DIAGNOSTICS OF COMPLEX AND RARE ABNORMALITIES USING ENSEMBLE DECOMPOSITION LEARNING

OLGA V. SENYUKOVA
Department of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russian Federation
olsen222@yandex.ru

VALERIY V. GAVRISHCHAKA
Department of Physics, West Virginia University, Morgantown, WV 26506 USA
gavrish@verizon.net

ABSTRACT
Diagnostics of complex and rare medical cases lacking clear symptoms of particular abnormalities is a very challenging problem. Case based reasoning (CBR) is known to provide helpful and generic guidelines for practitioners as well as for the design of medical expert systems dealing with such non-standard diagnostic problems. Single example learning (SEL) algorithms offer more formal machine learning framework for classification of rare and novel classes. However, both CBR and SEL approaches require significant number of well-studied examples and a set of objective features capable to provide robust matching of novel examples with the previous ones. In practice, data sets for well-defined abnormalities suited for quantification with existing indicators are often limited. However, significant amount of valuable clinical information from cases labeled only as normal or abnormal without particular diagnosis remains underutilized. Recently, we have demonstrated that this information can be effectively employed to produce powerful normal-abnormal meta-classifiers using ensemble learning techniques applied to existing physiological indicators. This is achieved by an optimal weighted combination of complementary indicators which are experts in different regimes of the considered biological complex system. Therefore, partial information of wide variety of dynamical regimes becomes implicitly encoded in the obtained ensemble of classifiers, while only aggregated output is used. Extraction of this underutilized knowledge could be formalized in terms of ensemble decomposition learning (EDL) and used for representation of complex and rare cases in terms of intrinsic dynamical regimes. Such representation could prove to be more accurate and robust compared to traditional CBR and SEL approaches. Illustrative application of the EDL approach to cardiac diagnostics based on HRV analysis is presented and discussed.

KEY WORDS
Rare case diagnostics, Ensemble learning, Boosting, Ensemble decomposition and single-example learning, Heart rate variability analysis, Cardiac diagnostics, Physiological models

1. INTRODUCTION

Diagnostics of complex and rare medical cases lacking clear symptoms of particular abnormalities is a very challenging problem. Case-based reasoning

(CBR) is known to provide helpful and generic guidelines for the design of medical expert systems dealing with such non-standard diagnostic problems [1]. Similarly, most practitioners implicitly follow case-based reasoning framework when they try to apply previously encountered cases to similar new situations.

Single-example learning (SEL) algorithms offer more formal machine learning framework for classification of rare and novel classes. 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. Usually they use prior knowledge obtained previously while learning other classes from large databases. It has been proved that human brain works the same way [2].

However, both CBR and SEL approaches have serious limitations when applied to our problem, the most important of which is requirement of a large number of examples of well-studied cases. In practice, data sets for well-defined abnormalities suited for quantification with existing indicators are often limited. However, significant amount of valuable clinical information from cases labeled only as normal or abnormal without particular diagnosis remains underutilized.

Recently, we have demonstrated that these coarsely classified data can be effectively employed to produce powerful normal-abnormal meta-classifiers using ensemble learning techniques, especially boosting, applied to existing physiological indicators which play the role of individual weak classifiers. This is achieved by an optimal weighted combination of complementary indicators which are experts in different regimes of the considered biological complex system. Therefore, partial information of wide variety of dynamical regimes becomes implicitly encoded in the obtained ensemble of classifiers. However, only aggregated output is used for normal-abnormal classification, while the rich internal structure of the ensemble is completely ignored.

Extraction of this underutilized knowledge could be formalized in terms of ensemble decomposition learning (EDL). Representation of complex and rare cases by the vector output of the ensemble of classifiers each element of which is an output of an individual classifier multiplied by its weight could prove to be more accurate and robust compared to significantly more coarse-grained representation typical for CBR and SEL approaches. Illustrative application of the proposed EDL approach to cardiac diagnostics based on HRV analysis is presented and discussed.

2. COMPLEX/RARE CASE DIAGNOSTICS: CHALLENGES AND EXISTING METHODS

The main challenge of the rare, complex, and emerging pattern/regime forecasting or classification is the absence of the statistically significant history of such events or cases. There were many attempts of tuning existing machine learning and statistical approaches, including boosting, to the specifics of the rare event/class prediction through objective function and data sample manipulations [3]. However, still too many examples are required to produce a model with satisfactory generalization. Alternative and more promising approach to learn novel/rare classes or patterns could be SEL frameworks pioneered in computer vision applications [2].

