Chapter 29
Hyperparameter Selection

Franck Dernoncourt, Shamim Nemati, Elias Baedorf Kassis and Mohammad Mahdi Ghassemi

Learning Objectives

High Level:

Learn how to choose optimal hyperparameters in a machine learning pipeline for medical prediction.

Low Level:

1. Learn the intuition behind Bayesian optimization.
2. Understand the genetic algorithm and the multistart scatter search algorithm.
3. Learn the multiscale entropy feature.

29.1 Introduction

Using algorithms and features to analyze medical data to predict a condition or an outcome commonly involves choosing hyperparameters. A hyperparameter can be loosely defined as a parameter that is not tuned during the learning phase that optimizes the main objective function on the training set. While a simple grid search would yield the optimal hyperparameters by trying all possible combinations of hyper parameters, it does not scale as the number of hyperparameters and the data set size increase. As a result, investigators typically choose hyperparameters arbitrarily, after a series of manual trials, which can sometimes cast doubts on the results as investigators might have been tempted to tune the parameters specifically for the test set. In this chapter, we present three mathematically grounded techniques to automatically optimize hyperparameters: Bayesian optimization, genetic algorithms, and multistart scatter search.

To demonstrate the use of these hyperparameter selection methods, we focus on the prediction of hospital mortality for patients in the ICU with severe sepsis. The

The original version of this chapter was revised: A chapter author's name Shamim Nemati was added. The erratum to this chapter is available at 10.1007/978-3-319-43742-2_30

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_29

outcome we consider is binary: either the patient died in hospital, or survived. Sepsis patients are at high risk for mortality (roughly 30 % [1]), and the ability to predict outcomes is of great clinical interest. The APACHE score [2] is often used for mortality prediction, but has significant limitations in terms of clinical use as it often fails to accurately predict individual patient outcomes, and does not take into account dynamic physiological measurements. To remediate this issue, we investigate the use of multiscale entropy (MSE) [3, 4] applied to heart rate (HR) signals as an outcome predictor: MSE measures the complexity of finite length time series. To compute MSE, one needs to specify a set of parameters, namely the maximum scale factor, the difference between consecutive scale factors, the length of sequences to be compared and a similarity threshold. We show that using hyperparameter selection methods, the MSE can predict the patient outcome more accurately than the APACHE score.

29.2 Study Dataset

We used the Medical Information Mart for Intensive Care II (MIMIC II) database, which is available online for free and was introduced by [5, 6]. MIMIC II is divided into two different data sets:

- the Clinical Database, which is a relational database that contains structured information such as patient demographics, hospital admissions and discharge dates, room tracking, death dates, medications, lab tests, and notes by the medical personnel.
- the Waveform Database, which is a set of flat files containing up to 22 different kinds of signals for each patient, including the ECG signals.

We selected patients who suffered from severe sepsis, defined as patients with an identified infection with evidence of organ dysfunction and hypotension requiring vasopressors and/or fluid resuscitation [7]. We further refined the patient cohort by choosing patients who had complete ECG waveforms for their first 24 h in the ICU. For each patient, we extracted the binary outcome (i.e. whether they died in hospital) from the clinical database. The HR signals were extracted from the ECG signals, and patients with low quality HR were removed.

29.3 Study Methods

We compared the predictive power of the following three sets of features to predict patient outcomes: basic descriptive statistics on the time series (mean and standard deviation), APACHE IV score and MSE. Since these features are computed on time series, for each feature set we obtained a vector of time series features. Once these features were computed, we clustered patients based on these vectors using spectral clustering. The number of clusters was determined using the silhouette values [8]. This allowed us to address the high heterogeneity of the data resulting from the fact that MIMIC patients came from different care units. Lastly, for each cluster, we trained a support vector machine (SVM) classifier. To classify a new patient, we computed the distance from each cluster center, and computed the output of each SVM classifier: to make the final decision on the predicted outcome, we computed a weighted average of the output of each SVM classifier, where the weights were the distance from each cluster center. This method of combining clustering with SVM is called transductive SVM. We used the area under the receiver operating characteristic (ROC) curve (AUROC, often named more simply and ambiguously AUC) as the performance metric for the classification. Figure 29.1 illustrates the functioning of transductive SVMs.

