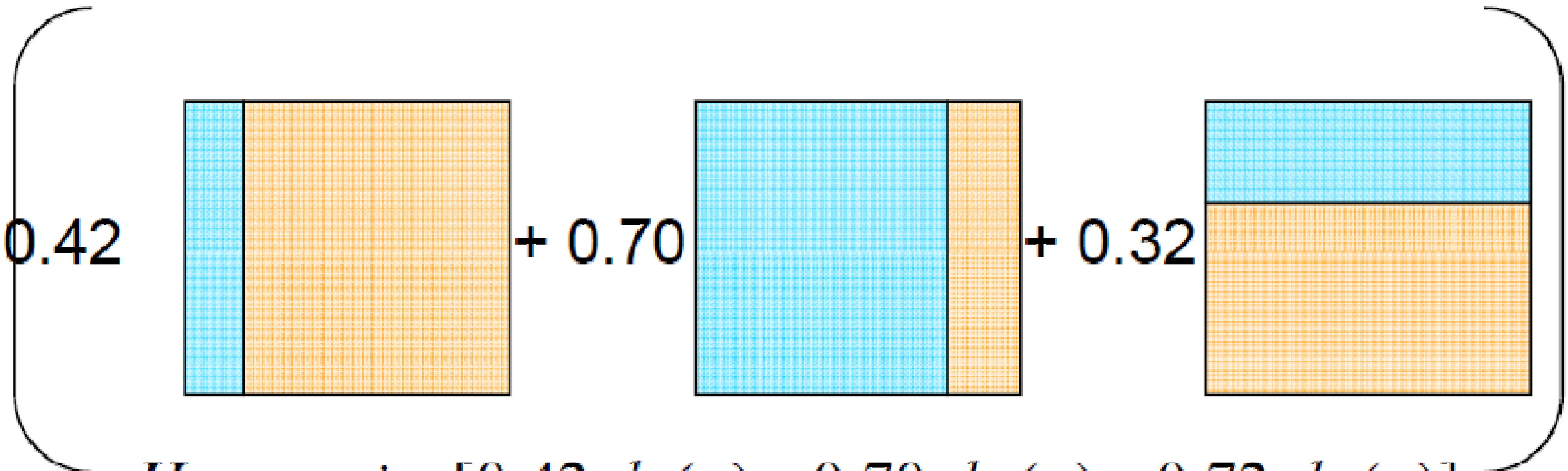


STOP

$\varepsilon_3 = 0.344$

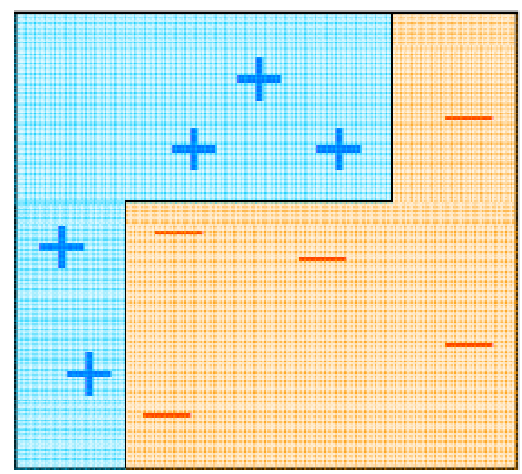
α_2=0.323

Final Model



$$H_{final} = sign[0.42 \cdot h_1(x) + 0.70 \cdot h_2(x) + 0.72 \cdot h_3(x)]$$

$$h_i(x) = \begin{cases} -1, x \rightarrow \Theta_0 \\ +1, x \rightarrow \Theta_1 \end{cases}$$



Ensemble Decomposition Learning

- Better performance of the final model is achieved by building and combining complementary models that are experts in different regions of feature space or regimes of the considered complex system
- Many unspecified regimes are learned implicitly
- However, only aggregated output H_{final} of the ensemble is used in the majority of applications, while the rich internal structure of the meta-model remains completely ignored

$$H_{final} = sign[\boxed{0.42 \cdot h_1(x) + 0.70 \cdot h_2(x) + 0.72 \cdot h_3(x)}]$$

