The Importance of Feature Selection Methods for the Error Prediction Process of a Digital Twin

Şebnem Özdemir^{1*}, Alptekin Erkollar², and Birgit Oberer²

1 Beykent University, 2 Sakarya University, * Corresponding author, sebnemozde@gmail.com

Abstract

The idea of building a digital twin is related to simultaneously creating a model that becomes a transportation vehicle for data within the information life cycle. In order to create such model, there should be well-defined feature space. Because of the "curse of dimensionality", while the complexity of the model exponentially increases, the accuracy rate of the model decreases. In this study, the importance of the methods chosen for dimensionality reduction while creating a model setup, which can predict the error on a digital twin, is presented with an exemplary implementation. Four different dimension reduction methods, PCA, Conventional PCA, WPCA, and Mars, were applied to dataset with 89016 observation values and 590 different attributes, in order to predict error via Non-linear SVM with Polynomial kernel. According to results WPCA and MARS methods, predicted the error more successfully than others. As a result, the feature extraction solutions, that the methods provide, affected the performance of the designed models.

Keywords: Data science, Digital twin, Feature selection, PCA, SVM.

Citation: Özdemir, Ş., Erkollar, A., Oberer, B. (2018, October) The Importance of Feature Selection Methods for the Error Prediction Process of a Digital Twin. Paper presented at the Fifth International Management Information Systems Conference. **Editor:** H. Kemal İlter, Ankara Yıldırım Beyazıt University, Turkey

Received: August 19, 2018, Accepted: October 18, 2018, Published: November 10, 2018

Copyright: © 2018 IMISC Özdemir et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

IMISC 2018 PAPER

The Importance of Feature Selection Methods for the Error Prediction Process of a Digital Twin

Abstract

The idea of building a digital twin is related to simultaneously creating a model that becomes a transportation vehicle for data within the information life cycle. In order to create such model, there should be well-defined feature space. Because of the "curse of dimensionality", while the complexity of the model exponentially increases, the accuracy rate of the model decreases. In this study, the importance of the methods chosen for dimensionality reduction while creating a model setup, which can predict the error on a digital twin, is presented with an exemplary implementation. Four different dimension reduction methods, PCA, Conventional PCA, WPCA, and Mars, were applied to dataset with 89016 observation values and 590 different attributes, in order to predict error via Non-linear SVM with Polynomial kernel. According to results WPCA and MARS methods, predicted the error more successfully than others. As a result, the feature extraction solutions, that the methods provide, affected the performance of the designed models.

Keywords

Data Science, Digital Twin, Feature Selection, PCA, SVM

Introduction

Working correctly and giving necessary reactions against external effects for a Cyber-Physical System (CPS) design is closely related with the level of success of the models, which the system components are designed with. Those models can be qualified as the ultra-high fidelity simulations, which include the machines in the real world, all applications regarding to this machines and the relationships between each other (Gabor, Belzner, Kiermeier, Beck, & Neitz, 2016). The primary function of these simulations also called digital twin is to actualize all events defined in the twin with the highest accuracy (Tuegel, Ingraffea, Eason, & Spottswood, 2011) (Glassen & Stargel, 2012). In addition to that mission, digital twins are also tasked to predict the possible behaviors while the system, which they are a part of, is operating. Just being designed with a high-level simulation is not enough for this function of a digital twin. It has to collect and process all required data for the system, which it is a part of, and increase the experience of the system regarding giving action to a reaction (Belzner, Hennicker, & Wirsing, 2015). Gaining and increasing experience in this way coincides with the definition of machine learning of Mitchell (1997) for a digital twin. So indeed, when the algorithm is considered in terms of experience and task, design of effective algorithms (Mohri, Rostamizadeh, &

