Leveraging Domain-expert Knowledge, Boosting and Deep Learning for Identification of Rare and Complex States

Rebecca Miao$^{2,*}$, Zhenyi Yang$^{2,b}$ and Valeriy Gavrishchaka$^{1,c}$

$^1$West Virginia University, Physics Department, Morgantown, WV, 26506 USA
$^2$Applied Quantitative Solutions for Complex Systems, 3800 Powell Ln, Suite 623, Falls Church, VA 22041 USA

Abstract. Identification of rare states in complex dynamical systems remains challenging due to data incompleteness and limited success in adopting standard machine learning algorithms for such problems. Although alternatives such as single-example learning (SEL) can learn novel classes from just a few training examples using similarities to reference classes for which large data sets are available, SEL performance quickly deteriorates with decreasing number of available reference classes. Previously, we have shown that with availability of just two broad classes, vector of boosting-ensemble components, discovered in supervised classification learning, provides universal state representation. Analytical models and other domain-expert knowledge can be directly incorporated within such ensemble decomposition learning (EDL) framework reducing requirements for data size. Alternatively, auto-encoders (AE), components of deep learning (DL) frameworks for automatic feature selection and dimensionality reduction, offer unlimited flexibility in discovering robust latent representations. However, AEs require large data sets and unlike boosting are less effective in utilization of existing models. We argue that using EDL vectors as inputs in standard AE architectures allows synergetic combination of the two approaches and discovery of robust state representations even with limited data. The framework is illustrated in the context of personalized medicine application using cardio data from www.physionet.org.

1. Introduction

Identification and forecasting of rare states in complex dynamical systems remains challenging due to data incompleteness and limited success in adopting standard machine learning algorithms for such problems. There are a number of approaches for tuning existing statistical approaches to the specifics of the rare event prediction through objective function and data sample manipulations [1]. However, too many training examples are still required to produce a model with satisfactory generalization.

Alternative and more promising approach to learn novel classes or patterns is single-example learning (SEL) frameworks pioneered in computer vision [2]. 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. SEL ideas agree well with functioning of the human brain [2]. When a human is already capable of distinguishing objects of different classes having seen numerous examples of them, he can learn to identify objects of a previously unseen class, being provided only one image of it. Various SEL approaches have been proposed [3 - 7]. Although diverse in technical details and
implementations, all of these techniques rely on different forms of generic or application-specific prior knowledge and constraints for a drastic reduction of training samples.

One of the generic techniques directly associated with SEL biological origin is single-example learning of novel classes using representation by similarity [3]. In this framework, a novel class is characterized by its similarity to a number of previously learned, familiar classes. If a system can already classify several classes with sufficient accuracy, it can be extended to classify an additional novel class using a single training example. However, performance of this approach could quickly deteriorate with decreasing number of reference classes and available data samples.

Domain-expert models and other knowledge obtained by deeper understanding of the considered application scope could play a key role in cases of severe data incompleteness because of natural dimensionality reduction and usage of domain-specific constraints, e.g., [8, 9]. Such simplified models are often biased and not capable to cover all possible regimes. We have proposed application of boosting-like algorithms for effective utilization of all available domain knowledge via discovery of compact ensembles of complementary low-complexity components [8 - 11]. This approach can tolerate severe data incompleteness and significantly increase accuracy and stability of individual base models. Such boosting ensembles could effectively utilize not just best-on-average models but all complementary domain-expert knowledge after proper parametrization and optimization [9 - 11].

Superior performance of these boosting ensembles is due to optimal combination of complementary models that are experts in different regimes of the considered complex system [10, 12, 13]. Therefore, partial information of many different regimes and states becomes implicitly encoded in the obtained ensemble. However, only aggregated output is used in the typical ensemble learning applications, while the rich internal structure of the ensemble is completely ignored. We have proposed extraction of this underutilized knowledge as a full vector of boosting ensemble components for generic state representation and referred to such approach as ensemble decomposition learning (EDL) [12]. EDL could be effectively used in rare, complex, and emerging regimes or patterns classification and forecasting [10, 12, 13]. Unlike traditional SEL approach, EDL does not require existence of many well-studied classes. Instead, even data from two broad classes (e.g., normal-abnormal in the biomedical context) could be used to build robust two-class ensemble classifier with implicitly encoded sub-classes or sub-regimes required for efficient operation of EDL framework [10, 12, 13].