In medicine, CBR guidelines are often used for diagnostics of rare and complex cases. CBR is implicitly used for discretional diagnostics by practitioners as well as in computerized medical expert systems [1]. In a broad sense, case-based reasoning means adapting old solutions to meet new demands or to explain and interpret new situations. Success of CBR-based solution critically depends on existence of previously studied cases similar to the currently considered and robust feature set that relates new example to the previous one. Although, CBR is more familiar for medical practitioners, SEL is more formalized and better suited for data-driven computerized applications. SEL could be used as part of our novel EDL approach described in the next section. Therefore, short summary of SEL fundamentals is presented next.

SEL is a trend in machine learning which agrees well with the way our brain works [2]. When a human is already capable of distinguishing objects of different classes having seen numerous examples of them, e.g. horses and cats, he can learn to identify objects of a previously unseen class, for example, opossums, being provided only one image of it. SEL techniques attempt to exploit this fact when teaching a machine. These techniques become more and more popular due to their intuitive biological rationale and effectiveness. Actually, the progress in machine learning as a whole leads to larger and larger databases of common objects and more efficient classification methods to distinguish them. Therefore, in some cases there is no need to create large training bases and employ sophisticated classification methods, if it is possible to adopt SEL techniques. Moreover, there are numerous situations when a sufficient number of training samples is simply not available.

Various approaches to address SEL problems have been proposed in previous works [4-8]. Although diverse in technical details and implementations, all of these techniques rely on different forms and representations of generic or application-specific prior knowledge and constraints for a drastic reduction of the required training samples. One of the generic techniques directly associated with SEL biological origin is single-example learning of novel classes using representation by similarity [4]. 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.

Thus, both CBR and SEL approaches require significant number of well-studied examples and a set of objective features capable to provide robust matching of novel examples with the previous ones. The proposed EDL approach described in the next section is capable to overcome these limitations.

3. COMPLEX/RARE CASE DIAGNOSTICS BASED ON ENSEMBLE DECOMPOSITION LEARNING

Several generic algorithmic approaches for the utilization of knowledge implicitly encoded in model ensembles could be proposed. Collectively, these algorithms could be called ensemble decomposition learning (EDL) techniques, since the extracted information is provided by the individual ensemble

constituents, $h_i(x)$, where x is an instance we want to classify, or their subgroups. This is in contrast to the classical usage of only aggregated information. For example, in the case of AdaBoost [9] it is given by

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

where $h_i(x)$ is a base classifier obtained at the i-th iteration, α_i is its weight and T is a number of iterations. Formally, one can introduce ensemble decomposition feature vector as follows:

$$D(x) = [\alpha_1 h_1(x), \alpha_2 h_2(x), \ldots, \alpha_T h_T(x)] \ . \qquad (2)$$

Each sample 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 $H(x)$. Indeed, boosting and similar algorithms construct local experts $h_i(x)$ for 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.

The implicitly encoded knowledge in boosting-based and other model ensembles could be used for different types of applications which may require appropriate algorithms for extraction and utilization of such knowledge. One of 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 SEL frameworks in general [4-8] and SEL based on representation by similarity in particular [4].

SEL method can be used in the context of EDL for boosting-based or similar model ensemble. The key difference is that in [4] each sample is represented by a vector of familiar classes classifier outputs and in our case it is represented by a vector of weighted base classifier outputs – ensemble decomposition vector (Eq. 2). 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) 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.

Ensemble decomposition vector can also be regarded as a similarity vector in [4]. Two samples x_1 and x_2 are considered to be similar if their ensemble decomposition vectors $D_1 = D(x_1)$ and $D_2 = D(x_2)$, are close to each other in some metric, for example, l_1 norm. Then, if we obtain an ensemble decomposition vector of an instance of some class (let's call it a training vector), other instances can be classified as belonging to this class or not according to the similarity of their ensemble decomposition vectors to the training vector. The

final decision can be based on a threshold δ which represents some critical value of l_1 norm of the difference vector – $\|D_1 - D_2\|_1$, so that if $\|D_1 - D_2\|_1 < \delta$, the instances are considered belonging to the same class. The presented EDL approach is generic and can be used in many different fields. Potential biomedical applications of this framework are outlined in the next section.

4. APPLICATION EXAMPLE

Technological advancements in clinical, portable, and wearable devices for real-time collection of physiological data provide new opportunities for computerized diagnostics in medicine. However, many conceptual and algorithmic challenges for accurate and robust quantitative modeling in such applications remain unresolved. Ensemble learning frameworks could be effective in alleviating and resolving many of these problems [10]. Recently, boosting-based framework was shown to 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 [11]. This approach could be more robust compared to traditional electrocardiogram (ECG) waveform analysis in the early stage of the developing cardiac abnormalities, for clinically significant pathologies lacking specific ECG signatures, and for express diagnostics from short RR segments using wearable devices.