Talwalkar, 2012), actualization of the learning as the machine's experiences are increasing in the light of these algorithms (Alpaydın, 2014) and the design of the software and programs which produce rules thanks to the dataset worked on (Harrington, 2012; Kodratoff & Michalski, 2014), adapt to the changes on the dataset and whose performance can also improves and gets better as their experience increases (Witten & Frank, 2005; Blum, 2007), lay a significant stress on the data for a responsive digital twin. Therefore, the digital twin can estimate how the system has to behave to tolerate the errors which happen while it is performing its tasks. In addition to tolerating the error in the production process, the design plan of the product should match up with the requirements and specifications. Thanks to the digital twin, the cost of producing a physical prototype in order to control such situation, is eliminated via the design of the digital prototype. Thus, it would be possible to make easier and more cost-effective validation and verification (V-V) than the classic method (Dahmen ve Rosmann; 2018). However, the model developed/used while predicting the errors, controlling the V-V and their consequences, has to deal with a huge number of features. The accuracy rate of the built model decreases when the number of features increases. This situation stated as "curse of dimensionality", describes the challenge in training the model as the predictor variables are added (Bellman, 1961). The main reason for this difficulty is the exponential increase of the complexity of the model concerning the number of features. One of the methods proposed as a solution to that problem is dimensionality reduction. In this study, the importance of the methods chosen for dimensionality reduction while creating a model setup, which can predict the error on a digital twin, is presented with an exemplary implementation.

Principal Component Analysis (PCA)

The main goal of Principal Component Analysis (PCA) is to perform dimensionality reduction in a multidimensional dataset. It is one of the frequently preferred methods to extract the features, which provide the most information-gain and reduce the number of dimensions (Da Costa, Alonso, & Roque, 2011; Jolliffe, 2002). Dimensionality reduction is performed by determining the features closely related with the target feature and specifying the attributes which provide the maximum information-gain about the target feature. PCA can be considered as a regression-based optimization problem (Kramer, 2011). Let there be *n* numerical variables in a dataset, V. PCA will calculate n principal components. Each of these PCs is a linear combination of original variables which includes coefficients equal to the eigenvectors of their correlation or covariance matrices. The first PC (*PC*₁) is as in the Equation 1 in the most general form (Jolliffe, 2002):

$$PC_1 = b_{11}(x_1) + b_{12}(x_2) + \dots + b_{1p}(x_p)$$
 Equation 1

where $b_{1p} p$. is the regression coefficient of the variable.

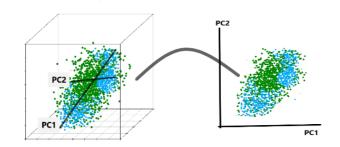


Figure 1. The design form of the dimensions in PCA method

Despite of the PCA method being frequently preferred, it can be seen that the method places the same importance on all of the observation values in some datasets and does not produce accurate results against the outliers and noise in the dataset. In return for this situation, different PCA based methods are proposed.

Conventional Principal Component Analysis

The primary goal of the Conventional PCA method is to represent the data with the maximum variance. For example, let $x_1, x_2, ..., x_n$ be (N) training sets and m represent the total mean of these training sets. In this case, the covariance matrix of the training set is defined as in the Equation 2 (Fan, Liu, & Xu, 2011).

$$C = \frac{1}{N} \sum_{i}^{N} (x_i - m) (x_i - m)^T = \frac{1}{N} X X^T$$
 Equation 2

X is defined as $X = [x_1 - m, x_2 - m, ..., x_N - m]$ in the above equation. However, the calculation of eigen decomposition of C is difficult when the dimensions of the covariance matrix, C, is oversize. As a solution to this problem, a new D matrix should be defined as $D = \frac{1}{N}X^TX$. The normalized eigenvectors of D are defined with v_i , those which belong to C are defined with u_i (Fan et al., 2011). However, u_i should be defined as a function of v_i (Equation 3). φ_i terms in the equation are the non-zero eigenvalues of both C and D.

$$u_i = \frac{1}{\sqrt{\varphi_i}} X v_i \qquad (i = 1, 2, ..., r)$$
 Equation 3

Conventional PCA extracts the features by transferring the random sample x into an r-dimensional space.

