Kernel spectral clustering for predicting maintenance of industrial machines

Rocco Langone[1], Carlos Alzate[1,2], Bart De Ketelaere[3], Johan A. K. Suykens[1]
[1]Department of Electrical Engineering (ESAT), SCD, KU Leuven, B-3001 Leuven Belgium
Email: {rocco.langone,carlos.alzate,johan.suykens}@esat.kuleuven.be
[3]Faculty of Bioscience Engineering, BIOSYST MeBioS Qualimatrics, KU Leuven, B-3001 Leuven Belgium
Email: bart.deketelaere@biw.kuleuven.be
[2]Smarter Cities Technology Center, IBM Research-Ireland
Email: carlos.alzate@ie.ibm.com

***Abstract*—Early and accurate fault detection in modern industrial machines is crucial in order to minimize downtime, increase the safety of plant operations, and reduce manufacturing costs. The process monitoring techniques that have been most effective in practice are based on the analysis of historical process data. In this paper we present a novel approach that uses Kernel Spectral Clustering (KSC) on the sensor data to distinguish between normal operating condition and abnormal situations. In other words, the main contribution is to show how KSC can be a valid tool also for outlier detection, a field where other techniques are more popular. KSC is a state-of-the-art unsupervised learning technique with out-of-sample ability and a systematic model selection scheme. Thanks to the abovementioned characteristics and the capability of discovering complex clustering boundaries, KSC is able to detect in advance the need of maintenance actions in the analyzed machine.**

I. Introduction

In industrial processes fault detection, isolation and diagnosis ensure product quality and operational safety. Traditionally, 4 ways to deal with sensory faults have been used [1],[2],[3]: corrective maintenance, preventive maintenance, manual predictive maintenance, and condition-based maintenance. The first type is performed only when the machine fails, it is expensive and safety and environment issues arise. Preventive maintenance is based on periodic replacement of components. The rough estimation of parts lifetime causes a non-optimal use of parts, and possible unexpected failures can still occur (with downtime, safety and environmental consequences). In predictive maintenance machines are manually checked with expensive monitoring hardware (termography, motor health, bearing health). 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 receiving increasing attention due to its many advantages. Machines status is automatically collected and centrally analyzed, and maintenance is planned based on the results of the analysis. 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 efficient way. With the development of information and sensor technology many process variables in a power plant can be sampled, like 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 behaviour 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), just to name the most widely used methods [4], [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 [6]. Within the unsupervised techniques, statistical models and models based on spatial proximity are the most popular. In this study, since the data are highly unbalanced for supervised learning, we use an unsupervised learning technique called kernel spectral clustering (KSC) [7]. KSC represents a spectral clustering formulation as a weighted kernel PCA problem with primal and dual representations, cast in the LS-SVM framework [8]. The KSC 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. The 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 to discover when the plant enters critical conditions, in an online fashion. To summarize, we show how KSC can be used in an effective way for outlier detection. Comparison with other techniques is out of scope of this paper. The remaining part of this paper includes: Section II, which summarizes the KSC model and the model selection scheme. Section III describes the data-sets used in the experiments. In Section IV the simulation results are presented. Finally, in Section V we give some conclusions and perspectives.

II. Kernel Spectral Clustering

A. Model

Spectral clustering methods make use of the eigenvectors of the Laplacian to find useful partitions of the data [9], [10], [11]. Given training data $\mathcal{D} = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d$ and the number of clusters k, the primal problem of spectral clustering via weighted kernel PCA is formulated as follows [7]:

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

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

where $e^{(l)} = [e_1^{(l)}, \ldots, e_N^{(l)}]^T$ are the projections, $l = 1, \ldots, k-1$ indicates 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 expressed by:

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

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

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

where Ω is the kernel matrix with ij-th entry $\Omega_{ij} = K(x_i, x_j)$, 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. $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the kernel function and captures the similarity between the data-points. In the experiments described in section IV we use the RBF kernel function given by $K(x_i, x_j) = \exp(-||x_i - x_j||_2^2/\sigma^2)$, where σ is the bandwidth parameter. The out-of-sample extension to new nodes is done by an (Error Correcting Output Codes) ECOC decoding procedure. The decoding scheme consists of comparing the cluster indicators obtained in the validation/test stage with the codebook and selecting the nearest codeword in terms of Hamming distance. The cluster indicators can be obtained by binarizing the score variables for test points as follows:

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

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

B. Tuning scheme

A proper way of choosing the tuning parameters in a kernel model is of a crucial importance. Usually the number of clusters in which to group the data and eventually the parameters of the kernel function have to be selected carefully to achieve good performances. In this section we describe the model selection scheme used in the experiments, namely the Balanced Line Fit BLF. The BLF criterion exploits the shape of the points in the projections space: it reaches its maximum value 1 when the clusters do not overlap, and in this ideal situation the clusters are represented as lines in this space. In particular the BLF is defined in the following way [7]:

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

where $\mathcal{D}^V$ represents the validation set and k as usual indicates the number of clusters. The linefit index equals 0 when the score variables are distributed spherically and equals 1 when the score variables are collinear (representing points in the same cluster). The balance index equals 1 when the clusters have the same number of elements 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 values 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 kernel machines using the training set
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.

III. Data-sets

The data are collected from a Vertical Form Fill and Seal (VFFS) machine used for filling and sealing packages in different industries, mainly food industry. The VFFS machine supplies film from a roll which is formed into a bag over the vertical cylinder. Pressing 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. From a market study conducted in the past at different industries using this kind of machine the dirt accumulation on sealing jaws seems to influence strongly the process quality. Therefore in the experiments that have been conducted the jaws were monitored to predict in advance the maintenance actions. In particular accelerometers mounted on the jaws in order to measure the dirt accumulation have been installed. A total of 2 experiments have been made and then we have 2 data-sets available:

- DS_I: this dataset consists of 771 events and 3 external maintenance actions. An event is related to a particular processed bag and takes place every two seconds. For each event we have two kinds of data: a feature variable named klasse that gives an instantaneous measure of sealing quality using PCA (klasse = 1 for a good sealed

bag and klasse = 0 for a bad sealed bag) and a 150-dimensional accelerometer signal (see top of Fig.1).

- DS_II: it contains a total of 11 632 processed bags and 10 maintenance events. 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.1). Also for this data-set we are provided with the klasse feature variable.

IV. Experimental results

In this section we show how KSC can be a valid method to perform just-in-time maintenance, not too soon in order to take full advantage of component lifetime but also not too late in order to avoid catastrophic failures and unplanned downtimes. In particular we are able to identify 2 regimes that we can interpret as normal behaviour and critical conditions (need of maintenance). Moreover a probabilistic interpretation of the results is also provided, which can better describe the degradation process experienced by the sealing jaws of the packing machine. We perform clustering both on the feature variable klasse and the raw accelerometer signals.

A. 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, like depicted in Fig. 2 (for the vibration signals) and Fig. 3 (for the feature variable klasse. We have then a total of 3 parameters to determine: the window size (i.e. the number of signals/features to take into account), the number of clusters k and the RBF kernel parameter σ (see section II-A). According to the BLF criterion the optimal window size is 40 and the optimal number of clusters is $k = 2$ for both the data-sets [1], while σ is data-set dependent. An example of the tuning procedure is shown in Fig 4.

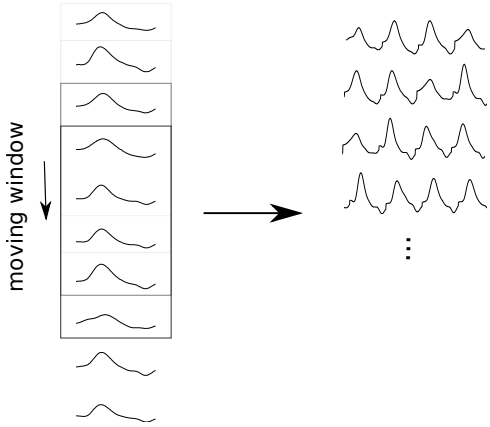


Fig. 2. Concatenation of accelerometer signals. After the windowing operation, each data-point is now a time-series of dimension $d = 40 \times 150$ for the first data-set and $d = 40 \times 190$ for the second data-set.

[1]The results are similar for the raw signals and the feature variable klasse.