- Our current research is intended to propose methods for extraction of that implicit knowledge
- We will refer to this group of methods as ensemble decomposition learning (EDL)

AdaBoost

- We illustrate the proposed technique using AdaBoost algorithm although it can be used with arbitrary ensemble learning techniques
- The AdaBoost algorithm presented by Freund and Schapire, 1997:

Observation weights: $w_i^{(0)} = 1/N$, where N is a number of training samples

For *m* = 1 to *M* {

a. Fit a classifier $T_m(\mathrm{x})$ to training data with $w_i^{(m)}$

b. Compute
$$err_m = \frac{\sum_{i=1}^{N} w_i^{(m)} I(y_i \neq T_m(\mathrm{x}_i))}{\sum_{i=1}^{N} w_i^{(m)}}$$

c. Compute $\alpha_m = \log((1-err_m)/err_m)$

d. Set $w_i^{(m+1)} = w_i^m \cdot \exp[\alpha_m \cdot I(y_i \neq T_m(\mathrm{x}_i))]$

}

Output $sign(\sum_{m=1}^{M} \alpha_m T_m(\mathrm{x}))$

Ensemble Decomposition Vector

- $H(\mathrm{x}) = \sum_{m=1}^{M} \alpha_m T_m(\mathrm{x})$ is a final aggregated classifier
- One can introduce ensemble decomposition feature vector as follows

$$D(\mathrm{x}) = [\alpha_1 T_1(\mathrm{x}), \alpha_2 T_2(\mathrm{x}), \ldots, \alpha_M T_M(\mathrm{x})]$$

- Each sample after ensemble classification procedure can be represented by this vector
- This vector can provide detailed and informative state representation of the considered system which is not accessible in the aggregated form given by $H(\mathrm{x})$
- $T_i(\mathrm{x})$ are local experts in different implicit regimes or domains of a whole feature space, which ensures good global performance of the final ensemble
- Therefore, it is reasonable to assume that, for similar samples from the same regime, meta-classifier would give similar decomposition vectors

EDL with Single-Example Learning

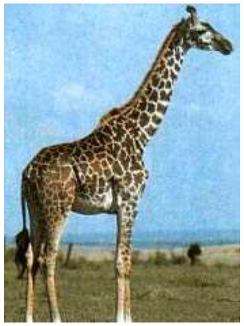
- One of the important multi-disciplinary areas where EDL could be effective is rare, complex, and emerging regimes or patterns classification and forecasting
- Natural choice of EDL tools for this set of problems could be single-example learning (SEL) frameworks
- SEL is a group of machine learning methods aimed at learning classifiers for novel classes by generalization from just one or a few training examples
- It may be useful in the case of rare and emerging pattern/regime forecasting because the main challenge is the absence of the statistically significant history of such events or cases

SEL: Representation by Similarity

- In [Bart, Ullman, 2005] a novel class is characterized by its similarity to a number of previously learned, "familiar" classes
- If there exist classifiers $C_i(\mathrm{x})$ for a number of classes i = 1,…,N, each sample can be characterized by a vector of classifier outputs of familiar classes' classifiers

$C_2(\mathrm{x})$=0.5

$C_1(\mathrm{x})$=0.4



$C_3(\mathrm{x})$=0.3

similarity vector:
$[C_1(\mathrm{x}), \mathrm{C}_2(\mathrm{x}), \mathrm{C}_3(\mathrm{x})]=$
$=[0.4, 0.5, 0.3]$

EDL and SEL: Representation by Similarity

- Ensemble decomposition vector

$$D(\mathrm{x}) = [\alpha_1 T_1(\mathrm{x}), \alpha_2 T_2(\mathrm{x}), \ldots, \alpha_M T_M(\mathrm{x})]$$

can be also regarded as a similarity vector.