Weighted Principal Component Analysis (WPCA)

WPCA method uses the distances between each of the test and training sets to calculate the weighted covariance matrix. It performs the feature extraction with that covariance matrix. Let y be the test set and $x_1, x_2, ..., x_n$ the training sets. At this point, the proposed WPCA weighted covariance matrix is calculated using Equation 4 (Fan et al., 2011).

$$C_{w} = \frac{1}{N} \sum_{i=1}^{N} x_{i}^{\prime} x_{i}^{\prime T}$$
 Equation 4

In this calculation, C_w is the weighted variance matrix, where $x'_i = w_i x_i$. The calculation of the weight coefficients, w_i , is given in the Equation 5.

$$w_i = exp\left(-\frac{-dist(x_i, y)}{\mu}\right) dist(x_i, y)$$
 Equation 5

 $dist(x_i, y)$ is the distance between x_i and y in the equation. \max_{dist} is the maximum of the distances between the test set and the training set. μ is a positive constant. As it can be understood from the calculation of WPCA, the weight coefficient of the closest training set to the test set is larger than the others. Therefore, it has a more significant effect on the variance matrix. The existence of such a training set restricts the effect of other training sets.

Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS) was proposed by Friedman (1991), and it can perform with relative ease even in the conditions where the data is large, and the number of variables is small. It applies the divide and conquer strategy (Zhang & Goh, 2016). MARS is a method, which uses nonlinear and nonparametric regression model and provides an opportunity for flexible modeling in high-dimensional data. The most general form of a MARS model is given in Equation 6 (Samui, 2013).

$$y = c_0 + \sum_{i=1}^{N} c_i \prod_{j=1}^{K_i} b_{ji} \left(x_{v(j,i)} \right)$$

$$w_i = exp \left(-\frac{-dist(x_i, y)}{\mu} \right) dist(x_i, y)$$

Equation 6

In the equation, y is the output variable, c_0 is the constant, c_i is the coefficient of the nonconstant basis function and $b_{ji}(x_{v(j,i)})$ is the truncated power basis function. v(j,i) is the indices of the independent variable in the *i*th term of the *j*th product. K_i is a parameter which limits the order of interaction (Friedman, 1991).

The Comparison of the Methods in the Prediction of Error

In this study, a jet dyeing machine of a factory in a stage of transition to CPS design in Massachusetts is used as a base. The factory sells plastic, plexiglass glasses ,and bottles with colored embossing special for Halloween to the large organizations as a promotional material every year. Dyeing faults happen in the products produced with combinations of 12 different colors of 42 different designs. The factory aims for the newly designed digital twin to predict the error and the system to behave in a way to minimize the error. In this study, feature extraction methods were used for determining the features with the most significant contribution to the error of the digital twin which has a design based on the production data of three years. The successes of the methods were compared using a support vector machine according to the performance of predicting the error through the extracted features. There are 89016 observation values and 590 different attributes in the dataset obtained in the study. A part of the base values of a raw sample taken from the dataset is shown in Figure 2.