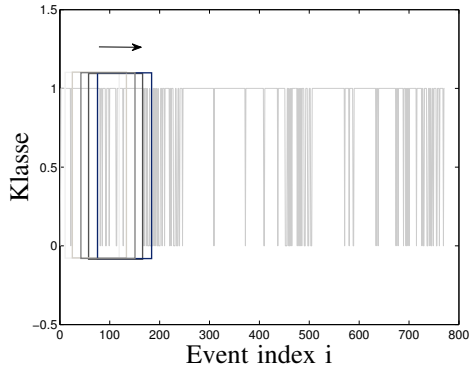


Fig. 3. Moving window with overlap on the feature variable klasse to create the dataset to be fed into the clustering model. The latter is a $N_{\text{test}} \times d$ data matrix, with $N_{\text{test}} = 771$ for data-set DS_I and $N_{\text{test}} = 11\,632$ for data-set DS_II, and $d = 40$ for both data-sets.

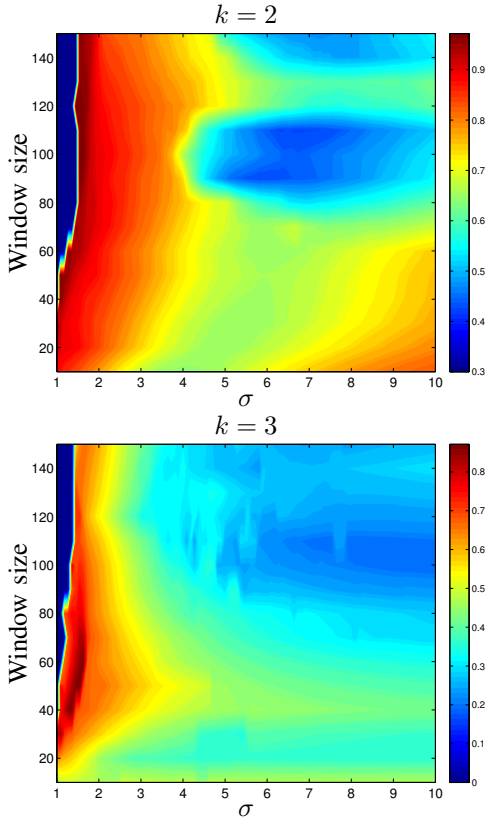


Fig. 4. Model selection surfaces for the first data-set, only the results for $k = 2$ and $k = 3$ are shown. If we consider more clusters ($k > 3$) the maximum value of the BLF decreases. The outcome is similar for the other data-set.