MSE may be understood as the set of sample entropy values for a signal which is averaged over various increasing segment lengths. The MSE, y, was computed as follows:

$$y_j^\tau = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i$$

where:

- x_i is the signal value at sample I,
- j is the index of the window to be computed,
- τ is the scale factor,
- Y is the length of sequences to be compared,
- Z is the similarity threshold.

Additionally, we have the following parameters:

- the maximum scale factor,
- the scale increase, which is the difference between consecutive scale factors,
- the similarity criterion or threshold, denoted r.

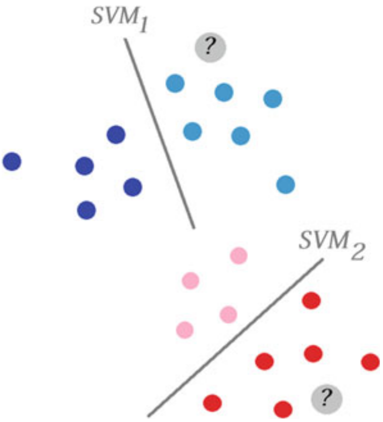

Fig. 29.1 Transductive SVM: clustering is performed first, then a convex combination of the SVM outputs is used to obtain the final prediction probability

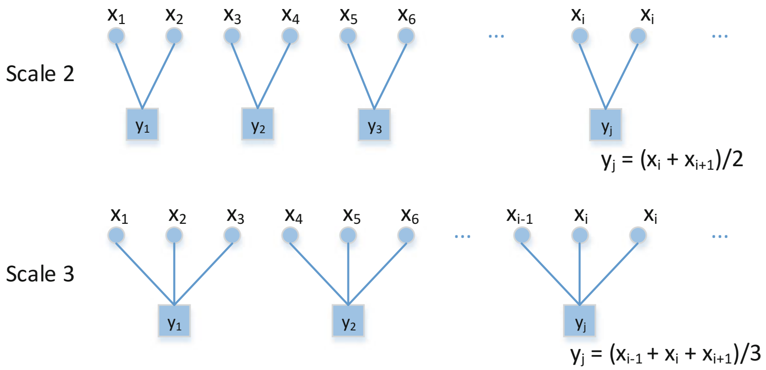

Fig. 29.2 Illustration of various scales from Costa et al. Only scales 2 and 3 are displayed. x_i is the signal value at sample i

Figure 29.2 shows how y is computed for different scales.

To select the best hyperparameters for the MSE, we compared three hyperparameter optimization techniques: Bayesian optimization, genetic algorithms, and multistart scatter search.

Bayesian optimization builds the distribution P($y_{\text{test}}|y_{\text{train}}$, x_{train}, x_{test}), where x_{train} is the set of MSE parameters that were used to obtain the y_{train} AUROCs, x_{test} is a new set of MSE parameters, and y_{test} is the AUROC that would be obtained using the new MSE parameters. To put it otherwise, based on the previous observations on MSE parameters and achieved AUROCs, the Bayesian optimization predicts what AUROC a new set of MSE parameters will yield. Each time a new AUROC is computed, the set of MSE parameters as well as the AUROC is added to x_{test} and y_{test}. At each iteration, we can either explore, i.e. compute y_{test} for which the distribution P has a high variance, or exploit, i.e. compute y_{test} for which the distribution P has a low variance and high expectation. An implementation can be found in [9].

A genetic algorithm is an optimization algorithm based on the principle of Darwinian natural selection. A population is comprised of sets of MSE parameters. Each set of MSE parameters is evaluated based on the AUROC it achieved. The sets of MSE parameters with low AUROCs are eliminated. The surviving sets of MSE parameters are mutated, i.e. each parameter is slightly modified, to create new sets of MSE parameters, which form a new population. By iterating through this process, the new sets of MSE parameters yield increasingly high AUROCs. We set the population size of 100, and ran the optimization for 30 min. The first population was drawn randomly.