ViV2V3V4Min. : $1.0.0667$ Min. : 1.040 Min. : 1.0230 Min. : 2.06636 Min. : 2159Min. : 12159Min. : 1215Min. : 100Hin. : $1.0.0667$ Min. : $1.0.2020$ Min. : 2.06636 Jet Qu:: 22666Jet Qu:: 22666Jet Qu:: 2266Jet Qu:: 2266Hedian : 2.076 Median : 2.076 Median : 1.0233 Median : 3014Mean : 12456Mean : 12201Median : 1225Mean : 0.3456 Mean : 9.162 Mean : 0.1047 Mean : 1.6245 Mean : 3014Mean : 2246Max. : 22183rd Qu:: 1591Max. : 3246Max. : 2218Srd Qu:: 0.1020 Max. : 10.0047 Max. : 90.4235 Nax. : 3256Max. : 2246Max. : 2218Srd Qu:: 0.1369 Srd Qu:: 0.1020 Max. : 10.0020 Max. : 90.4235 Nax. : 3266Max. : 2218Min. : 0.0000 Max. : 1.0000 Max. : 10.0000 Max. : 10.0000 Max. : 10.0000 Max. : 10.0000 Jet Qu:: 1.0177Jet Qu:: 100Min. : 10.0010 Jet Qu:: 10.9224 Mean : 10.0224 Median : 0.0216 Max. : 0.0024 Mean : 1.0168Median: 101.51Median: 0.1218 Mean : 10.0224 Mean : 10.0216 Max. : 0.0277 Srd Qu:: 0.0277 Max. : :14Na's : :14Na's : :14Na's : :0Na's : :0Na's : :0Na's : :0Srd Qu:: 0.03776 Na's : :14Na's : :14Na's : :0Na's : :0Na's : :0Na's : :0Na's : :0Srd Qu:: :0.0377Srd Qu:: 0.03807 Median : :1.111Leg U: :0.003400Min. : :0	
Ist Qu.:2866 Ist Qu.:2181 Ist Qu.:1082 Median :0.2334 Median :0.2364 Median :0.0355 Median :0.0101 Median :0.0101 Median :0.0101 Median :0.01011 Mean : 0.0101 Mean : 0.020 Max : 10.0101 Mean : 0.04355 NA* 10 Na* 114 NA* NA* 117000 Max : 10.0101 Max : 10.0000 Max : 10.0000 Max : 10.0001 Max : 10.0101 Max : 10.0001 Max : 10.0001 <td< td=""></td<>	
Hedian: 13011 Median: 12010 Hedian: 1201 Median: 1201 Median: 1201 Median: 1201 Mean: 1205 Mean: 10.1345	
Mean :3014 Mean :2201 Mean :1396 Mean :2201 Mean :1396 Mark :33rd Qu.:0.1228 3rd Qu.:0.1221 Max :0.032 Max. :2315 Max. :3715 Max. :221967 Max. :10.0020 Max. :0.0320 Max. :0.04250 Min. :0.6815 Min. :100 Min. :0.0000 Ist Qu.:0.1211 Min. :0.0010 Ist Qu.:0.0108 Median :0.0006 Ist Qu.:0.0108 Median :0.0006 Max. :0.00047 Mean :1017 ist Qu.:10.0 3rd Qu.:0.1218 Mean :0.0216 Mean :0.0216 Mean :0.0206 Max. :0.0206 Max. :0.0206 Max. :0.0206 Max. :0.0206 Max. :0.0206 Max.	
3rd Qu.:2057 3rd Qu.:218 3rd Qu.:1012 3rd Qu.:218 3rd Qu.:1012	
Max. :3356 Max. :2315 Max. :3715 Max. :17150 Ma	
V5 V6 V7 V8 V5/5 Math 10.0002 Math 10.0032 Math 10.0032 Math 10.0037 Math 10.0037 Math 10.0037 Math 10.0037 Math Math 10.0057 Math Math 10.0057 Math Math 10.0057 Math 10.00570 Math 10.00286	
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	
