LS-SVM based regression and spectral clustering for predicting maintenance of industrial machines

Rocco Langone[1,5], Carlos Alzate[1,4], Bart De Ketelaere[2], Jonas Vlasselaer,[3] Wannes Meert,[3] Johan A. K. Suykens[1,5]

[1]*Department of Electrical Engineering (ESAT), SCD, KU Leuven, B-3001 Leuven Belgium*
Email: {rocco.langone,carlos.alzate,johan.suykens}@esat.kuleuven.be

[2]*Faculty of Bioscience Engineering, BIOSYST MeBioS Qualimatrics, KU Leuven, B-3001 Leuven Belgium*
Email: bart.deketelaere@biw.kuleuven.be

[3]*Department of Computer Science, KU Leuven, Celestijnenlaan 200a B-3001 Leuven Belgium*
Email: {jonas.vlasselaer,wannes.meert}@cs.kuleuven.be

[4]*Smarter Cities Technology Center, IBM Research-Ireland*
Email: carlos.alzate@ie.ibm.com

[5] *iMinds Future Health Department*

Abstract

Accurate prediction of forthcoming faults in modern industrial machines plays a key role in reducing production arrest, increasing the safety of plant operations, and optimizing manufacturing costs. The most effective condition monitoring techniques are based on the analysis of historical process data. In this paper we show how Least Squares Support Vector Machines (LS-SVMs) can be used effectively for early fault detection in an online fashion, a field where other techniques are more popular. In particular, in the first part we present an unsupervised approach that uses Kernel Spectral Clustering (KSC) on the sensor data coming from a packing machine, in order to distinguish between normal operating condition and abnormal situations. Basically, we describe how KSC is able to detect in advance the need of maintenance actions in the analyzed machine. In the second part we illustrate a nonlinear auto-regressive model (NAR), thus a supervised learning technique, in the LS-SVM framework. We show that we succeed in modeling appropriately the degradation process affecting the machine, and are capable to accurately predict the evolution of dirt accumulation in the sealing jaws.

Keywords: kernel spectral clustering; LS-SVMs; NAR; time-series; condition

monitoring; maintenance; fault detection.

1. Introduction

In industrial processes, the detection and analysis of faults ensure product quality and operational safety. Traditionally, four ways to deal with sensory faults have been used [18],[19],[20]: corrective maintenance, preventive maintenance, manual predictive maintenance, and condition-based maintenance. Corrective maintenance is performed only when the machine fails, it is expensive and safety and environmental issues arise. Preventive maintenance is based on periodic replacement of components, which are then utilized in a non-optimal way. In predictive maintenance, machines are manually checked with expensive monitoring techniques. In this case the components are replaced according to their real status, but the operations are labor intensive and prone to human errors. Condition-based maintenance is increasingly receiving attention due to its many advantages. A machine's status is automatically inspected and maintenance is planned accordingly. The continuous monitoring of machine parts leads to reliable and accurate lifetime predictions, and maintenance operations can be fully automated and implemented in a cost effective way. Nowadays, many process variables can be measured. In a power plant, for example, these are temperature, pressure, flow rate etc. These measurements give an information on the current status of a machine and can be used to predict the faults and plan an optimal maintenance strategy. When a component starts degrading, the related sensor reading shows a deviation from its normal behavior and this can indicate an incoming failure of the component. So far, process models based on the sensor data have been constructed by using exponentially weighted moving average, cumulative sum, principal component analysis (PCA), etc. [7], [5]. Moreover, the problem of discovering the incoming faults can be seen as a special case of outlier detection, since an outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism. In this field, supervised, semi-supervised and unsupervised methods are employed [8].

In the first part of this work we describe the use of kernel spectral clustering (KSC, [2]) for identifying in advance when the analyzed machine enters critical conditions. A similar analysis has been carried out in [9], this paper represents an extended version.

KSC is a spectral clustering formulation with a primal-dual setting. The KSC method has two main advantages, a systematic model selection scheme for the

correct tuning of the parameters and the extension of the clustering model to out-of-sample points. These aspects are rare in standard clustering techniques, where heuristics and trial-error approaches are more common. In KSC, a clustering model can be trained on a subset of the data and then applied to the rest of the data in a learning framework. The out-of-sample extension allows then to predict the memberships of a new point thanks to the model learned during the training phase. In this way, once a model of the operation of a machine has been constructed, we can use it in an online fashion to discover when the plant enters critical conditions.

In the second part we utilize a nonlinear auto-regressive model (NAR [10]) to catch the deterioration phenomenon acting on the machine. Like KSC, the model is cast in the LS-SVM framework [16, 15] and, once properly trained, it can be used in an online manner thanks to the out-of-sample extension ability. The main advantages of constructing the NAR model in the LS-SVM framework consist of an high quality generalization capability [17], the flexibility in the model design and a clear procedure for model selection.

The remainder of this paper is structured as follows: Section 2 summarizes the KSC model and the related model selection scheme. Section 3 explains the basics of the LS-SVM regression method and how it can be used to formulate a NAR model. Section 4 describes the data-sets used in the experiments. In Section 5 the simulation results are presented. Regarding KSC, a comparison with k-means, which is the most popular method used in industrial applications [4] is also depicted. For what concerns the LS-SVM NAR model, we show that it outperforms a linear auto-regressive model (AR) and gives very good predictions up to 10 steps ahead. Section 6 presents a general discussion. Finally, Section 7 concludes the paper.

2. Kernel Spectral Clustering

2.1. Model

Spectral clustering methods use of the eigenvectors of the Laplacian to properly group the data-points [6, 21, 13]. Given training data $\mathcal{D} = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d$ and the number of clusters k, the kernel spectral clustering formulation is [2]:

$$\min_{w^{(l)},e^{(l)},b_l} \frac{1}{2}\sum_{l=1}^{k-1} w^{(l)^T} w^{(l)} - \frac{1}{2N}\sum_{l=1}^{k-1} \gamma_l e^{(l)^T} D^{-1} e^{(l)} \tag{1}$$

$$\text{such that } e^{(l)} = \Phi w^{(l)} + b_l 1_N \tag{2}$$

where $e^{(l)} = [e_1^{(l)}, \ldots, e_N^{(l)}]^T$ are the projections, $l = 1, \ldots, k-1$ indicate the score variables needed to encode the k clusters to find, $D^{-1} \in \mathbb{R}^{N \times N}$ is the inverse of the degree matrix D, Φ is the $N \times d_h$ feature matrix $\Phi = [\varphi(x_1)^T; \ldots; \varphi(x_N)^T]$ and $\gamma_l \in \mathbb{R}^+$ are regularization constants. The clustering model is formulated as:

$$e_i^{(l)} = w^{(l)^T} \varphi(x_i) + b_l, i = 1, \ldots, N \tag{3}$$

where $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^{d_h}$ is the mapping to a high-dimensional feature space, b_l are bias terms, with $l = 1, \ldots, k-1$. The projections $e_i^{(l)}$ represent the latent variables of a set of $k-1$ binary clustering indicators given by sign$(e_i^{(l)})$. The set of binary indicators form a code-book $\mathcal{CB} = \{c_p\}_{p=1}^k$, where each code-word is a binary word of length $k-1$ representing a cluster. After constructing the Lagrangian and solving the Karush-Kuhn-Tucker (KKT) conditions for optimality the following dual problem is obtained:

$$D^{-1} M_D \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)} \tag{4}$$

where Ω is the kernel matrix with ij-th entry $\Omega_{ij} = K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$, D is the graph degree matrix which is diagonal with positive elements $D_{ii} = \sum_j \Omega_{ij}$, M_D is a centering matrix defined as $M_D = I_N - \frac{1}{1_N^T D^{-1} 1_N} 1_N 1_N^T D^{-1}$, the $\alpha^{(l)}$ are dual variables, $\lambda_l = \frac{N}{\gamma_l}$. $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is the kernel function and captures the similarity between the data-points. In all the experiments we use the RBF kernel function defined by $K(x_i, x_j) = \exp(-||x_i - x_j||_2^2 / \sigma^2)$, with σ meaning the bandwidth parameter. The out-of-sample extension to new nodes is done by an Error Correcting Output Codes (ECOC) decoding procedure. In the decoding process the cluster indicators obtained in the validation/test stage are compared with the code-book and the nearest code-word in terms of Hamming distance is selected. The cluster indicators can be obtained by binarizing the score variables for test points in the following way:

$$\text{sign}(e_{\text{test}}^{(l)}) = \text{sign}(\Omega_{\text{test}} \alpha^{(l)} + b_l 1_{\text{Ntest}}) \tag{5}$$

with $l = 1, \ldots, k-1$. Ω_{test} is the $N_{\text{test}} \times N$ kernel matrix evaluated using the test nodes with entries $\Omega_{\text{test,ri}} = K(x_r^{\text{test}}, x_i)$, $r = 1, \ldots, N_{\text{test}}$, $i = 1, \ldots, N$.

2.2. Tuning scheme

The number of clusters in which to group the data and the parameters of the kernel function (if any) have to be selected properly to ensure good performances.

The model selection scheme used in the experiments is Balanced Line Fit (BLF). The BLF criterion reaches its maximum value 1 when the clusters are well separated, and in this ideal situation they are represented as lines in the projection space. The analytical formula of BLF is [2]:

$$\text{BLF}(\mathcal{D}^V, k) = \eta \text{linefit}(\mathcal{D}^V, k) + (1 - \eta)\text{balance}(\mathcal{D}^V, k) \tag{6}$$

where $\mathcal{D}^V$ indicates the validation set and k the number of clusters. The linefit index equals 0 when the score variables are distributed spherically and equals 1 when the score variables related to points in the same cluster are collinear. The balance index equals 1 when the clusters have the same size and tends to 0 in extremely unbalanced cases. The parameter η controls the importance given to the linefit with respect to the balance index and takes a value in the range $[0, 1]$. The BLF can be used to select the number of clusters and the kernel tuning parameters in the following way:

1. Define a grid of values for the parameters to select
2. Train the related KSC model
3. Compute the memberships of the validation points by means of the out-of-sample extension
4. For every partition of the validation set calculate the related score in terms of BLF
5. Choose the model with the highest score.

An example of the model selection procedure on a toy data-set is depicted in Fig. 1. It can be noticed that the maximum value of the BLF criterion is reached when the points in the projection space have a strong line structure, corresponding to optimal parameters.

3. LS-SVMs for regression

3.1. Primal-Dual Formulation

Given training data $\{x_i, y_i\}_{i=1}^N$, with $x_i \in \mathbb{R}^{d_x}$ and $y_i \in \mathbb{R}$, the regression problem in the primal space can be expressed as follows [16]:

$$\min_{w,e,b} \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2 \tag{7}$$

$$\text{such that } y_i = w^T \varphi(x_i) + b + e_i, i = 1, \ldots, N. \tag{8}$$

The expression $\hat{y} = w^T\varphi(x) + b$ indicates the model in the primal space, and the objective function (7) is in fact a ridge regression cost function in the case of zero bias term. With γ we indicate the regularization parameter which controls the trade-off between the model complexity and the minimization of the training error. If we construct the Lagrangian we have:

$$\mathcal{L}(w, e, b, \alpha) = \frac{1}{2}w^T w + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N}\alpha_i(w^T\varphi(x_i) + b + e_i - y_i)$$

where α_i are the Lagrange multipliers. The KKT optimality conditions are:

$\frac{\partial\mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N}\alpha_i\varphi(x_i)$,
$\frac{\partial\mathcal{L}}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma e_i$,
$\frac{\partial\mathcal{L}}{\partial b} = 0 \rightarrow \sum_{i=1}^{N}\alpha_i = 0$,
$\frac{\partial\mathcal{L}}{\partial \alpha_i} = 0 \rightarrow w^T\varphi(x_i) + b + e_i - y_i = 0$.

Eliminating the primal variables $e = [e_1; \ldots; e_N]$ and w leads to the following linear system in the dual problem:

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N^T & \Omega + I/\gamma \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right] \tag{9}$$

where $y = [y_1; \ldots; y_N]$, $1_N = [1; \ldots; 1]$, $\alpha = [\alpha_1; \ldots; \alpha_N]$. The term Ω means the kernel matrix with entries $\Omega_{ij} = \varphi(x_i)^T\varphi(x_j) = K(x_i, x_j)$. With $K : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ we denote the kernel function which maps the input points into an high dimensional feature space. By using a radial basis function (RBF) kernel, expressed by $K(x_i, x_j) = \exp(-||x_i - x_j||_2^2)/\sigma^2$, one is able to construct a model of arbitrary complexity. Finally, after solving the previous linear system, the LS-SVM model for function estimation becomes:

$$\hat{y} = \sum_{i=1}^{N}\alpha_i K(x, x_i) + b. \tag{10}$$

3.2. NAR model

A nonlinear auto-regressive model (NAR) describes time-varying phenomena specifying that the output variable depends non-linearly on its own previous values [10]:

$$\hat{y}_{k+1} = f(y_k, \ldots, y_{k-p}) \tag{11}$$

where $\hat{y}_{k+1}$ is the predicted future value based on the previous p values, $f : \mathbb{R}^{d_y \cdot p} \rightarrow \mathbb{R}$ is the nonlinear mapping, with d_y indicating the dimension of the output variable y. The parameter p is also called order of the NAR model and has to be properly tuned. In the LS-SVM framework the NAR model can be described by:

$$\hat{y}_{k+1} = w^T \varphi(y_k, \ldots, y_{k-p}) + b \tag{12}$$

For multi-step ahead prediction a recursive approach can be used:

$\hat{y}_{k+1} = f(y_k, \ldots, y_{k-p})$
$\hat{y}_{k+2} = f(\hat{y}_{k+1}, \ldots, y_{k-p+1})$
...
$\hat{y}_{k+m} = f(\hat{y}_{k+m-1}, \ldots, y_{k-p+m})$

where m is the number of steps ahead.