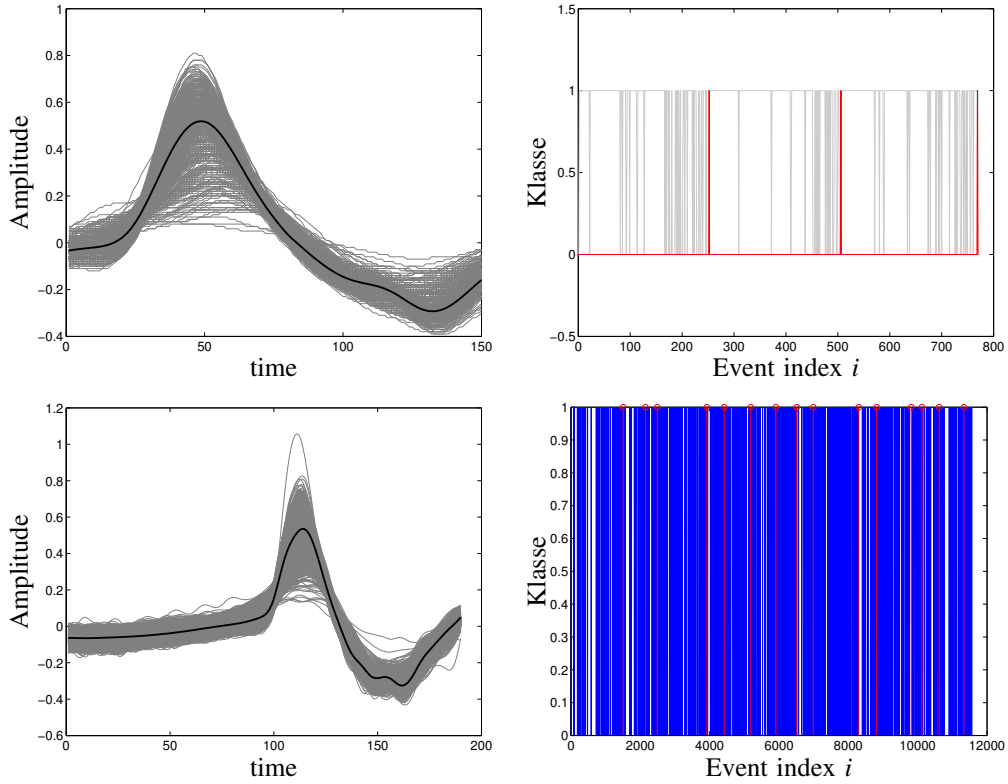


Fig. 1. **Top:** Accelerometer signals (with average in bold) and the feature variable klasse for the first data-set (maintenance in red). For what concerns the vibration signals, the plot is related to the whole data, which includes normal operating conditions and maintenance. **Bottom:** Accelerometer signals (with average in bold) and the feature variable klasse for the entire data-set DS_II (best visible in colors).

B. Hard clustering

Here we present the clustering output of the KSC model when tested on the two data-sets under investigation. By using the accelerometer signals we obtain good results on DS_I and DS_II, while if we use the feature variable to feed the clustering model we have meaningful results only on DS_I. This can be explained by considering that the feature variable klasse is extracted by using PCA. Since kernel spectral clustering can be seen as a form of weighted kernel PCA, the feature extraction and clustering processes are performed by the same model when it is fed by the vibration signals. In Fig. 5 an example of KSC prediction for the first dataset is shown. We can interpret the clusters as normal behaviour and maintenance cluster. In fact, if we examine the prototypes of the two clusters we understand how they are related to the frequency of bad sealed bags. The normal behaviour prototype describes a window of 40 perfected sealed bags, with the variable $1-$ klasse taking always the value zero. On the other hand the maintenance cluster is characterized by a big number of bad sealed bags (klasse $= 0$, $1-$ klasse $= 1$). Notice that the KSC model is able to predict some minutes in advance the maintenance actions before they are actually performed by the operator.

C. Soft clustering

In the previous section we demonstrated the effectiveness of KSC in predicting in advance the maintenance events. Neverthless 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 II-A). The latent variable provides a more informative output which can be analyzed in order to produce a better prognostic output. The latent variable for the data-set DS_I can be seen in the top of Fig. 6. The black dots show the latent variable value when moving the sliding window. Maintenance is predicted when the value becomes positive (the red zone). The latent variable 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. Since the range in which the latent variable takes 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, we rescale the score variable between 0 and 1. This transformation is based on the structure of the latent variable space. Since every cluster (normal operation and maintenance) is ideally represented as a line in the latent variable space and decisions are taken based on binarization, it makes sense to consider points far away from the origin as prototypical for the cluster they are in. These points have more certainty to belong to the given cluster because they are further from the decision

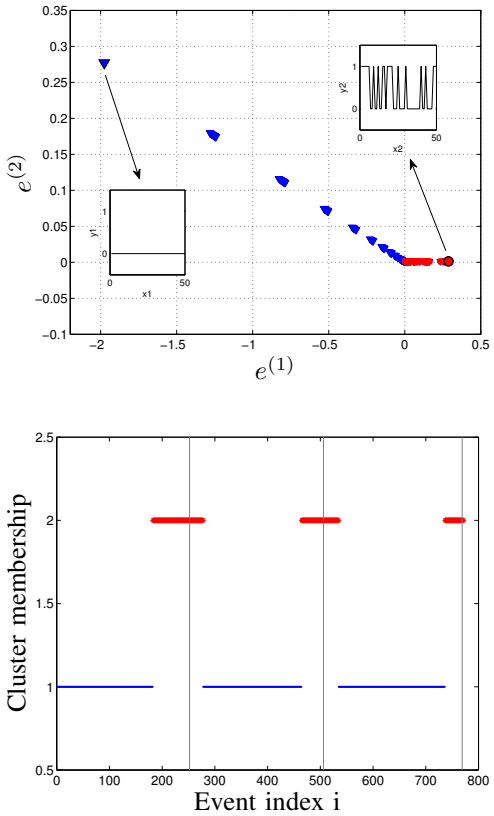


Fig. 5. **Top:** Visualization of the projected variables on *validation* data and the corresponding cluster prototypes for $k = 2$ clusters, when clustering the klasse variable. Note the (desirable) strong line structure of the clusters (corresponding to high value of the BLF). The tips of the lines are the prototypes of the normal condition (blu) and *maintenance* cluster (red circles). **Bottom:** Clustering results for the whole dataset. Cluster 2 represents predicted maintenance events. The vertical gray lines show the true maintenance. Similar results are obtained when KSC is fed by the raw accelermeter signals.

boundaries [12]. 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. 6 (first data-set) and in Fig.8 (second data-set). The value can now be considered as a "probability" to maintenance [13].