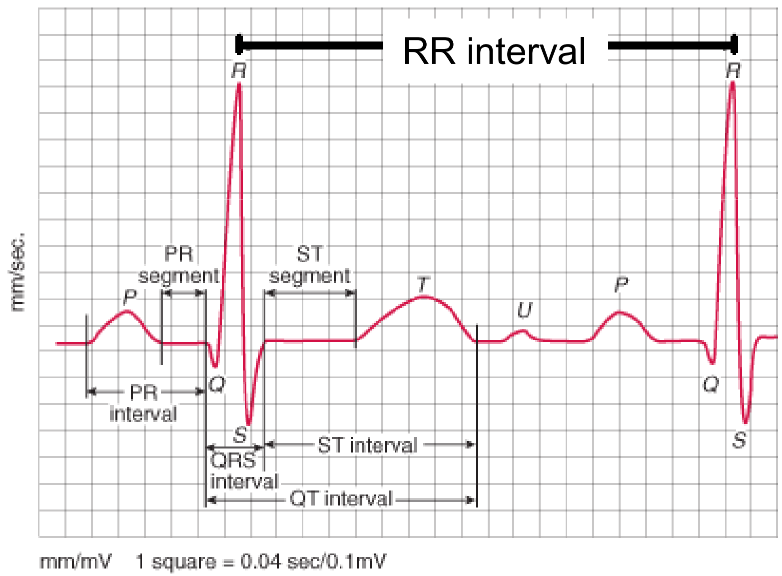
- Two samples x_1 and x_2 are considered to be similar if their ensemble decomposition vectors $D(x_1)$ and $D(x_2)$ are close to each other in some metric, for example, l_1 norm, i.e.

$$\left\| D(x_1) - D(x_2) \right\| < \delta$$

- This approach can be especially useful in applications where significant limitation of data with clear class labels makes it impossible to provide adequate number of reference classes required for standard SEL techniques

Application Example

- Previously we showed that boosting-based framework can be effective for combination of complementary heart rate variability (HRV) nonlinear indicators for express cardiac diagnostics from short beat-to-beat interval (RR) time series

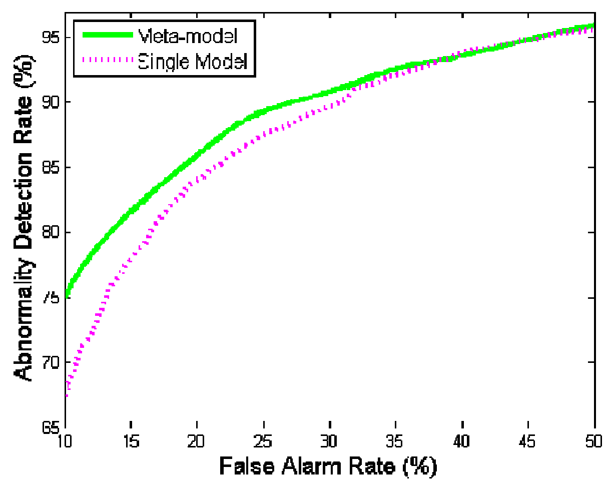


- We will demonstrate how the proposed combination of EDL and SEL techniques can be used for classification of rare pathological cases which is beyond the initial objective of a traditional boosting-based classifier

Normal/Abnormal Meta-Classifier

- At first, we obtain boosting-based multi-abnormality meta-classifier based on two well-known nonlinear dynamics (NLD) measures: detrended fluctuation analysis (DFA) and multi-scale entropy (MSE) [Gavrishchaka, Senyukova, CMLS-2011]

- Significant performance gain is achieved by boosting-based combination of complementary base models