3.3. Model selection

In order to build-up a valid LS-SVM NAR model for time-series prediction, we have to choose very carefully the parameters (γ, σ, p). The parameters γ and σ are selected using 10-fold cross-validation. The Coupled Simulated Annealing (CSA) is used to minimize the mean square error (MSE) in the cross-validation process. Also the simplex algorithm [12] is used afterward to further improve the search of the optimum. CSA leads to an improved optimization efficiency because it reduces the sensitivity of the algorithm to the initialization parameters while guiding the optimization process to quasi-optimal runs [22]. The optimal order p of the NAR model is found using a grid search approach.

4. Data-sets

The data are collected from a Vertical Form Fill and Seal (VFFS) machine used to fill and seal packages in different industries, but mainly in the food industry. The VFFS machine supplies film from a roll which is formed into a bag over the vertical cylinder. Sealing jaws close the bag at the bottom before it is filled. At the end of the cycle, the bag is sealed and cut off with a knife. Since the dirt accumulation on sealing jaws seems to influence strongly the process quality, in the experiments that have been conducted the jaws were monitored to predict in advance the maintenance actions. Maintenance consists of stopping the machine

and cleaning the sealing jaws. A total of three experiments have been performed. In the first two experiments only accelerometers mounted on the jaws are used in order to indirectly detect the dirt accumulation. In the third experiment the machine is also equipped with a thermal camera, which directly measures the dirt accumulation in terms of number of hot area pixels in the acquired images. This has resulted in three data-sets:

- DS_I: this data-set consists of 771 events and 3 external maintenance actions. An event is related to a particular processed bag and takes place every two seconds. Each event is associated with a 150-dimensional accelerometer signal (see top of Fig.2).

- DS_II: it contains a total of $11\,632$ processed bags and 15 maintenance actions. Here the vibration signals used to monitor the dirt accumulation in the jaws are 190-dimensional time-series (as shown in the bottom part of Fig.2).

- DS_III: there are $3\,519$ processed bags and 11 maintenance actions. For this data-set we are given, beyond the the accelerometer signals, also the thermal camera measurements, as depicted in Fig. 3.

5. Experimental results

Here we present the analysis of the data described in the previous Section. In the first part (unsupervised learning) we illustrate the performance of KSC and we compare it with k-means [11], the most popular data clustering method for industrial applications. The raw accelerometer signals are used to construct the clustering models. In the second part (supervised learning) we explain the time-series prediction on the thermal camera data by means of an LS-SVM NAR model. The comparison with a linear model and the standard baseline is also discussed.

5.1. Unsupervised learning

In this section it is shown how KSC can be used to perform just-in-time maintenance, not too early to take full advantage of component lifetime but also not too late to avoid catastrophic failures and unplanned downtime. In particular, 2 regimes were identified, where one of them can be interpreted as normal behavior and the other as critical conditions (need of maintenance). Moreover, a probabilistic interpretation of the results is also provided, which better describes the degradation process experienced by the sealing jaws of the packing machine. We

perform clustering on the raw accelerometer signals. Finally, we compare KSC with k-means.

5.1.1. Model selection

In order to catch the ongoing deterioration process of the jaws we need to use historical values of sealing quality in our analysis. For this purpose we apply a windowing operation on the data (Fig. 4). Therefore we have a total of 3 parameters to determine: the window size (i.e. the number of signals to concatenate), the number of clusters k and the RBF kernel parameter σ. According to the BLF criterion the optimal window size is 40 and the optimal number of clusters is $k = 2$ for the three data-sets, while σ is data-set dependent.

5.1.2. Hard clustering

In Fig. 5 the KSC prediction for the data-set DS_I is shown. We can interpret one of the clusters as normal behavior and the other as maintenance cluster. Notice that the KSC model is able to predict some minutes in advance the maintenance actions before they are actually performed by the operator. Concerning the data-set DS_II we can draw the same comments: KSC is very accurate in predicting the worsening of the packing process around the actual maintenance events (see top of Fig. 6). Finally Fig. 7 illustrates the results on data-set DS_III. In this case KSC predicts the need of maintenance also in zones where maintenance has not been really performed. As we will see later, surprisingly it would have been more logical to perform maintenance as suggested by KSC and not as actually done by the operator.

5.1.3. Soft clustering

In the previous section we demonstrated the effectiveness of KSC in predicting in advance the maintenance events. Nevertheless the predicted output is binary (it goes suddenly from normal operation to maintenance). An output of this form does not provide a continuous indicator of the incoming maintenance actions. To solve this issue we can use the latent variable $e(x)$ instead of the binarized clustering output $\text{sign}(e(x))$ (see section 2.1). The latent variable provides a more informative output which can be analyzed in order to produce a better prognostic output. Since the range from which it takes a value depends on many factors (e.g., the kernel and its parameters, the number of training data points), the interpretability might be difficult. To improve it, values are normalized between 0 and 1. This transformation is based on the structure of the latent variable space. As already mentioned in 2.2, in this space every cluster is ideally represented as

a line. The tips of the lines can be considered prototypes of their cluster, since they have more certainty to belong to it because they are further from the decision boundaries [1]. Thus, the distance from every point to the cluster prototype can be seen as a confidence measure of the cluster membership. The transformed latent variable is depicted at the bottom of Fig. 5 (data-set DS_I), Fig. 6 (data-set DS_II) and Fig. 7 (data-set DS_III). The value can be considered as a soft membership or "probability" to maintenance [3]. The latter increases as the number of faulty bags in the window increases. The value can decrease since the window can move onto zones with good seals after a period of bad seals. Maintenance is predicted when the probability reaches the value 1.

For data-sets DS_I and DS_II it can be noticed how KSC is able to discover from the vibration signals registered by the accelerometers the dirt accumulation in the jaws that leads to the maintenance actions. This is very surprising because clustering is an unsupervised technique, and thus does not make use of any information on the location of the maintenance actions (like it occurs for classification). For what concerns data-set DS_III, also regions where no maintenance actions have been really performed appear associated with high probability to maintenance. To understand these unexpected outcome, we can compare bottom of Fig. 7 (discovered degradation) with bottom of Fig. 3 (measured degradation by means of the number of hot area pixels in the images captured by the thermal camera). We can recognize similar patterns, meaning that KSC was able to catch the degradation process also in this last experiment. Thus, probably the operator should have performed the maintenance operations in a different way, being coherent with his behavior in the first two experiments.

5.1.4. k-means clustering

k-means clustering is a standard method for finding clusters in a set of data points. After choosing the desired number of cluster centers, the k-means procedure iteratively moves the centers to minimize the total within cluster variance. k-means has several drawbacks:

- the results are strongly influenced by the initialization
- the number of clusters should be provided by the user (this is done heuristically using a trial and error approach)
- it can discover only spherical boundaries.