The multistart scatter search is similar to the genetic algorithm, the only difference residing in the use of a deterministic process to identify the individuals of the next population such as gradient descent.

Figure 29.3 summarizes the machine learning pipeline presented in this section.

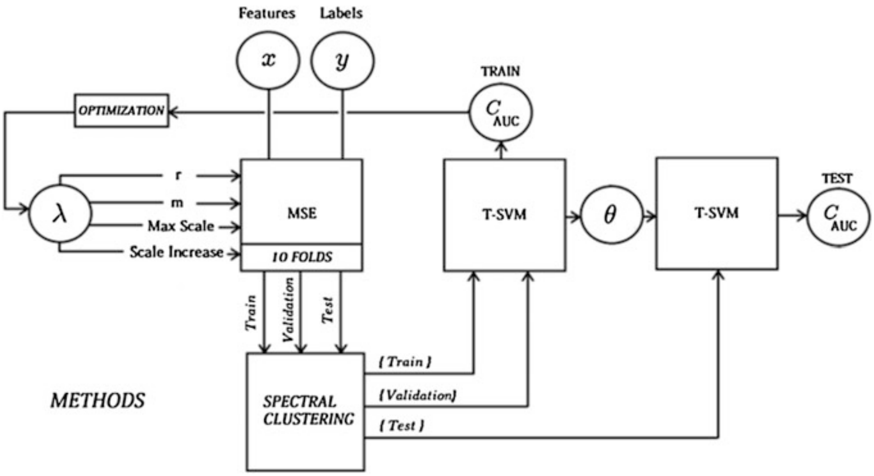

Fig. 29.3 The entire machine learning pipeline. The MSE features are computed from the input x using the parameters r, m, max scale and scale increase. 10 folds are created

The data set was split into testing (20 %), validation (20 %) and training (60 %) sets. In order to ensure robustness of the result, we used 10-fold cross-validation, and the average AUROC over the 10 folds. To make the comparison fair, each hyperparameter optimization technique was run the same amount of time, viz. 30 min.

29.4 Study Analysis

Table 29.1 contains the results for all three sets of features we considered. For the MSE features, Table 29.1 presents the results achieved by keeping the default hyperparameters, or by optimizing them using one of the three hyperparameter optimization techniques we presented in the previous section.

The first set of features, namely the basic descriptive statistics (mean and standard deviation), yields an AUROC of 0.54 on the testing set, which is very low since a random classifier yields an AUROC of 0.50. The second set of features, APACHE IV, achieves a much higher AUROC, 0.68, which is not surprising as the APACHE IV was designed to be a hospital mortality assessment for critically ill patients. The third set of features based on MSE performs surprisingly well with the default values (AUROC of 0.66), and even better when optimized with any of the three hyperparameter optimization techniques. The Bayesian optimization yields the highest AUROC, 0.72.

Table 29.1 Comparison of APACHE feature, time-series mean and standard deviation features, and MSE feature with default parameters or optimized with Bayesian optimization, genetic algorithms, and multistart scatter search, for the prediction of patient outcome

	Max scale	Scale increase	*r*	*m*	AUROC (training)	AUROC (testing)
Time series: mean and standard deviation					0.56 (0.52–0.56)	0.54 (0.45–0.60)
APACHE IV					0.77 (0.75–0.79)	0.68 (0.55–0.77)
MSE (defaults)	20	1	0.15	2	0.77 (0.73–0.78)	0.66 (0.60–0.72)
MSE (Bayesian)	17.62 (8.68)	2.59 (0.93)	0.11 (0.07)	2.58 (0.85)	0.77 (0.69–0.79)	0.72 (0.63–0.78)
MSE (genetic)	23.54 (14.34)	2.56 (1.12)	0.18 (0.15)	2.07 (0.70)	0.77 (0.67–0.84)	0.67 (0.44–0.78)
MSE (multi-start)	19.03 (12.57)	2.35 (0.87)	0.18 (0.128)	2.53 (0.87)	0.73 (0.69–0.76)	0.69 (0.53–0.72)