Recent advancements in deep learning (DL) frameworks based on deep neural networks (DNN) drastically improved accuracy of machine learning in image recognition, natural language processing and other applications [14 – 16]. Although universal capabilities of multi-layer NNs are well known [17, 18], the key advantage of DL is systematic approach for independent training of groups of DNN layers. This includes unsupervised training of auto-encoders (AE) for hierarchical representation of raw input data and supervised re-training of several last layers in the transfer learning that compensate for data incompleteness in a particular application [14], [19 – 23].

The general approach of component-wise training and hierarchical representations in DL is very similar to key ideas in boosting algorithms [24 - 27]. Operational challenges of DNN training such as finding optimal combination of numerous hyper-parameters, including DNN structure, are alleviated by fast technological advancements in massively parallel CPU/GPU systems. However, in cases of severe data limitations and/or absence of relevant problem for the transfer learning, advantages of DNN-based DL are drastically reduced. For example, pure data-driven auto-encoders dealing with high-dimensional raw input data require large amount of data for effective operation even when stacked shallow auto-encoders are used [21].

Existing domain-expert models obtained by deeper understanding of the considered domain could play a key role in applications with severe incompleteness of training data due to natural dimensionality reduction and usage of domain-specific constraints. However, in contrast to boosting-based frameworks, comprehensive incorporation of this domain knowledge into standard DNN-based DL is problematic, except for straightforward guidance in factor selection [27]. While several interesting approaches for combination of the best features of boosting and DNN frameworks have
been described in the literature [28 - 32], they did not focus on the problem of comprehensive integration of domain-expert models and other knowledge into such hybrid frameworks.

Recently we have proposed a hybrid framework combining boosting algorithm applied to domain-expert knowledge with DNN [27]. This framework can potentially tolerate severe data limitations and effectively leverage advantages of existing domain-expert knowledge, boosting-based ensemble learning and DNNs. In particular, instead of AE layers, we have proposed to use ensemble of complementary domain-expert model discovered by boosting as input to classification part of DNN, i.e. multi-layer perceptron (MLP), for further supervised training. In such hybrid system boosting warranties maximum utilization of domain-expert knowledge in situations with very limited data where AEs become ineffective. In turn, DNN may further increase accuracy of compact boosting ensemble by uncovering more subtle patterns such as non-linear mixed terms that were not fully explored by boosting formulation limited to linear combination of complementary models.

In the context of rare states identification, we propose a different hybrid framework that leverages domain-expert knowledge, boosting and DL. AEs, components of DL frameworks for automatic feature selection and dimensionality reduction, discover robust latent representations that potentially can be used for robust quantification of rare and complex states. However, AEs require large data sets and unlike boosting are less effective in utilization of existing models. We argue that using EDL vectors as inputs in standard AE architectures allows synergistic combination of the two approaches and discovery of robust state representations even with limited data. Similar to combination of boosting and DNNs for classification discussed in [27], AE could further enhance state representations offered by EDL vectors through additional nonlinear encoding and flexible mixing of boosting ensemble components. Encouraging results are illustrated in the context of personalized medicine application using cardio data from www.physionet.org

2. Limitations of existing approaches for rare states identification and advantages of novel hybrid framework

There are different approaches for adjustment of standard algorithms to the specifics of rare class prediction through objective function and data sample manipulations such as direct increase of rare example weight in the classification loss function and various oversampling techniques [1]. However, many training examples are still required to produce a model with satisfactory generalization. SEL frameworks pioneered in computer vision offer a promising alternative [2].

One of the effective SEL approaches is based on representation by similarity [2, 3]. In this framework, a novel class is characterized by its similarity to several previously learned classes. If a system is already trained to classify several classes using sufficient amount of training data it can be extended to classify an additional novel class using just one or several examples of this class. Assume that \( n \) well-learned classes \( C_i \) exist, \( F_i \) is a classifier for the \( i \)-th class, \( C(p) \) is the class to which a pattern \( p \) belongs, and \( F_i(p) \) is the output of \( F_i \) on pattern \( p \) (i.e., probability-like number that \( p \) belongs to \( C_i \)). \( F(p) \) can be interpreted as the similarity of pattern \( p \) to the class \( C_i \), i.e.:

\[
F(p) = [F_1(p), F_2(p), ..., F_n(p)]
\]

Other instances of class \( C \) are expected to have similarity/feature vectors (Eq.1) resembling that of \( E \). Therefore, nearest-neighbour classification can be used. New pattern \( p \) is classified as belonging to class \( C \) when the distance between their similarity vectors and that of \( E \) is below a chosen threshold. This approach could quickly deteriorate with decreasing number of reference classes and/or deficit of data for some of the important reference classes.