D. Non-decreasing probability

As can be seen from Fig. 6 (bottom) and Fig. 8 the value of the transformed latent variable is not always increasing. As already pointed out, this is due to the fact that the moving window can enter zones with good seals after a period of bad seals but not enough to trigger maintenance. Since an increasing output is very desirable as it resembles a degradation of the sealing quality over time, we can incorporate this property into the clustering model. For this purpose we post-processed the probability to maintenance output by keeping the maximum value found until event index i. The result is shown in Fig. 7 for data-set DS_I and in Fig. 9 for data-set DS_II. We can notice 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).

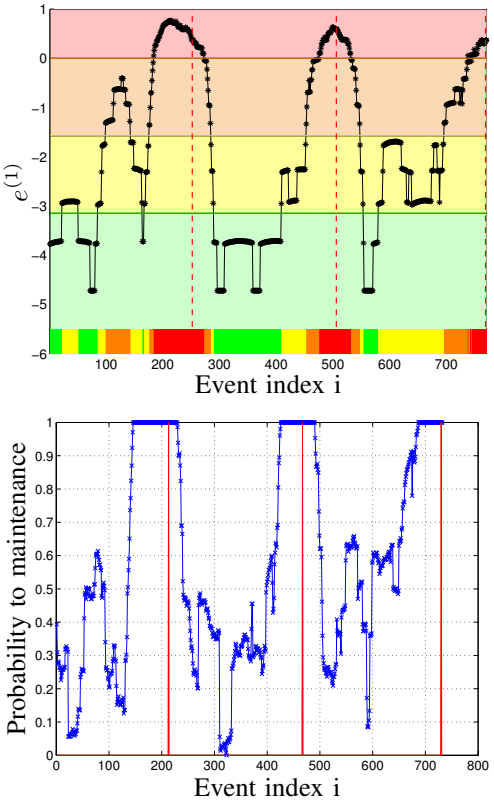


Fig. 6. **Top:** Latent variable for dataset DS_I. Maintenance is predicted when the value becomes positive. Actual maintenance events are depicted as red dashed lines. The sequence at the bottom indicates the transitions of the latent variable into the colored zones. This information can be used as an input feature for further analysis. **Bottom:** The latent variable is rescaled between 0 and 1 and can be interpreted as probability to maintenance in the sense explained in section IV-C.

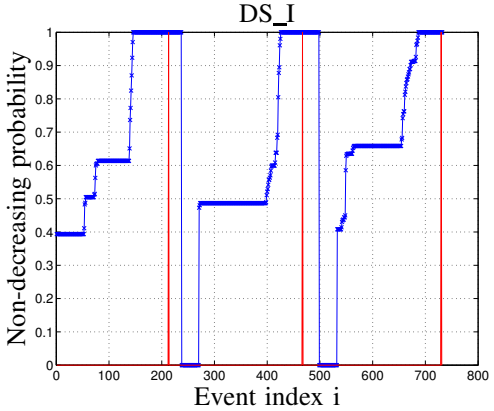


Fig. 7. Non-decreasing probabilistic output for the first data-set. Actual maintenance actions events are represented by the three red lines.

V. Conclusion

Predictive maintenance of industrial plants is receiving increasing attention in the last years due to its many advantages,

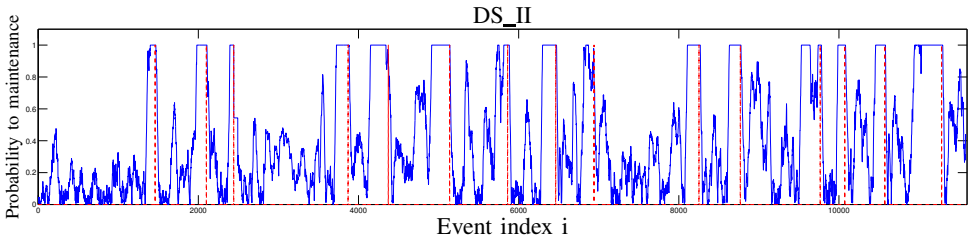


Fig. 8. Soft output in terms of probability for the second data-set. Actual maintenance in red.