Despite these disadvantages, it is still widely used since it works effectively in many scientific and industrial applications. Since it is the most popular method

used in the industrial sector, here we present the results of k-means applied on the three data-sets described in Section 4. Before discussing the results, it is worth to mention that, thanks to the model selection scheme of KSC described in Section 2.2, we give optimal parameters to k-means (number of clusters = 2 and window size of concatenated accelerometer signals = 40). Fig. 8 visualizes the outcomes of the hard and soft clustering. Similarly to KSC, the soft clustering results are based on the distance between the final centroids and the data-points in the input space (see [3]). Concerning data-sets DS_I and DS_III the results of KSC and k-means are very similar, while in the data-set DS_II analysis k-means performs worse than KSC, indicating the need of maintenance where it was not really performed. In this case k-means would suggest too many maintenance actions, which is not cost-effective. Finally, in table 1 the performance of KSC and k-means are evaluated according to a standard internal cluster quality measures, that is the mean silhouette index [14].

5.2. Supervised learning

In this paragraph we present the results of applying the NAR model explained in 3.2 for predicting the evolution of the hot area pixels in data-set DS_III. The optimal parameters obtained using the 10 fold cross-validation procedure summarized in Section 3.3 are: $\gamma = 1356.7$, $\sigma^2 = 301.6$. Moreover the optimal model order, tuned using a grid search approach, is $p = 5$, as depicted in Fig. 9. After selecting the optimal parameters, we performed multi-step ahead recursive prediction. The NAR model performed very well up to 10 steps ahead prediction, i.e. about 25 seconds since a bag is processed every $2 - 3$ seconds. Afterward the mean absolute error between outputs and actual values start increasing progressively, even if the predictions remains good up to 1 minute ahead prediction. In Fig. 10 we visualize the results, summarized also in Table 2, where a comparison with the standard baseline and a linear autoregressive model (AR) is shown.

6. Discussion

In the first part of this work we explained the use of KSC, which has been chosen instead of classification because of the high unbalance of the data (few maintenance events compared to normal operating condition). After applying a windowing operation on the data in order to catch the deterioration process affecting the sealing jaws, we showed how KSC is able to recognize the presence of at least two working regimes of the VFFS machine, identifiable respectively as normal and critical operating condition. Moreover we proposed also a soft

clustering output that can be interpreted as "probability" to maintenance, and it is more directly related to the degradation phenomenon affecting the jaws. KSC is also compared with the standard clustering method used in industrial applications, namely k-means. While in data-sets DS_I and DS_III they give quite similar results, in case of data-set DS_II k-means performs worse than KSC, since it predicts maintenance in regions where it was not actually performed by the operator.

In the second part a LS-SVM NAR model for predicting the evolution of the dirt accumulation in the jaws has been constructed. The dirt accumulation is directly measured by means of the number of hot area pixels present in the images obtained by the thermal camera. First we selected the parameters of the model using 10-fold cross-validation with coupled simulated annealing for γ and σ^2 and a grid search approach for the optimal model order. Then we trained the NAR model and we tested its ability in predicting the evolution of the number of hot area pixels. The performance is very good up to 10 steps ahead prediction (around 25 seconds), even if the forecast remains good up to 1 minute ahead. Moreover, the NAR outperforms the linear auto-regressive model (AR).

To summarize, both the LS-SVM NAR model and KSC, thanks to their forecasting capabilities, could help to optimize the timing of maintenance actions for the machine under study. In particular, a company who would like to use the LS-SVM techniques described in this work to maximize production capacity has two choices:

- option 1 (the cheapest): install on every packing machine only accelerometers and use the proposed unsupervised learning method, that is KSC, to predict in advance, in an online fashion, when the machine starts entering critical conditions. In this way the operator can perform maintenance at the right time and avoid long stops of the production.

- option 2: install the thermal camera and use the LS-SVM NAR model to predict up to 1 minute ahead the dirt accumulation on the jaws. Since the thermal camera costs more than the accelerometers, this option is more expensive. Nevertheless it is more reliable because it is a direct prediction of the deterioration while KSC, as surprisingly effective as it is completely unsupervised, infers the degradation indirectly from the vibration signals.

7. Conclusions

Predictive maintenance of industrial plants is receiving increasing attention in the last years due to its many advantages, like cost efficiency,reduced ecologi-

cal impact etc. It is based on constant monitoring the health of machines joined with advanced signal processing and prognostics techniques for optimal maintenance management. In this paper we proposed a least squares support vector machines (LS-SVM) framework for maintenance strategy optimization based on real-time condition monitoring of an industrial machine. We presented an unsupervised approach through kernel spectral clustering (KSC), and a supervised learning method, namely nonlinear auto-regression (NAR). In the first case we used the data collected by accelerometers positioned on the jaws of a Vertical Form Fill and Seal (VFFS) machine, while for the time-series analysis with the NAR model data acquired by a thermal camera (directly measuring the dirt accumulation in the jaws) are processed. Both the unsupervised and the supervised LS-SVM showed high predictive performance, and can help to schedule maintenance on the packing machine in order to maximize production capacity.

Acknowledgements

This work was supported by Research Council KUL: ERC Adv. A-DATADRIVE-B, GOA/11/05 Ambiorics, GOA/10/09 MaNet , CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants;Flemish Government:FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC) IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940); Contract Research: AMINAL; Other:Helmholtz: viCERP, ACCM, Bauknecht, Hoerbiger. Carlos Alzate is a research scientist at IBM's Smarter Cities Technology Center in Dublin, Ireland, and a postdoctoral fellow of the Research Foundation - Flanders (FWO). Bart De Ketelaere is Industrial Research Manager sponsored by the Industrieel Onderzoeksfonds (IOF) of the KU Leuven. Wannes Meert is postdoctoral researcher in Machine Learning in the Declarative Languages and Artificial Intelligence group of the Departement of Computer Science at the KU Leuven. Johan Suykens is a professor at the KU Leuven, Belgium. The scientific responsibility is assumed by its authors.

References

[1] Alzate, C., Suykens, J. A. K., 2010. Highly sparse kernel spectral clustering with predictive out-of-sample extensions. In: Proc. of the 18th European Symposium on Artificial Neural Networks (ESANN 2010). pp. 235–240.

[2] Alzate, C., Suykens, J. A. K., February 2010. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2), 335–347.

[3] Ben-Israel, A., Iyigun, C., Jun. 2008. Probabilistic d-clustering. J. Classif. 25 (1), 5–26.

[4] Berkhin, P., 2006. A survey of clustering data mining techniques. Grouping Multidimensional Data, 25–71.

[5] Choi, S. W., Yoo, C. K., Lee, I.-B., 2003. Overall statistical monitoring of static and dynamic patterns. Ind. Eng. Chem. Res. 42, 108 – 117.

[6] Chung, F. R. K., 1997. Spectral Graph Theory. American Mathematical Society.

[7] Kourti, T., MacGregor, J. F., 1995. Process analysis, monitoring and diagnosis, using multivariate projection methods. Chemometrics and Intelligent Laboratory Systems 28 (1), 3 – 21.