Min. : 0.6815 Min. : 100 Min. : 0.0000 lat Qu.: 11.502 lat Qu.: 0.0138 lat Qu.: 0.0034 lat Qu.: 1.0177 lat Qu.: 10.017 lat Qu.: 10.121 Median : 0.0204 Median : 0.0168 Median : 0.0034 Mean: 1.3168 Median : 100 Median : 10.151 Median : 0.1224 Mean : 10.642 Mean : 0.0204 Median : 0.0168 Median : 0.0054 Max. : 114.5356 Max. : 100 Max. : 0.1218 Mean : 10.0161 Mean : 0.1218 Mean : 0.0217 Max : 0.0200 Max. : 0.0206 NA*s : 114	
1st Qu:: 1007 1st Qu:: 100 1st Qu:: 100 <t< td=""></t<>	
Hedlan: 1.3168 Hedlan: 100 Hedlan: 100.151 Hedlan: 10.124 Mean 110.642 Mean 100.168 Mean 100.054 Mean: 4.1970 Mean: 100.111 Mean: 10.1218 3rd Qu:17.081 3rd Qu:0.0226 Max. 10.0258 3rd Qu:10.0277 3rd Qu:0.0.0203 3rd Qu:0.0.0055 Max. 111.5356 Max. 110.528 Max. 10.128 3rd Qu:17.081 3rd Qu:0.0.0277 3rd Qu:0.0.0237 Max Qu:0.0.0235 Max. 10.028 Max. 10.0286 Max. 10.00170 Max. 10.000100	
Jackan I. 100 Jackan I. 100 <th jackn<="" td="" th<=""></th>	
Max. Max. <thmax.< th=""> Max. Max. <thm< td=""></thm<></thmax.<>	
NA's NA's NA's NA's NA's NA's 164 164	
Nin V10 V11 V12 V52 V53 V54 V555 Nin. :-0.053400 Nin. :-0.0349000 Nin. :-0.05554 Nin. :0.0554 Nin. :0.0576 Nin. :0.000100 Nin. :0.000100 lst du::14.11 lst du::-0.01800 lst du::10.9581 lst du::10.01670 Nin. :0.000100 Nin. :0.000100 Median::-0.001300 Median::-0.000300 Median::0.9581 lst du::14.18 lst du::0.01160 lst du::0.001600 Median::-0.001300 Median::-0.000400 Median::0.9581 Median::72.29 Median::0.0502 Median::0.01160 lst du::0.003600 Mean:::-0.001458 Mean:: :0.001458 Mean:: :0.5001 Mean:: :0.003847 3rd du::1.517 3rd du::0.005400 3rd du::0.9848 Max. :0.303 Mean:: :0.47660 Max. :0.104500 Max :1 NA's :2 NA's :2 NA's :1 NA's :1 NA's :1 NA's :1	
Min. 1.191 Min. 1-0.034000 Min. 1-0.034000 Min. 1-0.034000 Min. 10.001700 List Qu:1.141 List Qu:-0.010800 List Qu:-0.0058000 List Qu:-0.0058000 List Qu:-0.005800 Median: List Qu:-0.005810 Median: List Qu:-0.005810 Median: List Qu:-0.005800 Median: List Qu:-0.005810 Max: List Qu:-0.0058100 Max:	
lst Qu::1.411 lst Qu::-0.010800 lst Qu::-0.0056000 lst Qu::0.9581 lst Qu::46.18 lst Qu::0.4979 lst Qu::0.01160 lst Qu::0.003100 Median :1.462 Median :-0.001300 Median : 0.0004000 Median :0.5658 Median : 77.29 Median :0.5001 Median :0.003600 Mean :1.462 Mean :-0.000840 Mean : 0.004150 Mean : 0.69644 Mean : 9.793 Mean : 0.5001 Mean : 0.003807 3rd Qu::1.517 3rd Qu:: 0.0054000 3rd Qu::0.0058000 3rd Qu::0.9848 Max. :10.564 Max. :0.5024 3rd Qu::0.01650 3rd Qu::0.004100 Max. :1.656 Max. : 0.0050000 Max. :0.9848 Max. :737.30 Max. :0.5096 Max. : 0.47660 Max. : 0.104500 Ma* :2 NA*s :2 NA*s :2 NA*s :2 NA*s :1 NA*s :1 NA*s :1	