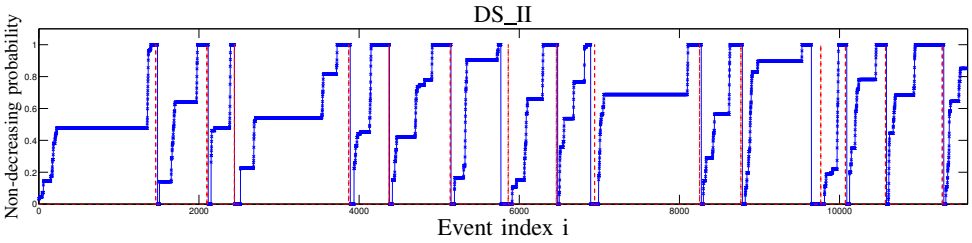


Fig. 9. Non-decreasing probabilistic output for the second data-set. Actual maintenance actions events are represented by the red lines.The model has good generalization capabilities giving a high probability value in the zones just before maintenance.

like cost efficiency and automation. It is based on constant monitoring the health of machines coupled with advanced signal processing, expert knowledge, system modelling, prognostics and maintenance management optimization. In this paper we proposed a model for maintenance strategy optimization based on real-time condition monitoring of an industrial machine. We used the data collected by accelerometers positioned on the jaws of a Vertical Form Fill and Seal (VFFS) machine. In particular, since the available data were very unbalanced (few maintenance events compared to normal operating condition) we proposed an unsupervised learning approach based on kernel spectral clustering (KSC). 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, identifiables respectively as normal and critical operating condition. Moreover we proposed also a soft clustering output that can be interpreted as "probability" to maintenance. In this way KSC could help to optimize the timing of maintenance actions for the machine under study. For example in future the KSC output could represent the input of a maintenance management model. By means of the latter the total profit per bag would be monitored continuously and updated at each newly produced bag. This would give the operator a direct overview of the performance of the machine. When the performance or profit generated by the machine drops, this information could be used by the operator to optimally schedule maintenance on the packing machine in a cost efficient way.

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 Industial Research Manager sponsored by the Industrieel Onderzoeksfonds (IOF) of the KU Leuven. Johan Suykens is a professor at the KU Leuven, Belgium. The scientific responsibility is assumed by its authors.

References

[1] V. Venkatasubramanian, R. Rengaswamy, and S. Kavuri, "A review of process fault detection and diagnosis. part i: Quantitative model-based methods," *Computers and chemical engineering*, vol. 27, no. 3, pp. 293–311, 2003.

[2] ——, "A review of process fault detection and diagnosis. part ii: Qualitative models and search strategies," *Computers and chemical engineering*, vol. 27, no. 3, pp. 313–326, 2003.

[3] ——, "A review of process fault detection and diagnosis. part iii: Process history based methods," *Computers and chemical engineering*, vol. 27, no. 3, pp. 327–346, 2003.

[4] T. Kourti and J. F. MacGregor, "Process analysis, monitoring and diagnosis, using multivariate projection methods," *Chemometrics and Intelligent Laboratory Systems*, vol. 28, no. 1, pp. 3 – 21, 1995.

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

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

[7] C. Alzate and J. A. K. Suykens, "Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 335–347, February 2010.

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

[9] F. R. K. Chung, *Spectral Graph Theory*. American Mathematical Society, 1997.
[10] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.
[11] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002, pp. 849–856.
[12] C. Alzate and J. A. K. Suykens, "Highly sparse kernel spectral clustering with predictive out-of-sample extensions," in *Proc. of the 18th European Symposium on Artificial Neural Networks (ESANN 2010)*, 2010, pp. 235–240.
[13] A. Ben-Israel and C. Iyigun, "Probabilistic d-clustering," *J. Classif.*, vol. 25, no. 1, pp. 5–26, Jun. 2008.