[8] Kriegel, H.-P., Kroger, P., Zimek, A., 2010. Outlier detection techniques. 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD).

[9] Langone, R., Alzate, C., De Ketelaere, B., Suykens, J. A. K., 2013. Kernel spectral clustering for predicting maintenance of industrial machines. In: IEEE Symposium Series on Computational Intelligence (SSCI) 2013.

[10] Ljung, L. (Ed.), 1999. System identification (2nd ed.): theory for the user. Prentice Hall PTR.

[11] MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: of California Press., U. (Ed.), Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. pp. 281–297.

[12] Murty, K. G. (Ed.), 1983. Linear programming. Wiley.

[13] Ng, A. Y., Jordan, M. I., Weiss, Y., 2002. On spectral clustering: Analysis and an algorithm. In: Dietterich, T. G., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Information Processing Systems 14. MIT Press, Cambridge, MA, pp. 849–856.

[14] Rousseeuw, P., Nov. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20 (1), 53–65.

[15] Suykens, J. A. K., Alzate, C., Pelckmans, K., 2010. Primal and dual model representations in kernel-based learning. Statistics Surveys 4, 148–183.

[16] Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J., 2002. Least Squares Support Vector Machines. World Scientific, Singapore.

[17] Van Gestel, T., Suykens, J. A., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G., de Moor, B., Vandewalle, J., 2004. Benchmarking least squares support vector machine classifiers. Machine Learning 54 (1), 5–32.

[18] Venkatasubramanian, V., Rengaswamy, R., Kavuri, S., 2003. A review of process fault detection and diagnosis. part i: Quantitative model-based methods. Computers and chemical engineering 27 (3), 293–311.

[19] Venkatasubramanian, V., Rengaswamy, R., Kavuri, S., 2003. A review of process fault detection and diagnosis. part ii: Qualitative models and search strategies. Computers and chemical engineering 27 (3), 313–326.

[20] Venkatasubramanian, V., Rengaswamy, R., Kavuri, S., 2003. A review of process fault detection and diagnosis. part iii: Process history based methods. Computers and chemical engineering 27 (3), 327–346.

[21] von Luxburg, U., 2007. A tutorial on spectral clustering. Statistics and Computing 17 (4), 395–416.

[22] Xavier-De-Souza, S., Suykens, J. A. K., Vandewalle, J., Bollé, D., Apr. 2010. Coupled simulated annealing. Trans. Sys. Man Cyber. Part B 40 (2), 320–335.

	DS_I	DS_II	DS_III
KSC	**0.29**	**0.25**	0.27
k-means	0.28	0.12	0.27

Table 1: **Cluster quality evaluation.** Mean Silhouette index, best performance in bold. In the case of data-set DS_II the low value of Silhouette indicates that k-means does not succeed in correctly separating the normal behavior and the maintenance cluster, as KSC on the other hand does.

	LS-SVM NAR	AR	Standard baseline
MAE	**227.03**	247.25	371.99
Percentage error	**1.42%**	4.11%	6.17%
R^2	**0.93**	0.92	0.77

Table 2: **Summary NAR performance on test data.** Best performance in bold. By standard baseline we mean a model that, at each step, use as prediction the previous value.

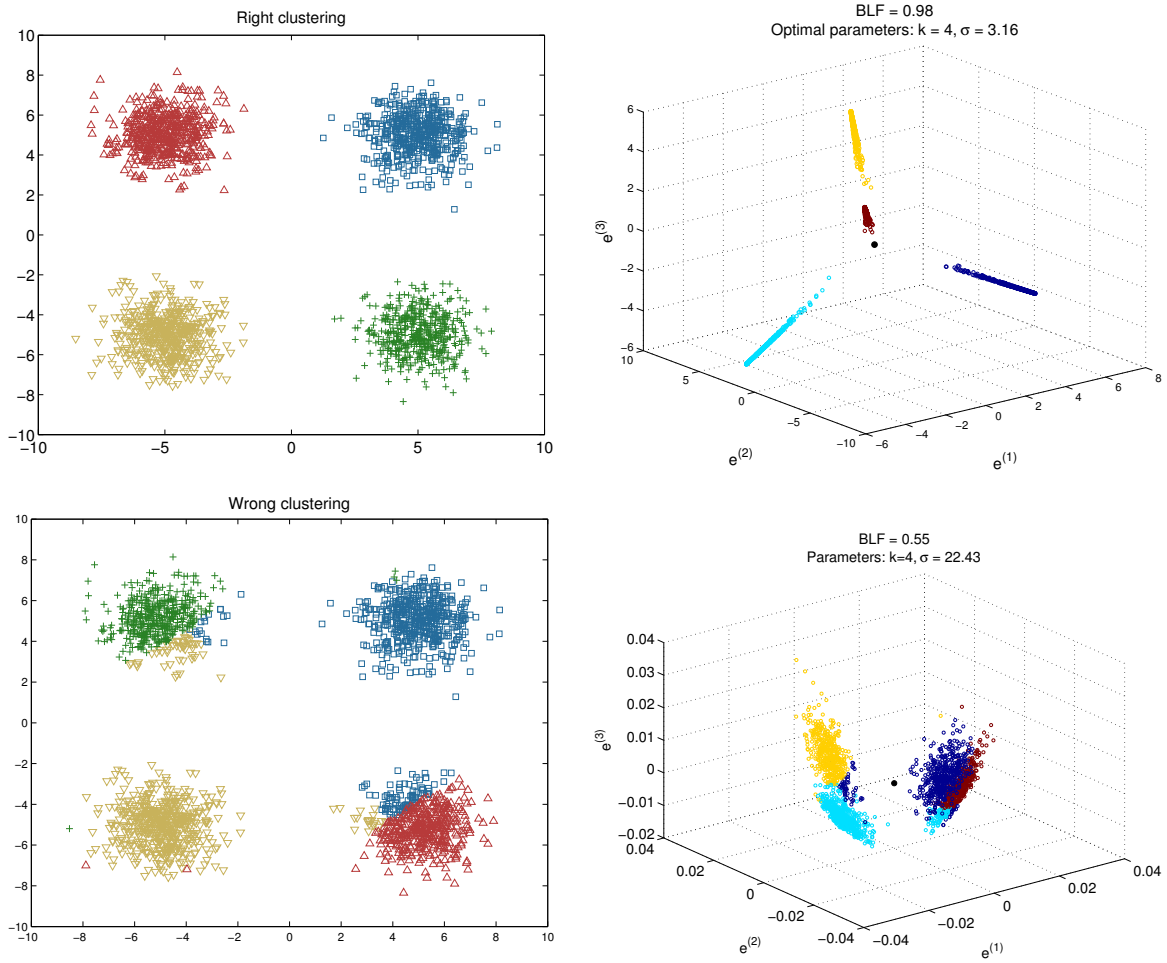


Figure 1: **Model selection illustrative example:** for well-chosen kernel parameters, the clusters for out-of-sample data are collinear in the projection space. **Top left:** Correct clustering of 4 Gaussian clouds. **Top right:** Projection space corresponding to an optimal σ value, for which the BLF is maximal. **Bottom left:** Wrong clustering results. **Bottom right:** Projection space for a wrong σ value.