We have proposed application of boosting-like algorithms for effective utilization of all available domain knowledge via discovery of compact ensembles of complementary low-complexity components [8 - 13]. This approach can tolerate substantial data incompleteness and significantly increase accuracy of individual base models as was demonstrated in cardiac diagnostics and in gait-based detection of neurological abnormalities [9 - 13]. Such boosting ensembles could effectively utilize all complementary domain-expert knowledge not just best-on-average models [8 - 13].
Next, we argued that utility of boosting ensembles extends beyond standard usage of the aggregated ensemble value in classification or regression problems [10, 12, 13]. Indeed, stable multi-regime performance of the ensemble suggests that boosting construct local experts for different implicit regimes or domains of a whole feature space, which ensures good global performance of the final ensemble. We have proposed utilization of this knowledge, implicitly encoded in model ensembles, and called this approach ensemble decomposition learning (EDL), since the extracted information is provided by the individual ensemble constituents (or their subgroups) [12].

EDL technique and its relation to SEL can be summarized as follows [12]. EDL is different from the classical usage of only aggregated ensemble information. For example, in the case of AdaBoost [24, 25], the final classifier (aggregated value) is given by:

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x) / \sum_{t=1}^{T} \alpha_t$$

(2)

Where $h_i(x)$ is a base classifier obtained at the $i$-th boosting iteration, $\alpha_i$ is its weight and $T$ is a total number of iterations. Formally, one can introduce ensemble decomposition feature vector as [12]:

$$D(x) = [\alpha_1 h_1(x), \alpha_2 h_2(x), ..., \alpha_T h_T(x)]$$

(3)

Each sample (state) after ensemble classification procedure can be represented by this vector. Although each individual component of this feature vector may not contain explicit and usable information, collectively, these values may provide detailed and informative state representation of the considered system which is not accessible in the aggregated form given by $H(x)$.

SEL method can be used in the context of EDL for boosting-based model ensemble. The difference is that in standard approach of representation by similarity (Eq.1), each sample is represented by a vector of familiar classes classifier outputs and in EDL it is represented by a vector of weighted base classifier outputs – ensemble decomposition vector (Eq. 3), which plays a role of similarity vector. Unlike traditional SEL, EDL does not require existence of many well-studied classes. Instead, even data from two broad classes could be used to build robust two-class ensemble classifier with implicitly encoded sub-classes required for efficient operation of EDL framework [12].

Robustness of EDL approach for representation of rare and complex physiological states in the context of cardiac diagnostics and personalized medicine was demonstrated in our previous works, where we used parametrized complexity measures adopted from non-linear dynamics (NLD) and linear spectral measures as boosting base models [9 - 13]. These measures represent part of the domain-expert knowledge for the variability analysis of physiological time series [33 - 38]. We found that with parametrized single measure, boosting was not able to achieve significant improvement over the best single model [9 - 13]. Only when two significantly different measures were used as base models, substantial increase of accuracy was achieved with boosting-based ensemble: on both, standard aggregated value approach and related EDL. Therefore, while EDL based on boosting ensemble of domain-expert base models could provide robust representation for rare and complex states using very limited data and coarse label set, its success depends on existence of diverse enough set of base models. When diversity of base models is limited, boosting may still discover complementary models through proper parametrization, however, linear form of discovered components combination without any cross terms could limit potential gains.

AEs, components of DL frameworks for automatic feature selection and dimensionality reduction, offer unlimited flexibility in discovering robust latent representations that potentially can be used for robust quantification of rare and complex states [19 - 21]. Classical “bottleneck” AE offers direct reduction of problem dimensionality and often interpreted as nonlinear principal component analysis (PCA) [19]. De-noising AE could often find even more robust (sparse) representations by increasing inner layer size beyond input dimensionality (over-specification) and adding masking or other noise to the input during training which forces AE to discover non-trivial lower dimensional manifolds as robust latent representations [20].