Here we follow [11] to obtain boosting-based multi-abnormality meta-classifier based on two well-known nonlinear dynamics (NLD) measures: detrended fluctuation analysis (DFA) [12] and multi-scale entropy (MSE) [13]. Normal-abnormal two-class classification framework is used to provide general abnormality warning irrespective of its specific type. As discussed earlier, 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. Then, we show that such meta-classifier could be potentially used by the earlier described EDL method for classification of rare pathological cases which is beyond the initial objective of this meta-classifier. The presented simplified analysis is for illustration of typical EDL operation only. Significantly more accurate meta-classifier and corresponding EDL-based classification could be obtained for larger pool of base models and more extensive RR data sets.

For the illustrative analysis presented in this section, we used RR data from 52 subjects with normal sinus rhythm, 27 subjects with congestive heart failure, and 48 subjects with different types of arrhythmia (MIT arrhythmia database) from www.physionet.org. Up to 24 hours of RR data for normal and congestive heart failure (CHF) subjects are available which results in ~7.3x106 of total number of beat-to-beat intervals. In addition, up to 30 min of RR data are available for each subject with arrhythmia. We also added 78 intervals (up to 30min duration each) from patients with supraventricular arrhythmias to combine 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. All presented performance measures in this section are for out-of-sample data only. Short RR segments (256 beats or <5 min) have been used.

Two-class classification receiver operating characteristic (ROC) curves with equally averaged CHF and arrhythmia detection rates for boosting meta-classifier and the best single model are shown in Fig.1. It demonstrates significant performance gain achieved by boosting-based combination of complementary base models using just two different types of NLD measures.

Fig.2a shows individual ROC curves for CHF-normal and arrhythmia-normal classification that further illustrate significant information content of the boosting meta-classifier for classification between normal state and different types of abnormalities. However, ROC curve for classification between CHF and arrhythmia presented in the same figure is almost diagonal indicating incapability of this meta-classifier to distinguish between two different types of abnormality. This illustrates that, in accordance with this meta-classifier original objective, it indeed provides clear differentiation between normal and abnormal cases. However, different types of abnormalities are not distinguished by a single aggregated output of the ensemble.

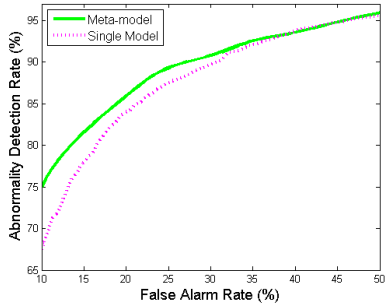


Fig.1 ROC curve with equally averaged CHF and arrhythmia detection rate: meta-model (solid line) and optimal single model (dotted line)

A discussed meta-indicator is capable to distinguish between normal states and multiple types of abnormal states using short RR segments. This means that multiple implicit regimes of different abnormalities and normal state are modeled by local experts, $h_i(x)$. Therefore, ensemble decomposition vector (Eq. 2) of this meta-classifier could be used for representation of various cases partially related to different features of the encoded types of abnormalities. For example, classification of rare or complex cases, lacking dedicated classifiers or specific diagnostic rules, could be based on the ensemble decomposition vector distance to a known example of such rare case as described in Section 3.

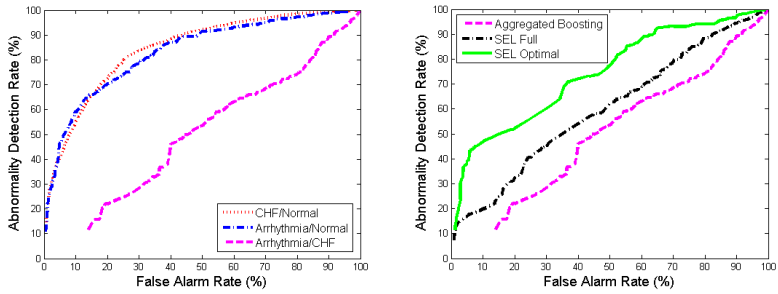
In this illustration, to approximate rare cases with arrhythmia-type signatures, we used out-of-sample arrhythmia data. We chose one arrhythmia sample as a reference ("training") example of such "rare" case and computed normalized distances of feature vectors of other arrhythmia, CHF and normal cases to this chosen example. Now instead of aggregated ensemble output this

distance to a "reference" case is used for classification. Similarly, ROC curves for this SEL-type classifier obtained in the context of EDL approach can be computed. The obtained curves for normal-abnormal classification (CHF-normal and arrhythmia-normal) are quite similar to those presented in Fig.2a. However, results for arrhythmia-CHF classification are very distinct and summarized in Fig.2b.