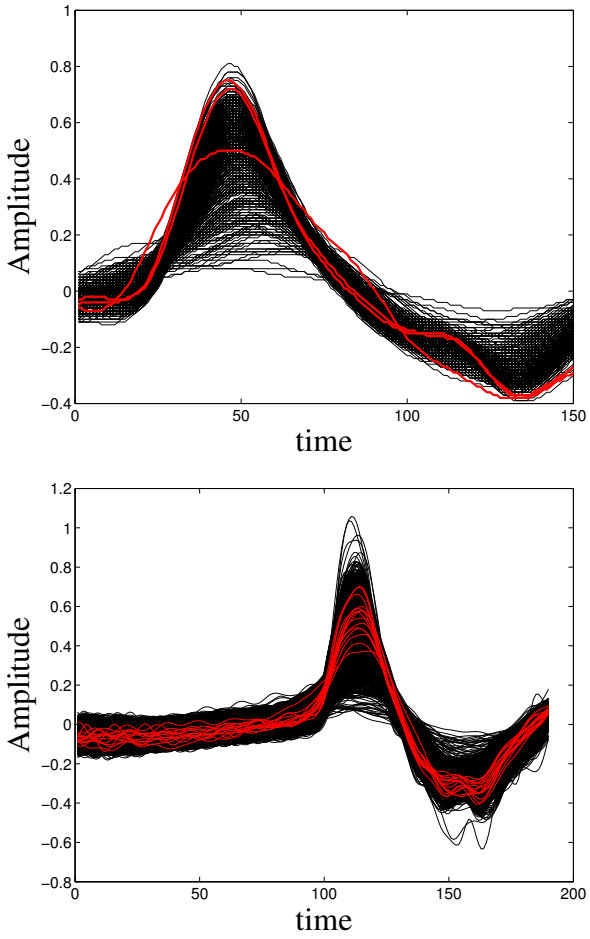


Figure 2: **Data-sets DS_I-DS_II. Top:** Accelerometer signals for the data-set DS_I. **Bottom:** Accelerometer signals for the entire data-set DS_II. The signals corresponding to maintenance actions are depicted in red (best visible in colors).

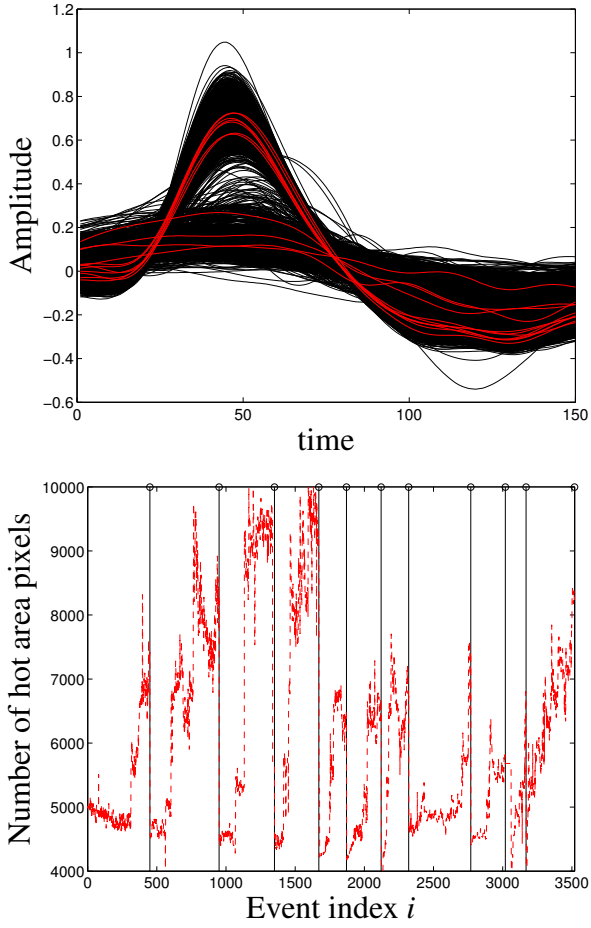


Figure 3: **Data-set DS_III. Top:** Accelerometer signals (signals corresponding to maintenance actions are pictured in red). **Bottom:** thermal camera data (the vertical black lines indicates true maintenance actions).

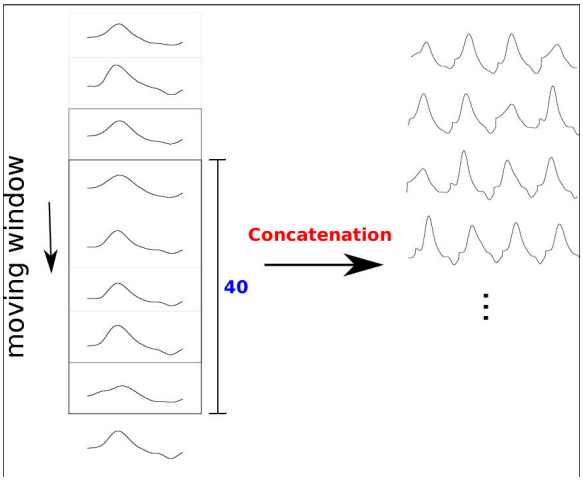


Figure 4: Concatenation of accelerometer signals. After the windowing operation, each data-point is now a time-series of dimension $d = 40 \times 150$ for the first and data-set DS_IIIs and $d = 40 \times 190$ for the data-set DS_II.

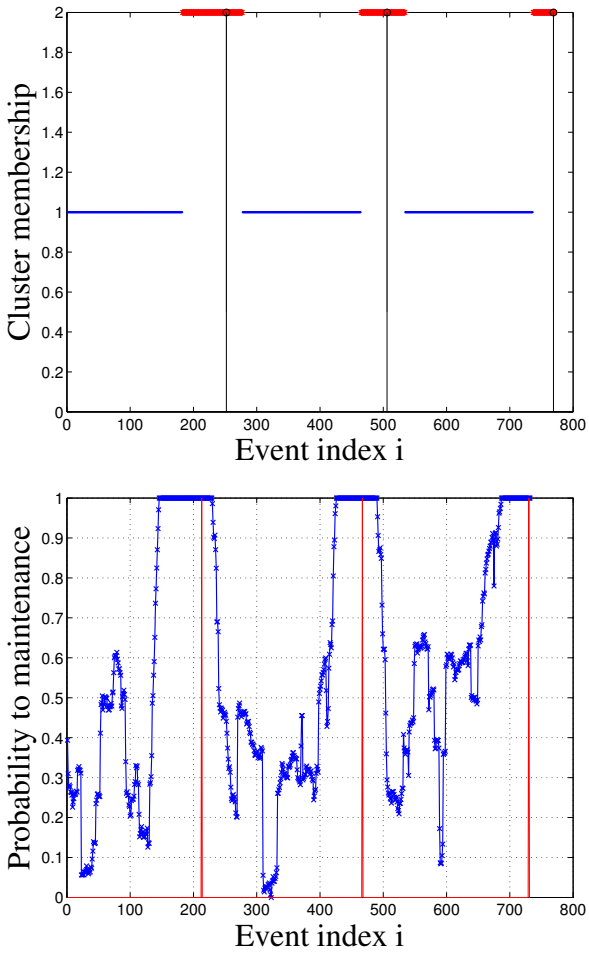


Figure 5: **KSC results data-set DS_I. Top:** Hard clustering results for the whole data-set. Cluster 2 represents predicted maintenance events. The vertical black lines show the true maintenance. **Bottom:** Soft clustering results in terms of probability to maintenance.