AEs are often introduced in the context of unsupervised pre-training when large percentage of the training data is unlabelled. However, when labels are available, unsupervised AEs could still be used in DNN training process since they offer effective regularization [20]. When part of training data is labelled, semi-supervised AEs with classification error term for labelled data added to standard
decoding error could be effectively used for forcing AE to find latent representations that are useful for subsequent classification problem with supervised training rather than just generic representations [39, 40]. Since employment of unlabelled data in applications with limited labelled data could be critical for the success of the final classification, effective semi-supervised approaches are often integrated within classification algorithms [41]. Recent generalizations of AE frameworks are more relevant to the problem of rare states identification and forecasting. For example, AEs with simultaneous minimization of standard decoding error and optimization of clustering capabilities based on latent representation is proposed in [42]. However, AEs require large data for effective operation and compared to boosting are less effective in utilization of the existing domain-expert models.

Using EDL vectors as inputs to standard AE architectures allows synergetic combination of the two approaches and discovery of robust state representations even with limited data. Similar to leveraging of boosting and DNNs for classification discussed in [27], AE could further enhance state representations offered by EDL vectors through additional nonlinear encoding and flexible mixing of boosting ensemble components. On the other hand, EDL brings all power of the boosting-based approach in discovery of all complementary domain-expert models. The proposed framework operation and encouraging results are illustrated in the context of personalized medicine application using cardio data from www.physionet.org.

3. Application example

In this section, we illustrate capabilities of the proposed hybrid framework in the context of personalized medicine and well-being maintenance applications based on heart rate variability (HRV) analysis, where HRV measures are computed from beat-to-beat (R-R) time series. The well-known NLD indicators applicable for HRV analysis are based on de-trended fluctuation analysis (DFA), multi-scale entropy (MSE), and multi-fractal analysis (MFA) including MFA extension of DFA [34 - 38]. The comparable performance is also demonstrated by advanced linear indicators based on power spectrum analysis of the RR time series [33]. In general, these HRV measures require long time series for stable calculation [9]. However, HRV indicators have to be computed on short segments in order to capture early signs of developing and/or intermittent abnormalities or to detect subtle initial effects of treatment procedures.

As shown previously, the boosting-based combination of these indicators are suitable for short RR time series [9, 10]. A natural choice of base models for boosting are low-complexity base classifiers, where each of the classifiers uses one complexity measure, \( \beta_i \), out of several choices [9,10]:

\[
(\beta \mid p_i, \gamma)
\]

Here \( \gamma \) is a threshold level (decision boundary) and \( p_i \) is a vector of adjustable parameters of the chosen measure. In our case, \( \beta_i \) may correspond, for example, to either DFA scaling exponent, slope of MSE curve, or power spectrum ratio. Applying boosting steps to a set of such base classifiers with different measures \( \beta_i \) and optimizing over \((p_i, \gamma)\) on each boosting iteration, we can obtain an ensemble of measures with significantly better accuracy and stability.

Relative comparison of \( H(x) \) values computed from short RR segments is the most direct way of express diagnostics and early detection of emerging pathologies [9, 10]. However, in the early stages of the developing and/or intermittent abnormality, a large fraction of short RR segments may be normal, which increases probability of missing such abnormalities. Stability and accuracy can be significantly increased by analysing global ECG dynamics characterized by the full distribution of aggregated values, \( H(x) \), or EDL vectors calculated on a large number of consecutive RR segments [10].

A proper representation of global cardio dynamics could be used for quick and objective matching of the current patient to former cases with known treatment plans and outcomes. Direct comparison of full ECG (RR) time series from two individuals is not effective due to high level of noise and natural
long-term variations of the ECG time series. Collection of consecutive EDL vectors provides effectively filtered representation that removes unimportant variations but preserves differentiation among key micro-states. By calculating Euclidean distances between each EDL vector of one subject with each EDL vector of another subject, distance matrix is obtained. Large distance matrix is often very noisy and usage of simple averages or medians from all cross-EDL distances as aggregated distance measure between subjects is not optimal [10].

The described challenge of handling distance matrix is similar to that encountered in financial applications dealing with quantification of the market state using large and noisy correlation matrices. Recently, it was shown that graph-based approaches such as Minimum Spanning Tree (MST) motivated by the human perception could offer significant advantages over more traditional approaches based on random matrix theory [43, 44]. A spanning tree is a connected graph containing all vertices of the original graph without loops [45]. MST derived by Prim's or Kruskal's algorithm [45] is a spanning tree with minimal length among all spanning trees connecting the nodes of the graph.