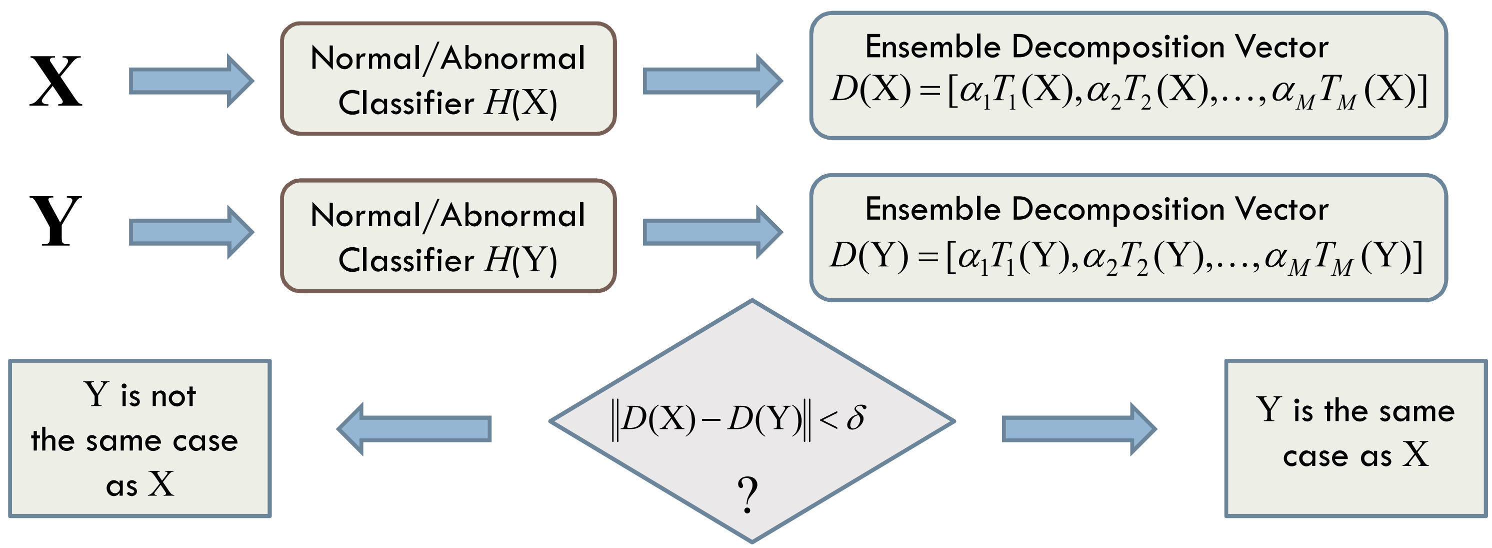


- Normal/abnormal two-class classification framework is used to provide general abnormality warning irrespective of its specific type

- Two-class formulation is tolerant to training data with vaguely specified or non-specific diagnoses, data incompleteness for certain abnormalities, and to complex cases of co-existing pathologies