Median: 1.462 Median: 1.0.003000 Median: 1.0.003600 Median: 1.72.25 Median: 1.0.502 Median: 1.0.003600 Mean: 1.463 Mean -0.000841 Mean: 1.0.001458 Mean: 1.0.001450 3rd 0.0.01502 Mean: 1.0.003600 Mean: 1.0.001450 3rd 0.0.01502 Mean: 1.0.003600 Mean: 1.0.001450 3rd 0.0.01502 Mean: 1.0.003600 Max. 1.0.053000 Max. 1.0.05000 Max. 1.0.05000 Max. 1.0.05000 Max. 1.0.05000 Max. 1.0.04500 Max. 1.0.04500 Max. 1.0.04500 Max 1.0.04500 Max 1.0.04500 Max 1.0.04500 Max 1.0.04500 Max	
Mean :1.463 Mean :-0.000841 Mean :0.0001458 Mean :0.9644 Mean :97.93 Mean :0.0011532 Mean :0.003847 3rd Qu.:1.517 3rd Qu.:0.009400 3rd Qu.:0.0059000 3rd Qu.:0.01530 3rd Qu.:0.010650 3rd Qu.:0.00100 Max. :1.65 Max. :0.079400 Max :0.0530000 Max. :0.9848 NA's :2 NA's :2 NA's :0.9848 Max. :0.5098 Max. :0.47660 Max. :0.04500 NA's :2 NA's :2 NA's :1 NA's :1	
3rd Qu.:1.517 3rd Qu.: 0.008400 3rd Qu.: 0.0059000 3rd Qu.: 0.9400 3rd Qu.: 0.005400 3rd Qu.: 0.005400 3rd Qu.: 0.005400 Max. :1.656 Max. : 0.074900 Max. : 0.0530000 Max : 0.9848 Max. : 10.5059 Max. : 0.47660 Max. : 0.104500 Ma* :2 NA*s :2 NA*s : 1 NA*s : 1 NA*s : 1	
Max. :1.656 Max. :0.074900 Max. :0.0530000 Max. :0.9848 Max. :0.35098 Max. :0.47660 Max. :0.104500 NA's :2 NA's :2 NA's :949 NA's :1 NA's :1	
NA's :2 NA's :2 NA's :2 NA's :2 NA's :440 NA's :1 NA's :1 NA's :1 NA's :1	
NA 5 1915 NA 5 11 NA 5 11 NA 5 11	
V13 V14 V15 V16 V17 V586 V587 V588 V589 Min.:182.1 Min.: 1.2.249 Min.:333.4 Min.: 1.4.70 Min	
Min. : 1.196 Min. :-0.01690 Min. :0.00320 Min. :0.001000	
Ist Qu.: 2.307 Ist Qu.: 0.01342 Ist Qu.: 0.01342 Ist Qu.: 0.01342	
Median 11975 Median 10 Median 10.507 Median 11222 Median 1 9.552 Median 2.758 Median 1 0.02050 Median 10.01480 Median 10.004600	
3rd On :202 0 3rd On :0 3rd On :10 862 3rd On :419 1 3rd On : 10 128 Mean : 3.068 Mean : 0.02146 Mean :0.01647 Mean :0.005283	
Max. :272.0 Max. :10 Max. :19.547 Max. :2249 Max. :122.868 3rd Qu.: 3.295 3rd Qu.: 0.02760 3rd Qu.: 0.02030 3rd Qu.: 0.006400	
NA's :2 NA's :3 NA's :3 NA's :3 NA's :3 NA's :3 NA's :3	
NA'S 11 NA'S 11 NA'S 11 NA'S 11 NA'S 11	
V590	
Min. : 0.00	
lst Qu.: 44.37	
Median : 71.90	
Mean : 99.67	
3rd Qu :: 114.75	
Max. :737.30	
NA's :1	

