Global Optimization Approaches for Parameter Tuning in Biomedical Signal Processing: A Focus on Multi-scale Entropy

Mohammad Ghassemi[1], Li-wei Lehman[1], Jasper Snoek[2] and Shamim Nemati[2]

Abstract

Many algorithms used for the analysis of physiological signals include hyper-parameters that must be selected by the investigator. The ultimate choice of these parameter values can have a dramatic impact on the performance of the approach. Addressing this issue often requires investigators to manually tune parameters for their particular data-set. In this study, we illustrate the importance of global optimization techniques for the automated determination of parameter values in the multi-scale entropy (MSE) algorithm. Importantly, we demonstrate that global optimization techniques provide an effective, and automated framework for tuning parameters of such algorithms, and easily improve upon the default settings selected by experts.

1. Introduction

The increasing use of very large data sets, and the advent of effective, scalable methods for analyzing them, have stimulated progress in many fields of research. In bio-medical research, this trend has been most apparent in genomics, but large-scale data resources are also beginning to have significant impact in basic and applied research in physiology and medicine, in which signal processing plays a key role [1].

PhysioNet is one example of a research resource for physiological signal processing [2]. Despite the rich volumes of high resolution waveform and time series data available in PhysioNet archives, countless research papers based on PhysioNet's data resources utilize only sparse, low-resolution subsets of available high-resolution vital signs and laboratory measurements(e.g., minimum, maximum, or average hear rate and blood pressure over a 24 hours period). This approach may be sub-optimal however as recent studies recommend that therapeutic interventions should not only aim at maintaining patient vitals within an acceptable static range, but also direct a patients trajectory towards healthy dynamical regimes with enhanced variability [3, 4].

As of last year, the two most frequently used techniques for quantifying time series variability in the physiological signal processing community were sample entropy (SE) and approximate entropy (AE) [5]. Multi-scale entropy (MSE) has been described as a more robust alternative for quantifying the dynamical activity of a physiological time-series. MSE may be understood as the set of sample entropy values for a signal which is averaged (or coarse-grained) over various increasing segment lengths. As demonstrated by Costa et al. [6], MSE is a more descriptive index of various types of signal variability than the SE of the original signal alone. A full description of the multi-scale entropy algorithm and it's advantages over SE can be found in [7]

To truly take advantage of MSE, however, requires the investigator to specify several parameter values, prior to analysis. These parameters include features which carry over from SE, such as the length of the sequences to be compared (commonly denoted m), a similarity threshold (commonly denoted r) as well as some MSE specific features such as the maximum time scale for which the SE is computed and the step size in the scale [7]. Existing techniques for principled selection of SE and MSE parameters include brute force and Monte Carlo techniques [8] among others [9] but manual tuning, or a simple reliance on default values, is also commonly employed.

We, like others in the community, believe that MSE and SE should not be a function of unprincipled parameter selection. [5, 9]. To address this issue, we propose use of recent advances in global optimization techniques for the identification of MSE parameters. In this paper, we directly compare parameters selected by the Multi-start Scatter-Search [10], Genetic Algorithm and the Bayesian Optimization [11] approaches to global optimization against the default values for MSE on an intensive care unit (ICU) sepsis dataset.

This work was supported in part by the James S. McDonnell Foundation Postdoctoral grant. The content of this article is solely the responsibility of the authors.

[1] M. Ghasemi and L. Lehman are with the Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. Correspondence email: ghassemi@mit.edu

[2] S. Nemati and J. Snoek are with the Harvard School of Engineering and Applied Sciences, 33 Oxford Street, Cambridge, MA 02138, USA. Correspondence email: shamim@seas.harvard.edu

2. Methods

2.1. Datasets

This study utilized retrospective data from a subset of ICU patients from the publicly available Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) database [12] that matched the definition of sepsis and severe sepsis as previously described by Mayaud et al. [13]. We selected $N = 118$ patients who had complete ECG waveforms for their first 24 hours in the ICU. We also collected the Acute Physiology and Chronic Health Evaluation IV (APACHE) scores for the cohort. The average heart rate (HR) time series over 10 second sliding windows (with no overlap) were extracted using peak detection with weighted averaging based on the quality of the individual heart beat waveform falling within each window.