MST provides robust low-dimensional representation of the original correlation or distance matrix. The aggregated MST measure is its length defined as

\[ L = \frac{1}{N-1} \sum_{(i,j) \in T} d_{ij} \]  

(5)

Here \((N-1)\) is the number of edges present in MST and \(d_{ij}\) is the distance between nodes \(i\) and \(j\). MST is used to measure similarity between global physiological states of two patients represented by the collection of EDL vectors (Eq.3). If the length of RR time series permits computation of \(N\) EDL vectors from \(N\) consecutive segments for each subject, we can create cross-subject distance matrix, where distances between all EDL vectors of one subject and all EDL vectors of another subject are computed. Information from \(N(N-1)/2\) numbers of distance matrix \(d_{ij}\) will be represented with \((N-1)\) edges of MST. Distance between EDL vectors \(i\) and \(j\), given by Eq.3, is defined as \(L_1\) or \(L_2\) norm in \(T\)-dimensional space. Distance between global physiological states of two subjects is given by the length Eq.5 of the obtained MST. MST-based distance measure is significantly more stable and informative than simple average or median of all distances between EDL vector pairs [10].

Even in the absence of data for direct classification for certain complex or rare states, one can train ensemble indicator using classes where large data sets are available, and use EDL vector based on this ensemble to represent wide range of other physiological conditions [12, 13]. Representing long RR time series with consecutive EDL vectors and using MST for computing aggregated distance Eq.5 between such collections, one can obtain fine differentiation between subjects even within the same abnormality type or normal condition [10]. As a proxy for such illustration we have used self-identification metrics, i.e. we rank distance of each subject to himself against distances to other subjects within the same abnormality type or normal condition. Minimal distance corresponds to rank 1 [10]. In the case of ideal self-identification, all ranking numbers would be 1. If our representation can be used for robust self-identification, i.e. if model ranking is significantly better (lower) than that obtained from random score, it is natural to assume that other subjects would have very similar psycho-physiological conditions and responses to personalized treatment and therapy [10].

Here we extend our analysis in [10] by applying classical bottleneck AE (h2o implementation) with “tanh” transfer function to EDL components. We used the same training/test data set as in our recent work [27] including long R-R records (up to 24 hours each) from http://www.physionet.org for 43 healthy subjects, 27 subjects with congestive heart failure (CHF), 84 subjects with long-term atrial fibrillation (LTAF), 12 subjects from Sudden Cardiac Death (SCD) database and collection of shorter records of miscellaneous abnormalities. Collection of 256-beat R-R segments obtained by 128-beat shifts (i.e., half overlapped) from R-R time series was used, where total number of R-R segments for calculation of DFA, LFHF, and MSE indicators was more than 1.35x10^5. For training, we used normal-abnormal balanced set of 256-beat segments, which includes just 25% of all data.

In the following, we illustrate AE capabilities in improving performance of MST-based distance for self-identification ranking beyond that obtained by pure EDL approach described in [10]. While EDL-based approach was shown to be effective for self-identification within normal subject population as
well as within populations of several abnormalities, here we report results only for population of normal subjects for simplicity, since directionality of the observed AE effects are similar for other groups and differentiation within normal subjects has its own value, e.g. in the case of developing abnormality, subtle psycho-physiological effects and overtraining detection [10, 13].

As in [10] we use collection of 50 consecutive EDL vectors for each subject, except that here base indicators are computed on half-overlapped R-R segments as in [27]. Subject-to-subject distance matrix is computed using simple Euclidean distance between pairs of EDL vectors from different subjects. The final subject-to-subject distance is given by normalized length of the MST tree constructed from the distance matrix using Kruskal’s algorithm [10]. Self-to-self distance for each subject is computed using distinct non-overlapping sets of 50 consecutive EDL vectors with maximum separation to avoid trivial case of measuring distance between the same set of EDL vectors [10].

EDL vectors considered here are based on combination of 10 boosting ensembles with very comparable performance (each consisting of 30 components) trained on slightly different balanced samples (i.e. equal number of R-R segments from healthy subjects with different abnormalities) where base models are based on parametrized LFHF, DFA and MSE measures [10]. This EDL representation is referred as “full boosting ensemble” in this analysis. EDL representation based on boosting ensemble (30 components) obtained from only one type of base model employing parametrized LFHF measure is called “LFHF ensemble”. Finally, equal-weight collection of all LFHF components (39 components) from the full boosting ensemble is called “LFHF components ensemble”.

