Research on Prediction of the Effects of Oil-Increasing Measures Driven by Data

Lu Yang[1], Kai Zhang[1](✉), Li- ming Zhang[1], Hua- qing Zhang[2], Xia Yan[1], Pi-yang Liu[3], and Jun Yao[1]

[1] School of Petroleum Engineering, China University of Petroleum (East China), Qingdao, China
zhangkai@upc.edu.cn

[2] College of Science, China University of Petroleum (East China), Qingdao, China

[3] Civil Engineering School, Qingdao University of Technology, Qingdao, China

Abstract. A large number of major oil fields in China have entered the late stages of development, and the decreasing production is increasingly unable to meet the continuously growing demand for energy. Therefore, it is crucial for oilfield production to accurately and rapidly predict the effects of production-increasing measures based on existing data. This paper comprehensively considers three types of data: geological static parameters, production dynamic parameters, and process parameters of measures. Advanced machine learning algorithms such as random forest (RF), support vector regression (SVR), and extreme gradient boosting (XGBoost) are separately used, together with data augmentation techniques and Bayesian optimization algorithms to construct the different enhancing production through measures prediction model. The best prediction model is optimized by comparing the scores of each model. The results of a comprehensive comparison of various models based on the mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R2) indicate that the model based on the extreme gradient boosting algorithm performs the best. The application of data augmentation and optimization algorithms significantly improves the model performance. The accuracy of predicting the oil production enhancement effect for a given measure can reach over 90%. Compared with traditional methods for predicting the effects of measures, this paper addresses the issues of long computational time in numerical simulations and difficulty in exploring the mechanism of

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20–22 September 2023.

This paper was selected for presentation by the IFEDC Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the IFEDC Technical Team and are subject to correction by the author(s). The material does not necessarily reflect any position of the IFEDC Technical Committee its members. Papers presented at the Conference are subject to publication review by Professional Team of IFEDC Technical Committee. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of IFEDC Organizing Committee is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of IFEDC. Contact email: paper@ifedc.org.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 21–30, 2024.
https://doi.org/10.1007/978-981-97-0272-5_2

oil production enhancement measures in depth, and achieves a rapid and accurate prediction of the multidimensional effect of measures for increasing oil production. This paper employs machine learning algorithms to fully explore the relationship between three types of data and oil production enhancement effects, accurately predicting the effect of measures for increasing oil production. It provides a technical foundation for selecting reasonable measures to increase oil production in oilfields and has certain guiding significance for actual production.

1 Introduction

As an irreplaceable strategic resource, petroleum plays a critical role in a country's power and economic development. Maintaining and increasing oil production has always been an important energy goal for nations. However, due to China's rapid economic growth over the past few decades, the country's consumption of petroleum has been steadily increasing. Nevertheless, as most of China's major oil fields have entered the middle and late stages of development, issues such as increased water content and reservoir damage have resulted in a decline in production capacity that is no longer sufficient to meet current energy demands [1]. This has led to a severe dependence on foreign oil and gas resources, with China's external oil dependency exceeding 70% in 2020. Large-scale oil imports could easily subject China to geopolitical risks, significantly threatening the country's energy security [2]. Therefore, implementing reasonable measures to increase oil production is imperative for China to address its energy gap, stabilize domestic economic development, and alleviate its energy crisis. However, with a wide variety of measures available and varying effectiveness, accurately predicting the effectiveness of such measures is crucial for oil field production.

In the field of measure effectiveness prediction, conventional methods such as the water flooding characteristic curve method and the Weng cycle method have limitations in their applicability due to various assumptions and complex formulas [3]. Although numerical simulations have been attempted to predict the effectiveness of measures, their applicability is restricted due to the complex mechanisms of measures to increase oil production and the expensive computations involved [4]. Research on machine learning-based measure effectiveness prediction is still in the exploratory stage, with a primary focus on production forecasting. There has been limited consideration of including process parameters in the evaluation of measures, as the limited sample size of measure wells restricts research in this direction to primarily fracturing methods [5].