For each MSE parameter we report their cross-fold mean and standard deviation (with standard deviation in parenthesis). For the reported AUROC, 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

29.5 Study Visualizations

Figure 29.4 provides an insight into the MSE parameters selected by the three hyperparameter selection techniques over the 10-fold cross-validation. Each point represents a parameter value optimized by a given hyperparameter selection technique for a unique data fold. For all 4 MSE parameters, we observe a great variance: this indicates that there is no clear global optimum, but instead there exist many MSE parameter sets that yield a high AUROC.

Interestingly, in this experiment the Bayesian optimization is more robust to the parameter variance, as shown by the confidence intervals around the AUROCs: most AUROCs reached by Bayesian optimization are high, unlike genetic algorithms and multistart scatter search. The two latter techniques are susceptible to premature convergence, while Bayesian optimization has a better exploration-exploitation tradeoff.

We also notice that the max scale and the *r* values reached by Bayesian optimization have a lower variance than genetic algorithms and multistart scatter search. One might hypothesize that heterogeneity across patients might be reflected more in the scale increase and *m* MSE parameters than in the max scale and *r* parameters.

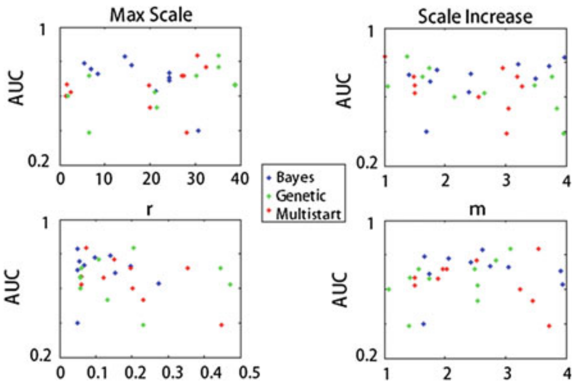

Fig. 29.4 The impact of the MSE parameters on the outcome prediction AUROC

29.6 Study Conclusions

The results of this case study demonstrate two main points. First, from a medical standpoint, they underline the possible benefit of utilizing dynamic physiologic measurements in outcome prediction for ICU patients with severe sepsis: the data from this study indeed suggest that utilizing these physiological dynamics through MSE with optimized hyperparameters yields improved mortality prediction compared with the APACHE IV score. Physiological signals sampled at high-frequency are required for the MSE features to be meaningful, highlighting the need for high-resolution data collection, as opposed to some existing methods of data collection where signal samples are aggregated at the second or minute level, if not more, before being recorded.

Second, from a methodological standpoint, the results make a strong case for the use of hyperparameter selection techniques. Unsurprisingly, the results obtained with the MSE features are highly dependent on the MSE hyperparameters. Had we not used a hyperparameter selection technique and instead kept the default value, we would have concluded that APACHE IV provides a better predictive insight than MSE, and therefore missed the importance of physiological dynamics for prediction of patient outcome. Bayesian optimization seems to yield better results than genetic algorithms and multistart scatter search.

29.7 Discussion

There is still much room for further investigation. We focused on ICU patients with severe sepsis, but many other critically ill patient cohorts would be worth investigating as well. Although we restricted our study to the use of MSE and HR alone, it would be interesting to integrate and combine other disease characteristics and physiological signals. For example, [10] used Bayesian optimization to find the most optimal wavelet parameters to predict acute hypotensive episodes. Perhaps combining dynamic blood pressure wavelets with HR MSE, and even other dynamic data as well such as pulse pressure variation, would further optimize and tune the mortality prediction model. In addition there exist other scores to predict group mortality such as SOFA and SAPS II, which would provide useful baselines in addition to APACHE [11].