EDL-based results similar to those reported in [10] are summarized in figures 1 and 2. Figure 1 illustrates that best-on-average single indicators of different types (LFHF, DFA and MSE-based) provide self-ranking noticeably better than “random model”. However, using proper parametrization and complementarity of different types of base models boosting can discover ensembles with much better self-ranking for the vast majority of subjects. Note that the final objective of self-ranking within the same group is not used at all during boosting training. Indeed, original boosting objective (loss function) relates to normal-abnormal classification and training set includes normal and abnormal R-R segments without any information from which particular subject it comes from.

Figure 1. MST-based self-identification ranking obtained from different EDL representations: full boosting ensemble (solid blue), single LFHF (solid red), single DFA (dashed red), single MSE differential (solid yellow) and single absolute MSE (dashed yellow). Dotted black line is median rank (proxy of random model).

Figure 2. MST-based self-identification ranking (sorted) obtained from different EDL representations: full boosting ensemble (blue), single LFHF (red), LFHF-only boosting ensemble (green) and collection of LFHF components from full boosting ensemble (yellow). Dotted black line is median rank.
Figure 3. MST-based self-identification ranking obtained from single LFHF (red), collection of 39 LFHF components from full boosting ensemble (yellow) and latent representation of two bottleneck auto-encoders based on LFHF components (solid and dashed green). Dotted black line is median rank.

Figure 4. MST-based self-identification ranking obtained from full boosting ensemble (blue), single LFHF (red), and latent representations of bottleneck auto-encoders based on LFHF components from full and LFHF-only boosting ensembles (solid and dashed green, respectively). Dotted black line is median rank.

For all specific abnormality types and generic normal-abnormal classification, DFA and LFHF indicators always showed significantly better performance compared to MSE. Also, application of boosting to parameterized DFA and LFHF base indicators did not show any noticeable gain compared to single DFA/LFHF model, which indicated insufficient variability of these models even after parameterization [9-13, 27]. However, when MSE-based indicators were included as base models in addition to DFA and LFHF measures, boosting was able to discover ensembles with normal-abnormal classification/ranking accuracy more than 10% higher than that of best single model.

These facts also remain valid for EDL-based ranking of the subjects within the same group as clear from figure 1 and 2. For example, figure 1 shows that single LFHF and DFA indicators are clearly better than single MSE indicators and boosting provides significant enhancement when all types of indicators are used as base models. On the other hand, figure 2 demonstrates that when only parametrized LFHF indicator is used as base model (without DFA and MSE), enhancement provided by boosting is much less pronounced. Even when all LFHF components from full boosting ensemble are collected within separate ensemble, self-ranking performance is still significantly inferior to the full boosting ensemble consisting of several types of base indicators. This illustrates a practical case with limitations of existing domain-expert models / knowledge in scope and variability which prohibits boosting to provide desired enhancement in performance via model combination.

AE can potentially enhance representation capabilities of any boosting ensemble. However, improvement of EDL representation by AE could be especially valuable in applications where variability and scope of existing set of domain-expert models is limited and boosting-based model combination cannot offer significant enhancement within EDL. In our case, this corresponds to ensembles consisting of only LFHF components shown in figure 2 by yellow and green lines. Ability of the AE to provide enhancement even in this case could be attributed to the well-known boosting feature of continuous margin increase even when formal training error stops decreasing. In our case, boosting also tries to find complementary set of LFHF indicators with different parameter sets, even though in-sample and out-of-sample performance of boosting ensemble do not show noticeable increase. This could be due to intrinsic boosting limitation of using linear combination of base models.

First we apply bottleneck AE to collection of 39 LFHF components from the full boosting ensemble and compare self-ranking based on individual LFHF components (yellow lines) with that based on AE latent representation (green lines) in figure 3. After trying several AE configurations (very coarse-grain search) we choose AE with 3 hidden encoding layers with decreasing number of nodes 20:10:5 (solid green line). Decoding layers used in AE training are symmetric to encoding ones, input layer has 39 nodes and latent representation is given by the most inner layer with 5 nodes.
Addition of more hidden layers did not show any significant improvement in performance. However, for reference, we also show results for AE with 4 hidden encoding layers 30:20:10:5 (dashed green line). Self-ranking from both AEs are comparable but noticeably better than standard single LFHF (red line) and any of individual LFHF components (yellow lines).