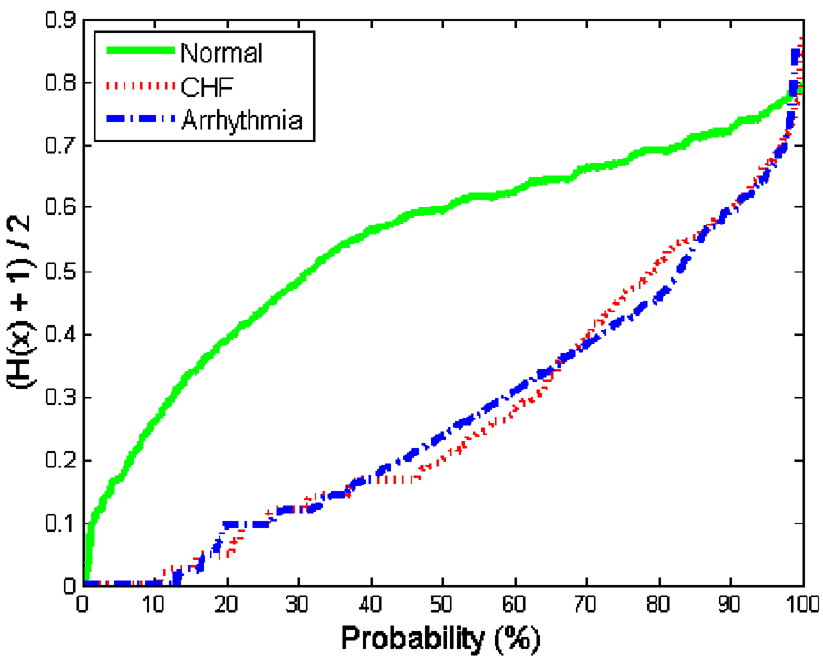
Classification of Rare Cases

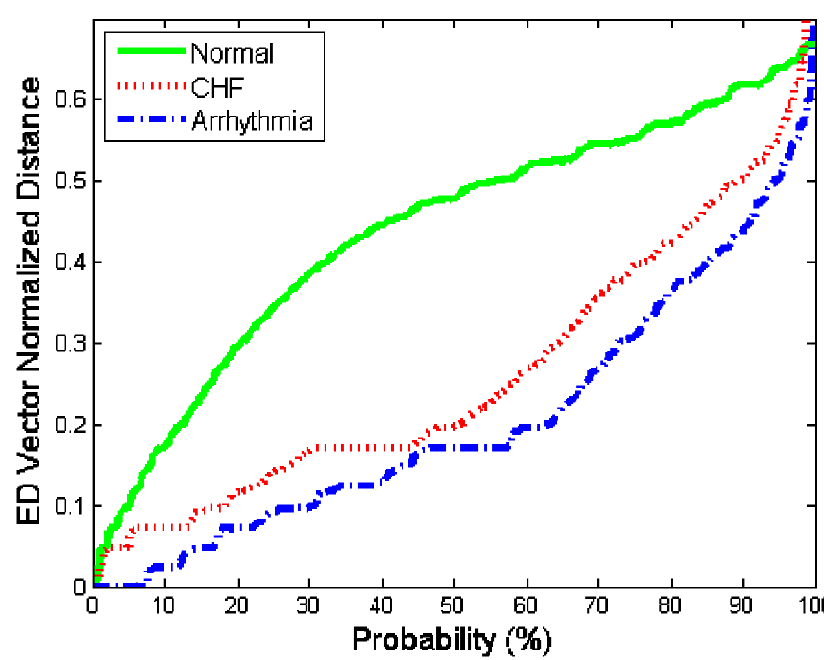
- The discussed meta-indicator is capable to distinguish between a normal state and multiple types of abnormal states – this means that multiple implicit regimes of different abnormalities and normal state are modeled by local experts $T_m(\mathrm{x})$
- Therefore, ensemble decomposition (ED) vector of this meta-classifier could be used for representation of different cases – for example, classification of rare or complex cases, lacking dedicated classifiers or specific diagnostic rules, could be based on the ED vector distance to a known example of such rare case



Comparison

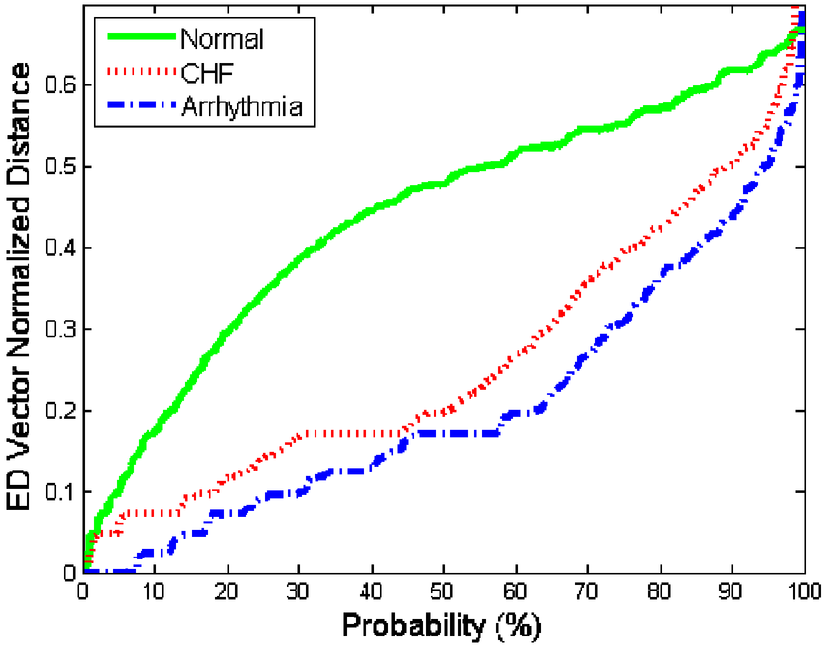
- Introduction of EDL provides better separation between different types of abnormalities (in our case, CHF and arrhythmia) than aggregated output
- There is even no need for large training bases for “familiar” classes as in [Bart, Ullman, 2005]