In recent years, the revolutionary development of artificial intelligence (AI) technology has attracted widespread attention from various industries due to its powerful generalization ability and rapid response speed [6]. The petroleum industry has also accumulated a large amount of historical data in production, and machine learning has shown great potential in the field of petroleum engineering [7]. As an alternative data-driven approach, machine learning can extract information from a large amount of historical data and construct regression or classification prediction models [8]. Many supervised machine learning methods, including linear regression, support vector machines, neural networks, etc., have been used to predict production decline, optimize water injection schemes, characterize reservoir permeability, and generate complex geological facies [9].

Based on the above content, in order to accurately and rapidly predict the effect of oil recovery measures, this paper proposes a data-driven approach for predicting the effect of oil recovery measures. Advanced machine learning algorithms, including Random Forest (RF), Support Vector Regression Machine (SVR), and Extreme Gradient Boosting (XGBoost), are used to explore the influence of three types of data on the effect of oil recovery measures, namely geological static parameters, production dynamic parameters, and process parameters. A prediction model is built, and data augmentation is employed to address the problem of insufficient samples, which improves the quality of the sample dataset. In the hyperparameter tuning stage of the model, the Bayesian optimization algorithm is introduced to solve the problem of difficult manual parameter tuning and further improve the model accuracy. After comparative experiments, the XGBoost algorithm-based oil recovery measure effect prediction model is selected, and the accuracy in the test set can reach over 90%.

2 Methodology

2.1 Feature Engineering

Feature engineering is the process of taking the raw input data and creating new features. To make the raw data more informative, it selects, extracts, and transforms meaningful features from the raw data. Feature engineering involve various techniques, including data cleaning, data normalization, data scaling, data augmentation, data encoding, dimensionality reduction, and feature selection. The source data for this study is the actual recorded data from the oil field, which has poor data quality. Therefore, feature engineering is crucial in processing the data. In addition to common data cleaning, normalization, and correlation analysis, this paper also employs the SMOTE oversampling technique as a data augmentation method.

Synthetic Minority Over-Sampling Technique (SMOTE)
SMOTE [10] is an approach to the construction of classifiers from imbalanced datasets is described. It is used to address the problem of imbalanced class distribution in data by synthesizing data through a combination of over-sampling the minority class and under-sampling the majority class [11]. The specific steps are as follows:

(1) For each sample x in the minority class, calculate its k-nearest neighbors to all samples in the minority class using Euclidean distance as the metric. The formula is:

$$d(s_l, s_k) = \sqrt{\sum_{j=1}^{m} \left(s_{lj} - s_{kj}\right)^2} \tag{1}$$

(2) Determine a sampling rate based on the imbalance ratio and set a sampling multiplier N. For each minority sample x, randomly select several samples from its k-nearest neighbors, denoted as $\mathrm{x_n}$.

(3) For each randomly selected neighbor $\mathrm{x_n}$, construct a new sample with the original sample according to the following formula.

$$x_{\text{new}} = x + \text{rand}(0, 1) \times (\tilde{x} - x) \tag{2}$$

2.2 Regression Prediction Algorithm

Support Vector Regression (SVR)

SVR [12] is a type of machine learning algorithm used for regression analysis. It is based on the Support Vector Machine (SVM) algorithm and is used to build models that can predict continuous output variables. The basic principle of SVR is to find a hyperplane in a high-dimensional space that best separates the data into different classes. In the case of regression, the hyperplane is used to predict the value of the outcome variable based on the input features. Therefore, the SVR problem can be formalized as:

$$\underset{w,b}{min}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} l_\epsilon(f(x_i), y_i) \tag{3}$$

In which, C is the regularization constant, lϵ is the ϵ-insensitive loss function. After introducing slack variables and Lagrange multipliers and taking partial derivatives, the formula of SVR can be expressed as:

$$f(x) = \sum_{i=1}^{m}\left(\widehat{\alpha}_i - \alpha_i\right)\kappa\left(x_i^T x\right) + b \tag{4}$$

where $\kappa\left(x_i^T x\right)$=$\varnothing(x_i)^T\varnothing\left(x_j\right)$ is the kernel function.