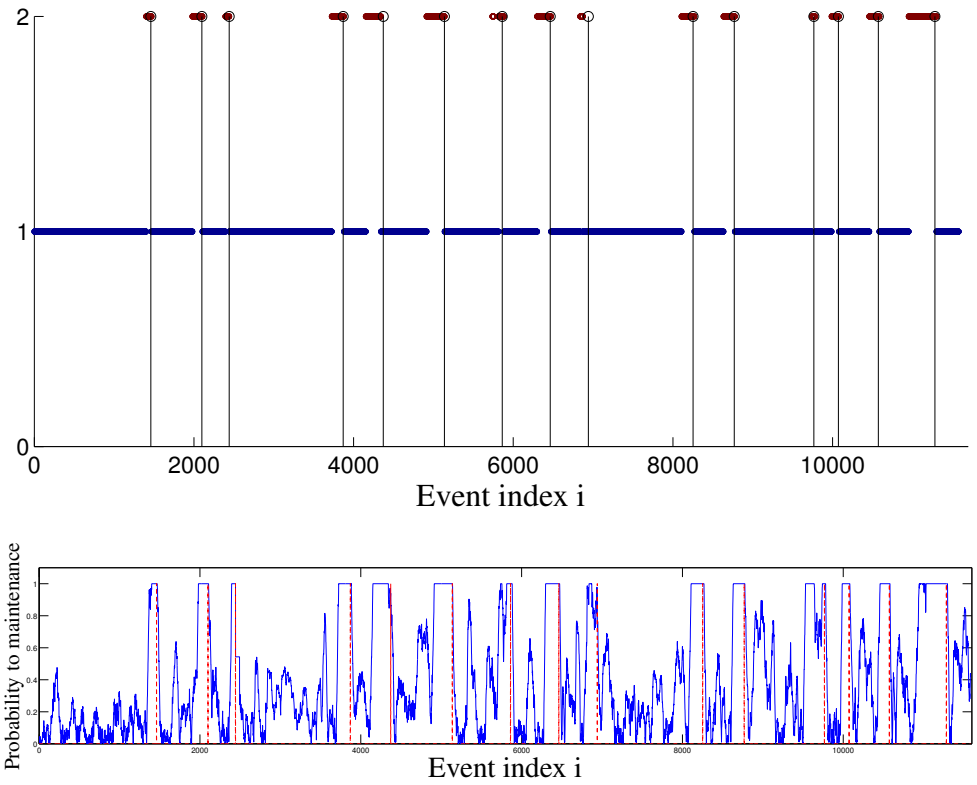


Figure 6: **KSC results data-set DS_II. Top:** Hard clustering, cluster 2 represents predicted maintenance events. **Bottom:** Soft clustering in terms of probability to maintenance.

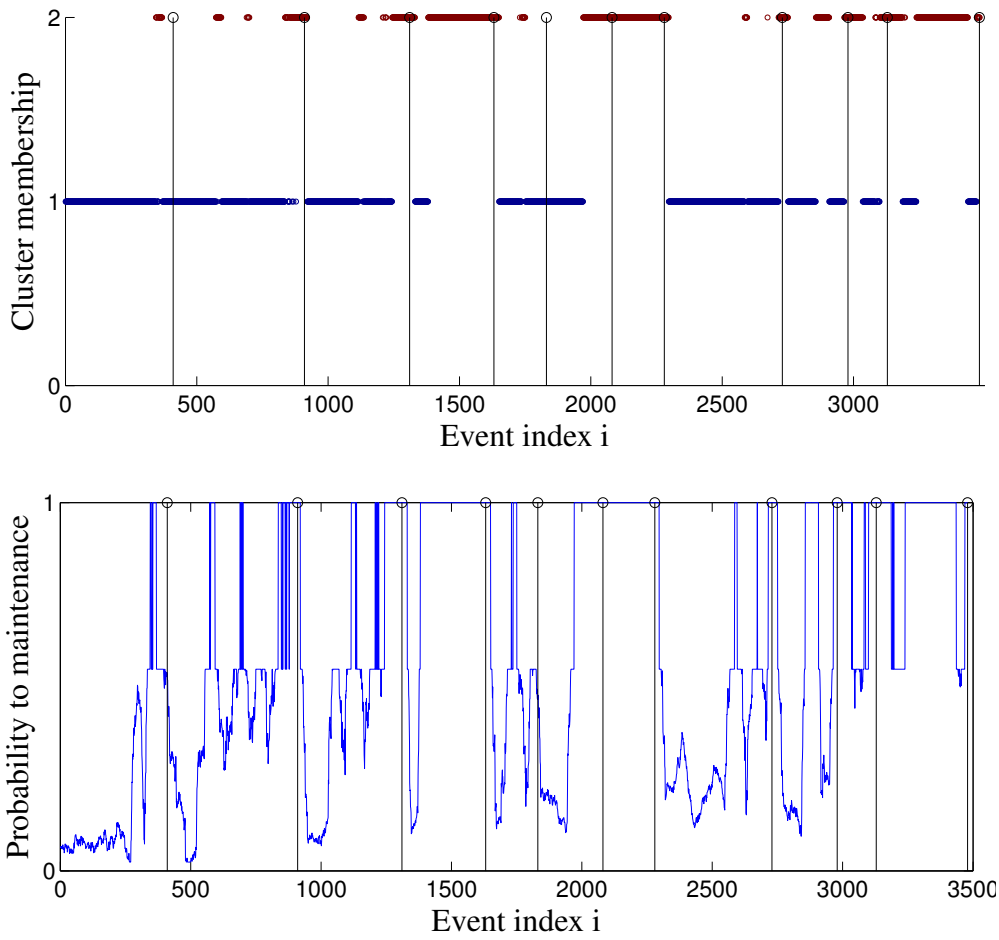


Figure 7: **KSC results data-set DS_III. Top:** Hard clustering, cluster 2 represents predicted maintenance events. **Bottom:** Soft clustering in terms of probability to maintenance.