2.2. Classification via Transductive SVM

The MIMIC ICU population constitute patients from several care units. Moreover, patients are subjected to a variety of interventions (ventilators, medications, etc.). This inherent heterogeneity raises an important question: how can one transfer knowledge about a given patient cohort to an unseen newly admitted patient? In this work, we assume that the patient population is made of multiple distinct clusters, and a new patient may belong to any one of the existing clusters with different probabilities. To facilitate patient classification under these assumptions, we use a transductive SVM (TSVM) approach to simultaneously discover patient clusters and classify new patients. First, each patient is represented by a vector of time series features (e.g., MSE coefficients). Next, using an appropriate similarity kernel (e.g., the radial basis function kernel) we construct a similarity matrix among the time series, and perform spectral clustering with automated cluster number determination via silhouette values [14]. Finally, a separate SVM classifier is fit to each cluster. Given a collection of SVM classifiers and the associated time series features (or support vectors) and a new patient time series, we proceed by estimating a probability that the patient feature vector belongs to any one of the SVM models. This is accomplished by calculating its average similarity to the time series within that model, and normalizing the resulting similarity vector. The final classification of the new patient is a convex combination of the outputs of the individual SVM classifiers, where weights are given by the normalized similarities.

2.3. Statistical Analysis

The data was randomly partitioned 10 times into testing (20%), validation (20%) and training (60%) sets. For each data partition, we used MSE values as the primary feature vector. Each training set was utilized to identify the coefficients for the TSVM classifier which performed optimally (as measured by AUC) on the corresponding validation set. The model was then evaluated on an unseen testing set. We report the statistical properties of these AUC measures across the 10 partitions.

As a performance baseline, we report AUC measures which result from selecting the default parameter values reported on PhysioNet. These were then compared to the AUC resulting from parameter selection via global optimization. Additionally, we compared the performance of MSE versus the time series mean and standard deviation by training the same TSVM model using these variables as the feature of interest. Lastly, we compared the performance of our trained models to a TSVM classifier utilizing the APACHE score feature provided to our cohort.

2.4. Optimization

2.4.1. Parameters for Optimization

MSE requires the specification of a maximum scale factor (default: 20), a difference between consecutive scale factors (default: 1), the length of sequences to be compared (default: 2) and a similarity threshold (default: 0.15). We determined the optimal setting of these parameters using the global optimization techniques described below. Global optimization algorithms typically require the explicit specification of bounds on possible hyper-parameter values. These bounds were selected as follows: Max Scale(1-40), Scale Increase(1-4) r(0.05-0.5), m(1-4). We performed global optimization on the MSE parameters using the Multi-start Scatter search algorithm, Genetic Algorithm, and Bayesian optimization approaches. We briefly describe these methods here but encourage the readers to refer to the original papers describing these algorithms in more detail.

2.4.2. Bayesian Optimization

Bayesian optimization is a methodology for global optimization that relies on building and querying a relatively inexpensive probabilistic surrogate of the more expensive objective function. In general, the surrogate is a Gaussian process, which when combined with observations yields a convenient posterior distribution over functions. Intuitively, the optimization routine proceeds by exploring through seeking regions of high posterior uncertainty in the surrogate and exploiting by evaluating regions with a promising expected value. At each iteration the routine proposes a set of hyperparameters that maximizes the expected improvement over the best result seen. An experiment is run with these hyperparameters and then the surro-

gate model is updated with the result. This process continues over several iterations until some threshold is reached, or a maximal number of iterations surpassed. We followed the implementation of [11] in our empirical analysis.

2.4.3. Genetic Algorithm

Genetic algorithms are an established method for global optimization that imitate the process of natural selection. In this approach, an initial collection of hyper-parameters are selected, and evaluated according to the objective function. The performance of each entity regulates it's propagation into subsequent generations. Genetic algorithms employ heavy use of randomization and have several parameters that may be tuned. The genetic algorithm was parameterized to allow for an infinite number of generations with a starting population size of 100 and a termination condition of run-time exceeding 30 minuets. Starting location for parameters were drawn from a uniform distribution over the hyper-cube defining our parameter space. All other optimization options were chosen following the defaults provided by Matlab, and may be found online [15].