In practical situation when only one type of domain-expert model (LFHF in our case) is available, one would use LFHF components discovered by boosting with a single base model. Therefore, AE would be applied to this LFHF ensemble. However, as shown in figure 4, final results from AE with 20:10:5 encoding layers is not sensitive to the choice of initial LFHF collections, i.e. from full boosting ensemble (solid green line) or from LFHF-only boosting ensemble (dashed green line). In both cases, self-ranking is much better than that from single LFHF (red line) and very comparable to EDL representation based on the full boosting ensemble (blue line). Thus, non-linear mixing effects introduced by AE were able to compensate for insufficient variability of the LFHF model within boosting framework alone.

4. Conclusions

We proposed a hybrid framework that leverages domain-expert knowledge, boosting and deep auto-encoders for rare and complex states identification. Existing analytical models and other domain-expert knowledge can be directly utilized within previously introduced EDL approach which results in significant reduction of requirements for training data size. On the other hand, AEs offer unlimited flexibility in discovering robust latent representations when large data sets are available. We argued that using EDL vectors as inputs in standard AE architectures allows synergetic combination of the two approaches and discovery of robust state representations even with limited data. This hybrid approach could be especially valuable in applications where variability and scope of existing set of domain-expert models is limited and boosting-based model combination cannot offer significant enhancement within EDL. Operations of the proposed framework were illustrated in the context of personalized medicine application using cardio data from www.physionet.org. Further enhancements using de-noising and semi-supervised AEs with multi-objective optimization of hyper-parameters are possible and will be reported elsewhere.

References

[1] Joshi M V, Kumar V and Agarwal R C 2001 Evaluating boosting algorithms to classify rare classes: Comparison and improvements Proc. of ICDM pp 257-64
[2] Edelman S 1999 Representation and recognition in vision, MIT Press
[3] Bart E and Ullman S 2005 Single-example learning of novel classes using representation by similarity Proc. of BMVC
[4] Miller E G, Matsakis N E and Viola P A 2000 Learning from One Example through Shared Densities on Transforms Proc. of CVPR 1 pp 464-71
[5] Li F 2003 A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories Proc. of ICCV 2 pp 1134-41
[6] Breuel T M 2003 A Bayesian Approach to Learning Single View Generalization in 3D Object Recognition
[7] Fink M 2004 Object Classification from a Single Example Utilizing Class Relevance Metrics Proc. of NIPS pp 449-56
[8] Gavrishchaka V V 2006 Boosting-Based Frameworks in Financial Modeling: Application to Symbolic Volatility Forecasting Econometric Analysis of Financial and Economic Time Series Advances in Econometrics 20 part 2 pp 123-51
[9] Gavrishchaka V V and Senyukova O 2013 Robust Algorithmic Detection of Cardiac Pathologies from Short Periods of RR Data Knowledge-Based Systems in Biomedicine and Computational Life Science, Studies in Computational Intelligence 450 T.D. Pham, L.C. Jaim, Ed. Heidelberg (Germany: Springer) pp 137-53
[10] Senyukova O, Gavrishchaka V, Sasonko M, Gurfinkel Y, Gorokhova S and Antsygin N 2016 Generic ensemble-based representation of global cardiovascular dynamics for personalized treatment discovery and optimization Computational Collective Intelligence: 8th Int. Conf., Proc. Part I pp 197-207 9875
[11] Gavrishchaka V, Senyukova O and Davis K 2015 Multi-complexity Ensemble Measures for Gait Time Series Analysis: Application to Diagnostics, Monitoring and Biometrics Advances in Experimental Medicine and Biology 823 et al Sun C, Bednartz T, Pham T D, Vallotton P, and Wand D Cham (Switzerland: Springer Int. Publishing) pp 107-26
[12] Senyukova O and Gavrishchaka V 2011 Ensemble decomposition learning for optimal utilization of implicitly encoded knowledge in biomedical applications Proc. of the 6th IASTED Int. Conf. on Computational Intelligence and Bioinformatics, CIB 2011 pp 69–73
[13] Senyukova O, Gavrishchaka V and Koepke M 2014 Universal Multi-complexity Measures For Physiological State Quantification in Intelligent Diagnostics and Monitoring Systems The First Int. Aizu Conf. on Biomedical Informatics and Technology: CCIS 404 pp 76-90 Springer-Verlag, Berlin
[14] LeCun Y, Bengio Y and Hinton G May 2015 Deep learning Nature 521 pp 436-44
[15] Deng L and Yu D June 2014 Deep Learning: Methods and Applications Foundations and Trends in Signal Processing 7 no3-4 pp 197-387
[16] Krizhevsky A, Sutskever I, E G and Hinton 2012 Imagenet Classification with Deep Convolutional Neural Networks Proc. of the Neural Information Processing System, et al Pereira F, C C J, Bottou B L, and Q K Weinberger pp 1097-105 2
[17] Kolmogorov A 1957 Dokl. Akad. Nauk. SSSR 114 pp 953-6
[18] Cybenko G Dec 1989 Math. Control Signals Systems 2 no3 pp 303-14
[19] Hinton G E and Salakhutdinov R July 2006 Science 313 pp 504-7
[20] Vincent P 2010 Journal of Machine Learning Research 10 pp 3371-408
[21] Gehring J, Miao Y, Metze F, and Waibel A 2013 Extracting deep bottleneck features using stacked auto-encoders Int. Conf. on Acoustics, Speech, and Signal Processing pp 3377-81
[22] Shin H, Roth H R, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D J and Summers R M May 2016 Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning IEEE Trans. Med. Image 35 no5 pp 1285-98
[23] Christodoulidis S, Anthimopoulos M, Ebner L, Christie A and Mougiakakou S Jan 2017 IEEE J. Biomed. Health Inform 21 no1 pp 76-84
[24] Schapire R E 1992 The design and analysis of efficient learning algorithms Ph.D. dissertation Massachusetts Institute of Technology Cambridge MA
[25] Friedman J, Hastie T and Tibshiranim R Oct 2001 The Annals of Statistics 28 no2 pp 337-407 Apr 2000
[26] Friedman J Annals of Statistics 29 no5 pp 1189-232
[27] Gavrishchaka V, Yang Z and Miao R and Senyukova O 2018 Advantages of Hybrid Deep Learning Frameworks in Applications with Limited Data Int. Journal of Machine Learning and Computing 8 no6 pp 549-58
[28] Che Z, Purushotham S, Kheiman R and Liu Y Dec 2015 Distilling knowledge from deep networks with applications to healthcare domain arXiv:1512.03542 [stat.ML] 11
[29] Moghim M, Saberian M, Yang J, Li L J, Vasconcelos N and Belongie S 2016 Boosted Convolutional Neural Networks Proc. of the British Machine Vision Conf. pp 24.1-24.13
[30] Schwenk H and Bengio Y Aug 2000 Neural Computation 12 no8 pp 1869-87
[31] Sharkey A J C 1999 Boosting Using Neural Networks Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems, A. J. C. ed Sharkey (London UK: Springer London) pp 51-78
[32] Shalev-Shwartz S Apr 2017 SelfieBoost: A Boosting Algorithm for Deep Learning arXiv:1411.3436v2 [stat.ML]
[33] Heart rate variability: standards of measurement, physiological interpretation, and clinical use
Circulation Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology 93 no5 pp 1043-65 Mar 1996