Figure 2. The base values of a raw sample taken from the dataset

In the dataset, the values measured during the production process, from the faultlessness of the printing of the product coming from the printing machine to the spraying speed of the color and dye according to the pattern in the productions before every Halloween, are given. In Figure 3, the distribution of the observation values in the sample taken from the dataset is given.

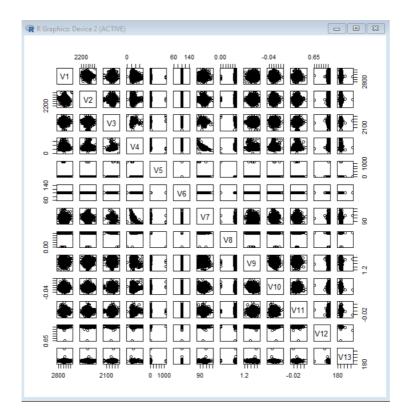


Figure 3. The distribution of the observation values in the sample taken from the dataset First of all, the values in the dataset were preprocessed using R ,and with the help of mice, VIM, Boruta packages ,and the missing values were completed. Non-linear SVM with polynomial kernel was used for predicting the error with the help of feature spaces extracted by applying standard PCA, conventional PCA, WPCA and MARS to the dataset. The existence or absence of error was tried to be predicted by splitting the data in feature space with the hold-out method (75%-25%). The accuracy values of error prediction of the SVM models designed in each feature space are given in the Table 1.

Feature Space (FS)	SVM Model Type	SVM Model	ACC
FS _{PCA}	Non-linear Polynomial kernel	<i>M</i> ₁	87,73%
FS _{CPCA}	Non-linear Polynomial kernel	<i>M</i> ₂	89,46%
FS _{WPCA}	Non-linear Polynomial kernel	<i>M</i> ₃	91,34%
FS _{MARS}	Non-linear Polynomial kernel	<i>M</i> ₄	91,18%

Table 1. The accuracy values of error prediction of the SVM models designed in eachfeature space

In Table 1, it can be seen that the success rates vary by the SVM models designed in the new feature spaces created with the feature extraction methods. It can be seen that the model (M_1) designed with standard PCA is not sufficient although it produces a nearly successful result. In this sense, it can be observed that the models (M_3, M_4) , designed with the help of features spaces

created with WPCA and MARS methods, can predict the error more successfully than others. In addition to the accuracy of the model, F-measure was also calculated as the integrated performance evaluation criterion. In these calculations, it was seen that M_3 (F=94,68%) has larger values than M_4 (F= 92,56%).

Results

The importance of the methods used in predicting the variables that cause the error, regarding the capability of digital twin to optimize the system's behavior to correct the error, for digital twin within CPS design to be able to predict the error is explained through this study. In applied practice, it has been understood that the standard PCA method fails to acquire the desired achievement. WPCA, which is the further developed version of this method, displays more accurate estimations of the error. However, it would be incorrect to present this method as the only method that should be used in twin design. PCA is affected by the variance condition of the dataset. Therefore, MARS method should be chosen in the design when the variance condition is not met.

Some constraints of the study are the prediction of erroneous conditions: color-visual discrepancy and dye bleeding in the embossment. Another constraint of the study is that the hold-out method is used in the design of SVM model. Instead of this method, more precise model design and accuracy prediction are possible with the k-fold cross validation methods.

As alternatives to standard PCA method, there are methods like kernel based PCA (Burges, 2010), sparse-data kernel based PCA (Li, Gao, 2011), singular value decomposition (SVD), SVD based PCA in literature. The feature extraction solutions that the methods provide purposefully differ by methods and they affect the performance of the designed models. For this reason, using only the production-oriented working principle design as a base in the design of a digital twin, and focusing only on the data collection and model design strategy in the design for prediction, shall cause making an imperfect design. In order that the action-reaction process of the system in production can work adequately, feature extraction should be done on the collected data and used on the actual factors which cause the trouble for the solution.