2.4.4. Multi-start Scatter Search Algorithm

Like the GA approach, the Scatter Search (ScS) algorithm iteratively establishes a set of possible solutions starting from a random set of starting location. Unlike GA, ScS use a deterministic process to identify the members of the next generation, such as gradient descent. Like the genetic algorithm, The Muti-start algorithm was parametrized to allow for an infinite number of iterations and a termination condition of run-time exceeding 30 minuets. Local minima were identified using constrained nonlinear minimization and the interior point algorithm [16]. All other optimization options were chosen following the defaults provided by Matlab, and may be found online [17].

3. Results

Table 1 provides a comparison of the APACHE and time-series mean and standard deviation features for the prediction of patient outcome. Table 2 provides a comparison of our method's predictive performance using MSE with default parameter values compared to parameters trained by various global optimization methods. Of the attempted optimization methods, Bayesian optimization provided the best selection of parameter values and resulting predictive performance.

In Figure 1 we illustrate the selected parameter values, and corresponding AUC of each global optimization approach across the ten testing folds. We highlight that the variability of the inferred parameter values is not constant across methods. For all but the *m* parameter, the estimates provided by Bayesian optimization are the most tightly clustered. For the *r* and *Max Scale* parameters, the difference in estimate variability is most apparent.

Table 1. *A comparison of the APACHE and time-series mean and standard deviation features for the prediction of patient outcome. Within each cell, upper values represent the 50th percentile across folds, with lower values (in parenthesis) representing the the 25th and 75th percentiles.*

	Time Series Mean + Std	**APACHE IV**
AUC (Training)	0.56 (0.52 - 0.56)	0.77 (0.75 - 0.79)
AUC (Testing)	0.54 (0.45 - 0.60)	0.68 (0.55 - 0.77)

Table 2. *A comparison of predictive performance using MSE with default parameter values compared to parameters trained by various global optimization methods. For each MSE parameter we report their cross-fold mean and standard deviation (with standard deviation in parenthesis).For the reported AUC, we report the 50th percentile in the top half of the cell and the 25th and 75th percentiles in the lower half of the cell.*

	MSE (Defaults)	**MSE (Bayesian)**	**MSE (Genetic)**	**MSE (Multi-Start)**
Max Scale	20	17.62(8.68)	23.54(14.34)	19.03(12.57)
Scale Increase	1	2.59(0.93)	2.56(1.12)	2.35(0.87)
r	0.15	0.11(0.07)	0.18(0.15)	0.18(0.1285)
m	2	2.58(0.85)	2.07(0.70)	2.53(0.87)
AUC (Training)	0.77 (0.73- 0.78)	0.77 (0.69 - 0.79)	0.77 (0.67 - 0.84)	0.73 (0.69 - 0.76)
AUC (Testing)	0.66 (0.60 - 0.72)	0.72 (0.63 - 0.78)	0.67 (0.44 - 0.78)	0.69 (0.53 - 0.72)

4. Discussion

There are several important points which are highlighted by our results. Firstly a comparison of Tables 1 and 2 demonstrates the importance of physiological dynamics (as measured by MSE) for the prediction of mortality in the sepsis cohort. When adequately tuned via optimization, we see that the MSE features facilitate classification performance which exceeds that provided by the APACHE or time-series mean and standard deviation features. This result is in agreement with the existing literature which shows that MSE based HR complexity may have a prognostic value beyond time series mean and variance [18]. Importantly, the performance of MSE above the widely employed APACHE scores highlights the need for severity of illness metrics which include measures of physiological