The scale of our experiments was satisfying for the case study's goals, but some other investigations might require a data set that is an order of magnitude larger. This might lead one to adopt a distributed design to deploy the hyperparameter selection techniques. For example, [12] used a distributed approach to hyperparameter optimization on 5000 patients and over one billion blood pressure beats. [13, 14] present another large-scale system to use genetic algorithms for blood pressure prediction.

Lastly, a more thorough comparison between hyperparameter selection techniques would help comprehend why a given hyperparameter selection technique performs better than others for a particular prediction problem. Especially, the hyperparameter selection techniques also have parameters, and a better understanding of the impact of these parameters on the results warrant further investigation.

29.8 Conclusions

In this chapter, we have presented three principled hyperparameter selection methods. We applied them to MSE, which we computed on physiological signals to illustrate their use. More generally, these methods can be used for any algorithm and feature where hyperparameters need to be tuned.

ICU data provide a unique opportunity for this type of research with routinely collected continuously measured variables including ECG waveforms, blood pressure waveforms from arterial lines, pulse pressure variation, pulse oximetry as well as extensive ventilator data. These dynamic physiologic measurements could potentially help unlock better outcome metrics and improve management decisions in patients with acute respiratory distress syndrome (ARDS), septic shock, liver failure or cardiac arrest, and other extremely ill ICU patients. Outside of the ICU, dynamic physiological data is routinely collected during surgery by the anesthesia team, in cardiac units with continuous telemetry and on Neurological care units with routine EEG measurements for patients with or at risk for seizures. As such the potential applications of MSE with hyperparameter optimization are extensive.

References

1. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR (2001) Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med 29(7):1303–1310
2. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L, Moody GB, Heldt T, Kyaw TH, Moody BE, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access ICU database. Crit Care Med 39(5):952–960. doi:10.1097/CCM.0b013e31820a92c6
3. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101(23): e215–e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]
4. Mayaud L, Lai PS, Clifford GD, Tarassenko L, Celi LA, Annane D (2013) Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension*. Crit Care Med 41(4):954–962
5. Ng AY, Jordan MI, Weiss Y et al (2002) On spectral clustering: analysis and an algorithm. Adv Neural Inf Process Syst 2:849–856
6. Snoek J, Larochelle H, Adams RP (2012) Practical Bayesian optimization of machine learning algorithms. Adv Neural Inf Process Syst 2951–2959
7. Dernoncourt F, Veeramachaneni K, O'Reilly U-M (2015) Gaussian process-based feature selection for wavelet parameters: predicting acute hypotensive episodes from physiological signals. In: Proceedings of the 2015 IEEE 28th international symposium on computer-based medical systems. IEEE Computer Society
8. Castella X et al (1995) A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study. Crit Care Med 23(8):1327–1335
9. Dernoncourt F, Veeramachaneni K, O'Reilly U-M (2013c) BeatDB: a large-scale waveform feature repository. In: NIPS 2013, machine learning for clinical data analysis and healthcare workshop
10. Hemberg E, Veeramachaneni K, Dernoncourt F, Wagy M, O'Reilly U-M (2013) Efficient training set use for blood pressure prediction in a large scale learning classifier system. In: Proceeding of the fifteenth annual conference companion on genetic and evolutionary computation conference companion. ACM, New York, pp 1267–1274
11. Hemberg E, Veeramachaneni K, Dernoncourt F, Wagy M, O'Reilly U-M (2013) Imprecise selection and fitness approximation in a large-scale evolutionary rule based system for blood pressure prediction. In: Proceeding of the fifteenth annual conference companion on genetic and evolutionary computation conference companion. ACM, New York, pp 153–154
12. Knaus WA et al (1981) APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 9(8):591–597
13. Costa M, Goldberger AL, Peng C-K (2005) Multiscale entropy analysis of biological signals. Phys Rev E 71:021906
14. Costa M, Goldberger AL, Peng C-K (2002) Multiscale entropy analysis of physiologic time series. Phys Rev Lett 89:062102