Random Forest(RF)

RF [13] is a popular ensemble learning algorithm used for classification, regression, and other machine learning tasks. The algorithm combines multiple decision trees to create a "forest" of trees that work together to make predictions. In regression problems, the output of each decision tree is averaged to obtain the final regression result [14]. The specific idea is as follows:

(1) Assuming that the training dataset contains N data objects, a training dataset is constructed by randomly sampling M samples with replacement using the bootstrap method, where each sample is not completely identical to the others.
(2) Assuming that each sample data has X features, a subset of x (x $< \; =$ X) features is randomly selected from all the features, and the best splitting attribute is chosen as the node to grow the CART decision tree, with k remaining constant during the tree growing process.
(3) Repeat the above steps to build n CART trees, and obtain the final prediction by averaging the outputs of these decision trees.

$$\widehat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b\left(x'\right) \tag{5}$$

eXtreme Gradient Boosting(XGBoost)

XGBoost [15] is a highly efficient gradient boosting decision tree algorithm that uses the ensemble idea - the Boosting idea - to integrate multiple weak learners into a strong learner through a certain method. Its algorithmic process is as follows:

(1) Set the model to begin with a constant value:

$$\widehat{f}_{(0)}(x) = \arg\underset{\theta}{min}\sum_{i=1}^{N} L(y_i, \theta) \tag{6}$$

(2) Calculate the gradients and hessians:

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)} \tag{7}$$

$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)} \tag{8}$$

(3) Train a base learner on the training set by solving the following optimization problem:

$$\hat{\phi}_m = \underset{\phi\in\Phi}{\arg min} \sum_{i=1}^{N} \frac{1}{2}\hat{h}_m(x_i)\left[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i)\right]^2 \tag{9}$$

$$\hat{f}_m(x) = \alpha\hat{\phi}_m(x) \tag{10}$$

(4) Modify the model:

$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x) \tag{11}$$

(5) Output:

$$\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x) \tag{12}$$

2.3 Optimization Algorithm

An optimization algorithm refers to the process of minimizing or maximizing an objective function, subject to given constraints, by finding one or more optimal or near-optimal solutions. This paper introduces a tree-structured Bayesian optimization algorithm to tune hyperparameters of the production enhancement effect prediction model. This method solves the problem of obtaining the optimal prediction model through manual tuning, providing a more efficient and effective approach.

Tree-structured Parzen Estimator(TPE)

TPE [16] uses two density functions to define $p(x|y)$:

$$p(x|y) = \begin{cases} l(x) & \text{if } y < y^* \\ g(x) & \text{if } y \geq y^* \end{cases} \tag{13}$$

In the above equation, $l(x)$ is established using the observation space $\{x^{(i)}\}$ and the corresponding loss $f(x^{(i)})$ is less than y*, while $g(x)$ is established using the remaining observations. The TPE-based method relies on a value of y* greater than the best observed value of $f(x)$, so that some points can be used to build $l(x)$. TPE adopts expected improvement (EI) as the acquisition function. However, since it is impossible to obtain the posterior probability $p(x|y)$, Bayesian formula is employed to transform the acquisition function:

$$\mathrm{EI}_{y^*}(x) = \int_{-\infty}^{y^*}(y^* - y)p(y|x)dy = \int_{-\infty}^{y^*}(y^* - y)\frac{p(x|y)p(y)}{p(x)}dy \tag{14}$$