Conclusion

The ultimate goal of digital twin for production process is create data-driven solutions incorporating advanced analytics. Thus that production process can be transformed from "react and repair" to "predict and prevent". Besides digital twins with predictive power will provide significant reduction to unplanned downtime and costs and also significant benefits and advantages during well construction and production, But the whole advantages of a digital

twin depends on its data strategy from collecting to the modelling. According to SIEMENS (2018), with insufficient data analysis strategy and adequate modeling, all the costs from predicting to modeling, from validation to verification will become unmanageable, when it's compared to the classic production process. So there should be such a strategy that necessary to properly verify that the model properly predicts the source of error and validate that the model adequately represents the reality.

References

Alpaydın, E. (2014). Introduction to Machine Learning. MIT Press.

- Bellman, R. (1961). Adaptive control processes: a guided tour. Princeton University Press.
- Belzner, L., Hennicker, R., & Wirsing, M. (2015). Onplan: A framework for simulation-based online planning. *Formal Aspects of Component Software*, 1-30.
- Blum, A. (2007). *Machine learning theory*. A., 2007, Machine learning theory, taken from http://www.cs.cmu.edu/afs/cs/user/avrim/www/Talks/mlt.pdf
- Burges, C. J. (2010). Dimension reduction: A guided tour. *Foundations and Trends in Machine Learning*, 275-365.
- Da Costa, J. F., Alonso, H., & Roque, L. (2011). A Weighted Principal Component Analysis and Its Application to Gene Expression Data. *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, 7, 245-252.
- Dahmen U., Roßmann J. (2018) Simulation-based Verification with Experimentable Digital Twins in Virtual Testbeds. In: Schüppstuhl T., Tracht K., Franke J. (eds) *Tagungsband des 3. Kongresses Montage Handhabung Industrieroboter*. Springer Vieweg, Berlin, Heidelberg
- Fan, Z., Liu, E., & Xu, B. (2011). Weighted Principal Component Analysis. Artificial Intelligence and Computational Intelligence, 569–574. https://doi.org/10.1007/978-3-642-23896-3_70
- Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics, 19*, 1-141.
- Gabor, T., Belzner, L., Kiermeier, M., Beck, M. T., & Neitz, A. (2016). A simulation-based architecture for smart cyber-physical systems. *Proceedings - 2016 IEEE International Conference on Autonomic Computing, ICAC 2016*, 374–379. https://doi.org/10.1109/ICAC.2016.29
- Glassen, E. H., & Stargel, D. (2012). The digital twin paradigm for future NASA and US air force vehicles. 53rd Structures, Structural Dynamics, and Materials Conference: Special Session on the Digital Twin, (s. 1-14). Honolulu, US.

Harrington, P. (2012). Machine Learning in Action. NY: Manning Publications.

Jolliffe, L. T. (2002). Principal Component Analysis. Springer.

- Kodratoff, Y., & Michalski, R. S. (2014). Research in machine learning: Recent progress, classification of methods, and future directions. Y. Kodratoff, & R. S. Michalski, *Machine learning: an artificial intelligence approach* (s. 3-30). Morgan Kaufmann.
- Kramer, O. (2011). Dimensionality reduction by unsupervised k-nearest neighbor regression. Machine Learning and Applications and Workshops, 10th International Conference on. IEEE.
- Mitchell, T. M. (1997). Machine Learning. McGraw-Hill Science/Engineering/Math.
- Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. The MIT Press.
- Samui, P. (2013). Multivariate adaptive regression spline (Mars) for prediction of elastic modulus of jointed rock mass. *Geotech Geol Eng*, 31, 249-253. doi:0.1007/s10706-012-9584-4
- Tuegel, E. J., Ingraffea, A., Eason, T., & Spottswood, S. M. (2011). Reengineering aircraft structural life prediction using a digital twin. *International Journal of Aerospace Engineering*, 1-14. doi:10.1155/2011/154798
- Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.
- Zhang, W., & Goh, A. T. (2016). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geoscience Frontiers*, 7, 45-52.