[34] Belair J, Glass L, Der Haien U and Milton J 1995 J. Dynamical Disease: Mathematical Analysis of Human Illness (New York: AIP Press)

[35] Peng C K, Havlin S, Stanley E H and Goldberger A L Sep 1995 Chaos 5 pp 82-7

[36] Costa M, Goldberger A L and Peng C K Feb 2005 Physical Review Letters E 71 021906

[37] Makowiec D, Dudkowska A, Zwierz M and Rynkiewicz A May 2006 Acta Physica Polonica B 37 no5 pp 1627-39

[38] Voss A, Schulz S, Schroederet R, Baumert M, M and Caminal P Jan 2009 Philosohical Transactions of the Royal Society A 367 pp 277-96

[39] Snoek J, Adams R P and Larochelle H Journal of Machine Learning Research 13 2012 pp 2567-88

[40] Taghanaki S A, Kawahara J, Miles B and Hamarneh G July 2017 Computer Methods and Programs in Biomedicine 145 pp 85-93

[41] Oliver A, Odena A, Raffel C, Cubuk E D and Goodfellow I J May 2018 Realistic Evaluation of Semi-Supervised Learning Algorithms hyperparamarXiv:1804.09170v2 [cs.LG] 23

[42] Yang B, Fu X, Sidiropoulos N D and Hong M Jun 2017 Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering arXiv:1610.04794v2 [cs.LG] 13

[43] Onnela J-P, Chakraborti A, Kaski K, Kertesz J and Kanto 2003 A Phys. Rev. E 68 056110

[44] Tumminello M, Lillo F and Mantegna R N 2010 J. Econ. Behav. Organ 75 pp 40-58

[45] Theodoridis S and Koutroumbas K 1998 Pattern Recognition Academic Press San Diego CA