In this equation, y* represents a threshold value. Let $\gamma = p(y < y^*)$ denote a certain quantile used in the TPE algorithm to partition $l(x)$ and $g(x)$. The value of γ is in the range of (0, 1). The final simplified formula is:

$$EI_{y^*}(x) = \frac{\gamma y^* \updownarrow(x) - \updownarrow(x)\int_{-\infty}^{y^*}p(y)dy}{\gamma\updownarrow(x) + (1-\gamma)g(x)} \propto \left(\gamma + \frac{g(x)}{\updownarrow(x)}(1-\gamma)\right)^{-1} \tag{15}$$

3 Experiment and Result

3.1 Introduction to the Dataset

This study collected data on all oil production enhancement measures implemented in a certain block of an oilfield from 2017 to the present, including acidification, unclogging, and water flooding. After selecting wells where the measures were effective and conducting data cleaning and correlation analysis, the sample size of wells subjected to acidification and unclogging measures was too small to support machine learning analysis. Therefore, this study ultimately chose water flooding measures as an example for oil production enhancement prediction. The sample database contains 147 wells that achieved oil production enhancement after water flooding measures were implemented. The input variables include eight geological static features, four production dynamic features, and two measure process features, as shown in Table 1.

Table 1. Feature presentation table.

Feature category	Feature name
Geological static features	Porosity of injection well, Permeability of injection well, Thickness of injection well, Oil temperature of injection well, Viscosity of crude oil of injection well, Permeability of production well, Porosity of production well, Distance between wells
Production dynamic features	Injection well pressure before the measure, Injection well daily volume before the measure, Daily fluid production before the measure, Daily oil production before the measure
measure process features	Measure chemicals, Chemicals injection volume

According to statistics, there are three different chemical types A, B and C in 147 profile control measures samples, and the sample sizes of different classes of chemicals are significantly different, among which 96 are of type A, 33 are of type B and 18 are of type C. Unbalanced sample distribution has a great influence on the learning and prediction of machine learning model. Therefore, this study adopted the data enhancement method of oversampling to expand the data, so as to ensure the same sample size of the three measures. Generate new 410 after SMOTE oversampling and screening and use this new 410 as the data set for the forecast model.

3.2 Experimental Setting

To select the optimal predictive model for effects of oil-increasing measures, this study compared three commonly used machine learning algorithms for regression problems in petroleum engineering: Support Vector Regression (SVR), Random Forest (RF), and XGBoost. Predictive models were constructed for each algorithm and their performance was evaluated under multiple loss functions. The best-performing predictive model was determined, and an optimization algorithm was introduced to fine-tune the model's hyperparameters, further enhancing its predictive accuracy.

3.3 Result

Based on three different algorithms, data-driven predictive models were developed for oil-increasing measures using geological static parameters, production dynamic parameters, and process parameters as inputs, and post-measure oil production as output. The training and testing sets were divided in a 9:1 ratio. The prediction results of the different algorithms are shown in Fig. 1. By comparing the performance of the models using the same testing set, it can be observed that the predictive accuracy of XGBoost and RF algorithms are significantly higher than that of SVR algorithm.