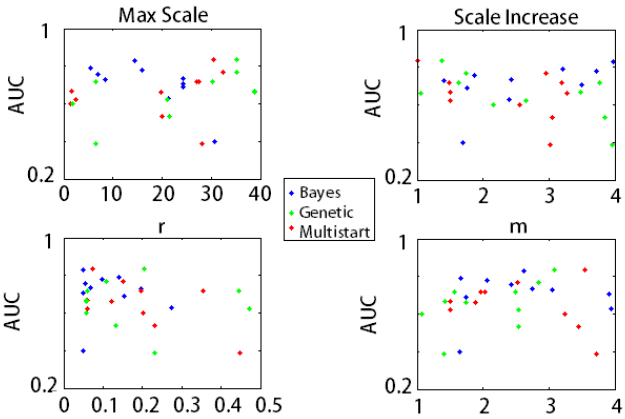

Figure 1. *Selected parameter values, and corresponding AUC of each global optimization approach, across the ten testing folds.*

dynamics.

For MSE to deliver meaningful entropy values across scales, high resolution waveforms are necessary and this often stands at odds with many conventional forms of clinical data collection, which sample at resolutions of minuets or higher. Hence, these results also motivate the value of high resolution data for prognostication, and patient monitoring in critical care settings.

The results in Table 2, clearly illustrate the beneficial effects of principled parameter selection on model performance. In general, global optimization approaches are best motivated for objective functions which are both costly to evaluate and whose performance is sensitive to parametrization. MSE is just one example of a method which requires such parametrization. Importantly, related work has called into question the setting of MSE parameter values, and eluded to the potential utility of an optimization approach, which we have now demonstrated [18]. Other commonly employed methods in biomedical signal processing may also gain from such an approach, whether it be selecting the number of Gaussian mixture components used to model a density function, the number of layers in a neural network, or the number of assumed source signals in an Independent Component Analysis.

It is important to note that while all three optimization approaches yielded parameter sets which resulted in 50th percentile AUC values above those provided by the defaults, Bayesian optimization was the clear victor, and the only method which facilitated MSE to outperform the APACHE score. This result is in line with recent literature, which has demonstrated the superiority of Bayesian optimization on several benchmark datasets [11].

The variation in estimated parameters values across our data folds shown in Fig. 1, illustrates the heterogeneous nature of ICU patients with homogeneous disease profiles. If there was a single optimal set of parameters for the entire cohort we would expect to see higher average AUC values near specific parameter values. Instead, Fig. 1 shows great variance in model AUC, across a range of parameter selections. Importantly, we see that BO was often more immune to estimate variability than the other approaches. The *max scale* and *r* values inferred by Bayesian optimization were far more tightly clustered than those provided by other methods. This fact, coupled with it's enhanced performance may indicate that patient heterogeneity manifests it's effect in some features of MSE more so than others.

It is important to note that for the same computational cost as measured in run-time, we see greater performance from BO as compared to the other techniques. These results however, should be understood in the context of the hyper-parameters we selected to govern each of the optimization approaches themselves. Indeed, It is likely that with longer run-times or denser initial sampling of the parameters space that the GA and Multi-start methods will converge to increasingly similar values as BO.

In this study, algorithms were executed sequentially but it is worth highlighting that these techniques are easily parallelized, which makes them particularly relevant in an age where massive computational clusters, such as Amazon's EC2, are increasingly available and affordable for researchers around the globe.

The importance of this approach extends beyond the immediate scope of this paper. Robust navigation and mining of physiologic time series databases, in general, requires finding similar temporal patterns of physiological responses. Detection of these complex physiological patterns not only enables demarcation of important clinical events but can also elucidate hidden dynamical structures that may be suggestive of disease processes. Some specific examples where this may also be useful include real-time detection of cardiac arrhythmia, sleep staging or detection of seizure onset. In all these cases, being able to identify a cohort of patients who exhibit similar physiological dynamics could be useful in prognosis and inform treatment strategies. Ultimately, we are hopeful that these results will encourage others in the biomedical signal processing community to employ global optimization techniques when performing parameter selection and facilitate more robust results.

ACKNOWLEDGMENT

We would like to acknowledge the support of the Salerno Foundation, the James S. McDonnell Foundation, and the National Institute of Health's Neuroimaging Training Program for their generous support of this research endeavor.