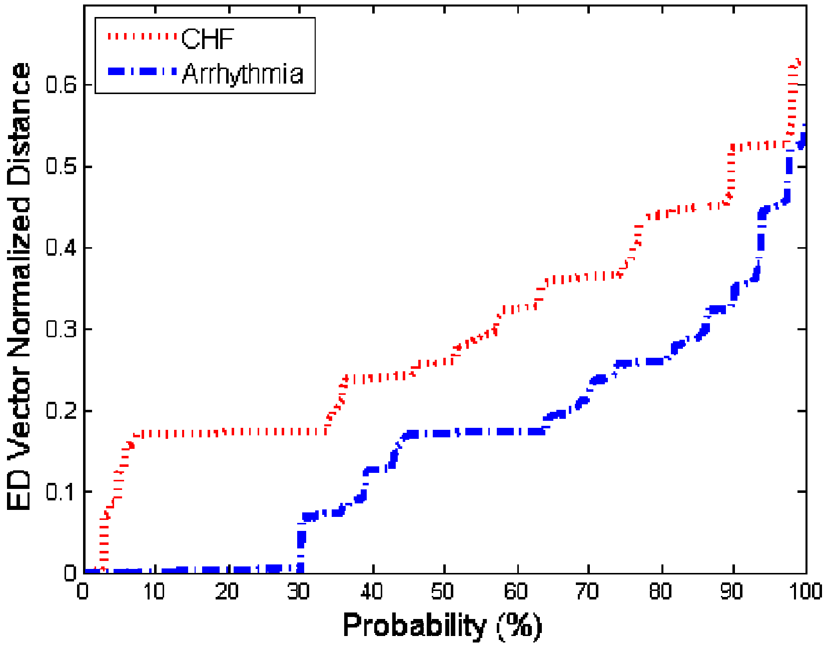




Improvement

The choice of an optimal sub-vector of the ensemble decomposition vector can significantly improve separation between classes





Implementation Details

- We used RR data from www.physionet.org:
  - 52 subjects with normal sinus rhythm
  - 27 subjects with congestive heart failure (CHF)
  - 48 subjects with different types of arrhythmia (MIT arrhythmia database)
- Up to 24 hours of RR data for normal and CHF subjects are available which results in ~7.3x106 of total number of beat-to-beat intervals
- We also added 78 intervals (up to 30min duration each) from patients with supraventricular arrhythmias to combine them with MIT arrhythmia data
- To illustrate universal multi-abnormality detection feature, we obtained two-class "normal-abnormal" meta-classifier using ~1/3 of available normal, CHF, and arrhythmia data for training and the rest data for testing
- All presented performance measures in this section are for out-of-sample data only
- Short RR segments (256 beats or <5min) have been used

Future Plans

- Development of methods for choosing an optimal sub-vector of an ensemble decomposition vector:
  - initial analysis of the base model types (MSE, DFA, MF-DFA, etc.)
  - choosing a subset of local models based on the measure which is the most diverse for the known abnormality types
  - introduction of new base model types
- Combining feature vectors from different ensemble algorithms, including different variations of boosting
- Expanding training data with more abnormality types
- Deployment of our models at www.aqscs.com to allow other researchers and practitioners to test utility of EDL/SEL approach for rare abnormality diagnostics and similar applications
- For questions and comments please contact us at olsen222@yandex.ru (Olga Senyukova) and info@aqscs.com (Valeriy Gavrishchaka)