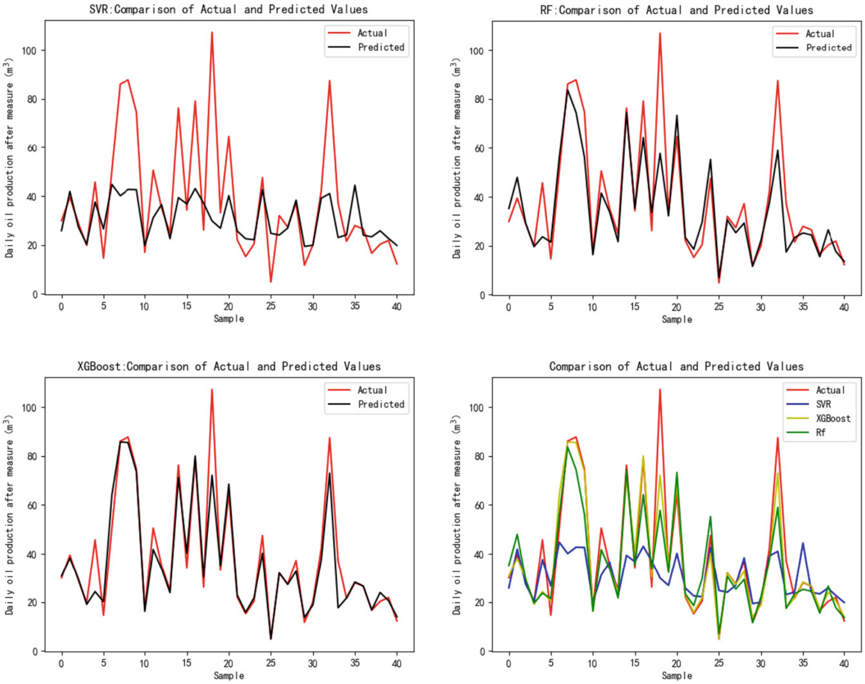

Fig. 1. The effect comparison of different algorithms

To conduct a precise analysis of the prediction performance of RF and XGBoost algorithms, this study comprehensively evaluated their performance using three loss functions: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2), as shown in Table 2. From the numerical results, it can be seen that XGBoost algorithm outperforms the other two algorithms. Therefore, XGBoost was selected for further research.

Table 2. Evaluation of three algorithms

Algorithms	MAE	MSE	R^2
SVR	13.12	453.82	0.26
RF	7.09	136.76	0.78
XBGoost	4.39	68.22	0.88

The above model obtained the optimal results by manually adjusting the hyperparameters after determining their approximate range using grid search. However, manual tuning of hyperparameters can hardly result in the best combination of model parameters, and there is still room for improvement in hyperparameter performance. Therefore, this study introduced a Bayesian optimization algorithm based on Tree-structured Parzen Estimator (TPE) to optimize the hyperparameters of the XGBoost prediction model. The final optimization result returned the maximum value of R2, and the optimized parameters are shown in Table 3.

Table 3. Hyperparameter optimization results

Hyperparameter	optimization results	Hyperparameter	optimization results
max_depth	5	n_estimators	498
learning_rate	0.0581	gamma	6.1491
colsample_bytree	0.9185	subsample	0.5102
reg_alpha	0.0196	reg_lambda	1.5302e-07

The performance of the TPE-XGBoost model for predicting the effect of enhanced oil recovery measures, incorporating the optimization algorithm, is shown in Fig. 2 and Table 4. It can be observed that the introduction of the optimization algorithm improves the performance of the predictive model, with the optimized model outperforming the non-optimized model under all three loss functions. The final predictive accuracy (R^2) can exceed 90%.