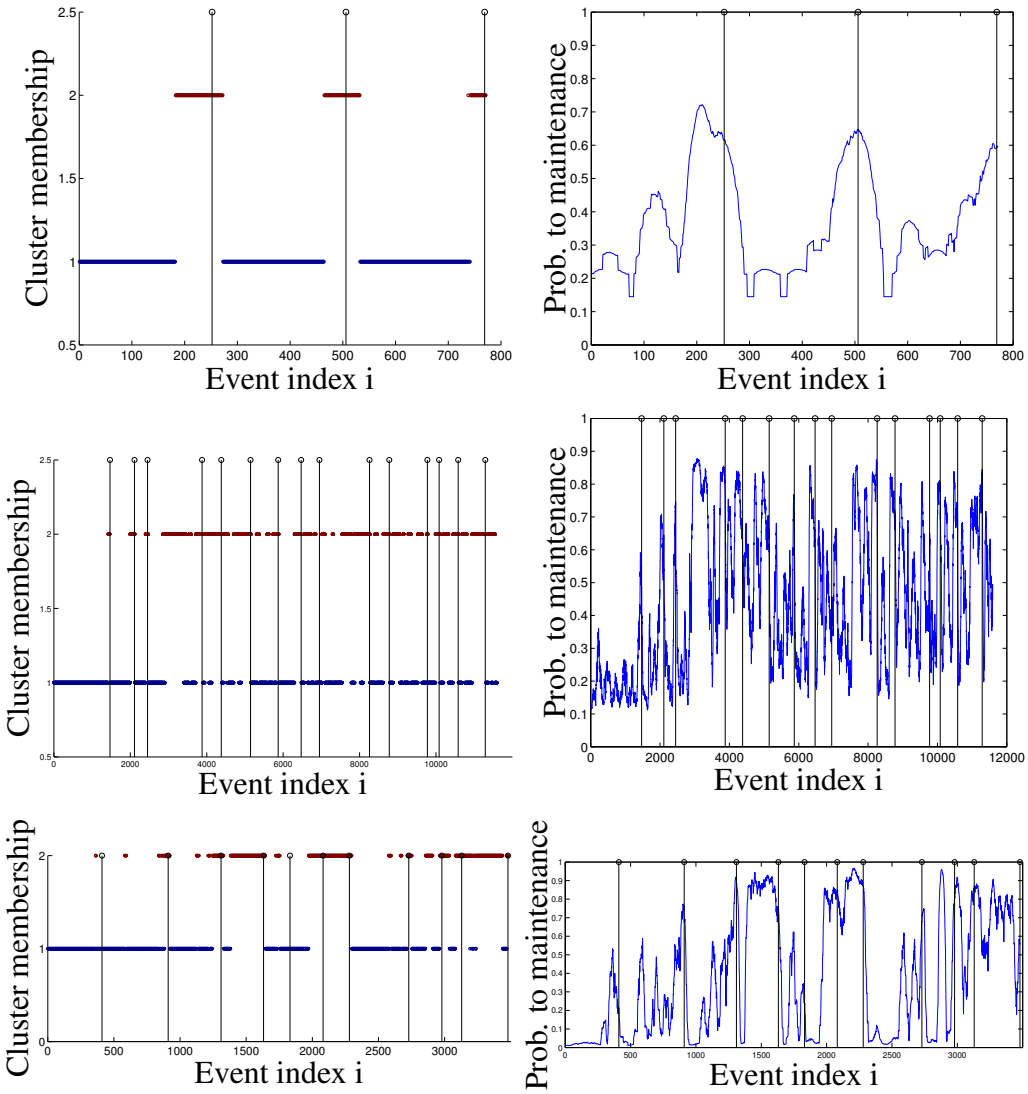


Figure 8: **k-means results.** Cluster 1 symbolizes the normal operating condition, cluster 2 represents predicted maintenance events. The vertical black lines show the true maintenance. **Top**: data-set DS_I. The results are similar to KSC outcomes, even if slightly worse since in the end there is a kind of false alarm (a single prediction of maintenance followed by normal behavior, before the final maintenance cluster). **Center**: data-set DS_II. In this case k-means performs much worse than KSC, suggesting maintenance in regions not corresponding to actual maintenance events. **Bottom**: data-set DS_III. The results of KSC and k-means are very similar. Even if both techniques predict maintenance when it has not been really performed, it is quite likely that in this case the operator should have performed the maintenance actions in a different way he did (see discussion in the end of Section 5.1.3).

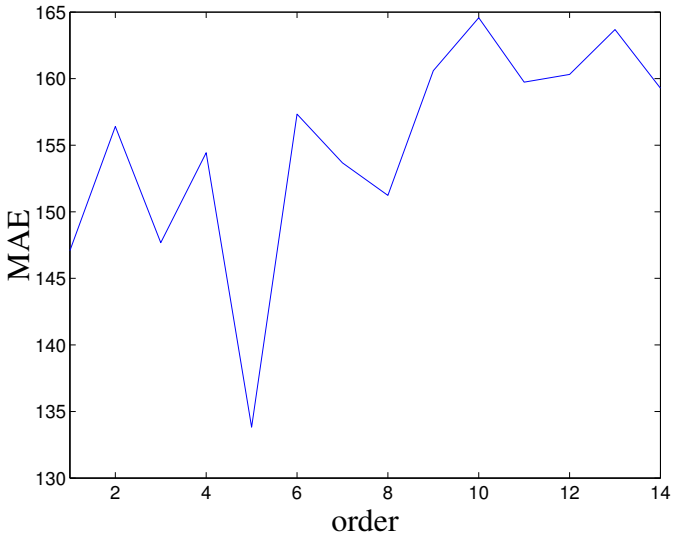


Figure 9: **Tuning order NAR.** Mean absolute error (MAE) between one step ahead prediction and true values with respect to the order of the NAR model. A clear minimum is present at $p = 5$, suggesting to consider this value as the optimal parameter to construct the final NAR model.

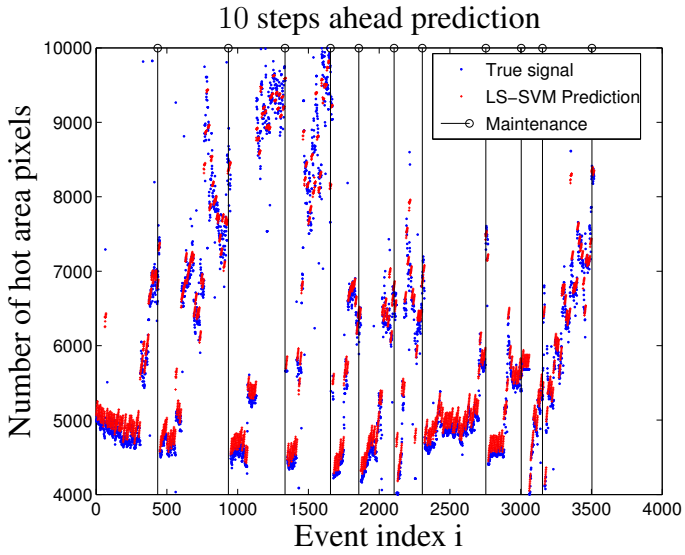


Figure 10: **Prediction NAR.** The NAR model is able to predict very well even 10 steps ahead the future behavior of the thermal camera data based on a window of the 5 past values.