In Fig.2b, we show arrhythmia-CHF classification ROC curves for three different classifiers:

- based on aggregated ensemble output as in Fig.2a
- SEL classifier using feature vector based on the full ensemble
- SEL classifier with optimal sub-vector based on the part of the ensemble.

We see that, SEL-based approach in the EDL context is capable to construct quite accurate classifier for the two classes that are almost indistinguishable when standard aggregated ensemble output is used. It should be noted that presented SEL classifiers are based on just one reference example which suggests suitability of this approach for diagnostic of complex and rare cases characterized by extreme limitation of the available data.



(a) (b)

Fig. 2 (a) ROC curves for CHF-normal (dotted line), arrhythmia-normal (dash-dotted line), and arrhythmia-CHF (dashed line) classification based on the aggregated output of the boosting ensemble;
(b) ROC curves for arrhythmia-CHF classification based on the aggregated output of the boosting ensemble (dashed line), SEL classifier with feature vector based on the full ensemble (dashed-dotted line) and with optimal sub-vector based on part of the ensemble (solid line)

The presented illustrative example also indicates that quick initial analysis of the base model types may suggest the choice of optimal sub-vector using just subset of the ensemble which could make EDL approach significantly more accurate. The obvious heuristic is to choose subset of local models based on measure which is the most diverse for known abnormality types. In this particular case, it is a slope of MSE curve.

Including more types of base models and combining feature vectors from different ensemble algorithms, including different variations of boosting, one can expect to obtain more practical and accurate decision-support classifiers based on EDL approach. Future research on this topic for biomedical and other applications is warranted.

5. CONCLUSIONS

Challenges of diagnostics of complex and rare medical cases lacking clear symptoms and limitations of the existing solutions have been discussed. A novel approach based on decomposition of model ensembles has been proposed. We have argued that many ensemble learning techniques including boosting implicitly encode detailed information about different local regions of global feature space, or regimes. However, only aggregated output is used in the majority of applications, while the rich internal structure of the ensemble is completely ignored. Extraction of this underutilized knowledge could be formalized in terms of ensemble decomposition learning techniques. We have outlined one of such frameworks based on existing single-example learning algorithms. Operational details and benefits of this approach in biomedical applications were presented and discussed.

REFERENCES

[1] KOLONDER Janet L. An Introduction to Case-Based Reasoning. Artificial Intelligence Review [J], 1992, 6, PP3-34.

[2] EDELMAN Shimon. Representation and recognition in vision. MIT Press, 1999.

[3] JOSHI Mahesh V; KUMAR Vipin; AGARWAL Ramesh C. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. Proceedings of ICDM, 2001, PP257-264.

[4] BART Evgeniy; ULLMAN Shimon. Single-example learning of novel classes using representation by similarity. Proceedings of BMVC, 2005.

[5] MILLER Erik G; MATSAKIS Nicholas E; VIOLA Paul A. Learning from One Example through Shared Densities on Transforms. Proceedings of CVPR, 2000, 1, PP464-471.

[6] FEI-FEI Li et al. A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories. Proceedings of ICCV, 2003, 2, PP1134 – 1141.

[7] BREUEL Thomas M. A Bayesian Approach to Learning Single View Generalization in 3D Object Recognition. 2003.

[8] FINK Michael. Object Classification from a Single Example Utilizing Class Relevance Metrics. Proceedings of NIPS, 2004, PP449—456.

[9] FREUND Yoav; SCHAPIRE Robert E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences, 1997, 55(1), PP119-139.

[10] GAVRISHCHAKA Valeriy V; KOEPKE Mark E; ULYANOVA Olga N. Ensemble Learning Frameworks for the Discovery of Multi-Component Quantitative Models in Biomedical Applications. Proceedings of ICCMS**,** 2010, 4, PP329-336.

[11] GAVRISHCHAKA Valeriy V; SENYUKOVA Olga. Robust algorithmic detection of the developed cardiac pathologies and emerging or transient abnormalities from short periods of RR data. Proceedings of CMLS, 2011, 1371, PP215-224.

[12] PENG Chung-Kang; HAVLIN Shlomo; STANLEY Eugene H; GOLDBERGER Ary L. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos [J], 1995, 5, PP82-87.

[13] COSTA Madalena; GOLDBERGER Ary L; PENG Chung-Kang. Multiscale entropy analysis of biological signals. Physical Review Letters [J], 2005, E 71, 021906.

[14] BISHOP Christopher M. Pattern Recognition and Machine Learning. Springer, 2006.

[15] SCHAPIRE Robert E. The Design and Analysis of Efficient Learning Algorithms. MIT Press, 1992.

[16] GAVRISHCHAKA Valeriy V. Boosting-Based Frameworks in Financial Modeling: Application to Symbolic Volatility Forecasting. Advances in Econometrics [J], 2006, 20B, PP123-151.