Table 4. The effect after hyperparameter optimization

Algorithms	MAE	MSE	R^2
XBGoost	4.39	68.22	0.88
TPE-XGBoost	4.11	58.19	0.91

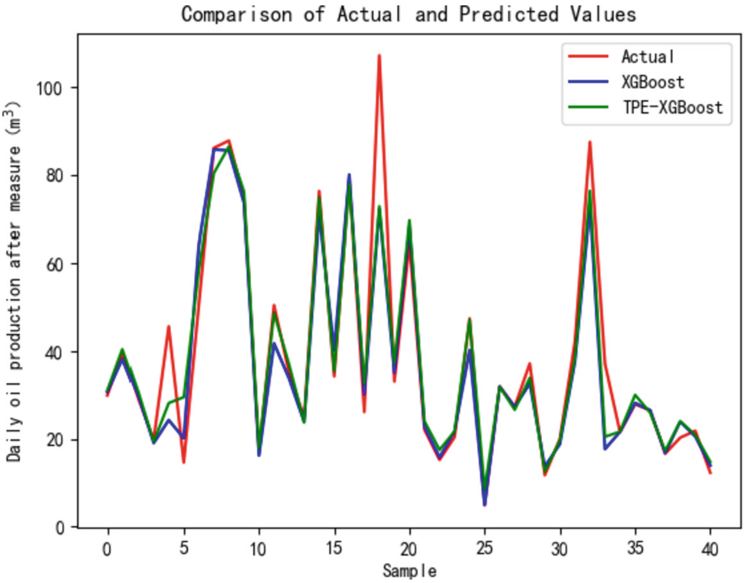

Fig. 2. The effect after hyperparameter optimization

4 Conclusion

This article proposes a data-driven method for predicting the effects of oil-increasing measures based on the TPE-XGBoost algorithm. This method first enhances the data samples to some extent, which alleviates the problem of insufficient sample size. At the same time, the model comprehensively considers three types of features: geological static parameters, production dynamic parameters, and measure process parameters, deeply mining their relationship with the effect of increasing oil, and automatically optimizing the model hyperparameters to achieve the prediction of daily oil production after the measures, which significantly improves the prediction accuracy compared with other algorithms and can reach over 90%. However, the current research is still limited by the insufficient quality of on-site data. In future research, in addition to obtaining high-quality data from the source, high-level feature engineering will also be the next research focus. In addition, incorporating economic indicators into machine learning is also a future direction of research.

Acknowledgments. This work is supported by the National Natural Science Foundation of China under Grant 52274057, 52074340 and 51874335, the Major Scientific and Technological Projects of CNPC under Grant ZD2019–183-008, the Major Scientific and Technological Projects of CNOOC under Grant CCL2022RCPS0397RSN, the Science and Technology Support Plan for Youth Innovation of University in Shandong Province under Grant 2019KJH002, 111 Project under Grant B08028.

References

1. Chen, S.Y., Zhang, Q., Mclellan, B., et al.: Review on the petroleum market in China: history, challenges and prospects. Pet. Sci. **17**, 1779–1794 (2020)
2. Gong, X., Sun, Y., Du, Z.: Geopolitical risk and China's oil security. Energy Policy **163**, 112856 (2022)
3. Alfarge, D., Wei, M., Bai, B.: Evaluating the performance of hydraulic-fractures in unconventional reservoirs using production data: comprehensive review. J. Nat. Gas Sci. Eng. **61**, 133–141 (2019)
4. Zhang, Q., Zhu, W., Liu, W., Yue, M., Song, H.: Numerical simulation of fractured vertical well in low-permeable oil reservoir with proppant distribution in hydraulic fracture. J. Petrol. Sci. Eng. **195**, 107587 (2020)
5. Hassan, A.M., Aljawad, M.S., Mahmoud, M.A.: Predicting the productivity enhancement after applying acid fracturing treatments in naturally fractured reservoirs utilizing artificial neural network. In: Abu Dhabi International Petroleum Exhibition & Conference. OnePetro (2021)
6. Vaishya, R., Javaid, M., Khan, I.H., Haleem, A.: Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab. Syndr. **14**(4), 337–339 (2020)
7. Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., Oza, H.: Application of machine learning and artificial intelligence in oil and gas industry. Petrol. Res. **6**(4), 379–391 (2021)
8. Koroteev, D., Tekic, Z.: Artificial intelligence in oil and gas upstream: trends, challenges, and scenarios for the future. Energy AI **3**, 100041 (2021)
9. Xue, L., Liu, Y., Xiong, Y., Liu, Y., Cui, X., Lei, G.: A data-driven shale gas production forecasting method based on the multi-objective random forest regression. J. Petrol. Sci. Eng. **196**, 107801 (2021)
10. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 1–48 (2019)
11. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
12. Awad, M., Khanna, R.: Support vector regression. Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, 67–80 (2015)
13. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
14. Grömping, U.: Variable importance assessment in regression: linear regression versus random forest. Am. Stat. **63**(4), 308–319 (2009)
15. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
16. Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. In: Advances in Neural Information Processing Systems, vol. 24 (2011)