An Online Feature Selection Methodology for Ball-Bearing Harmonic Frequencies Based on HMMs

Carlos Puerto-Santana[1,2](✉), Pedro Larrañaga[2], Javier Diaz-Rozo[1], and Concha Bielza[2]

[1] Aingura IIoT, San Sebastian, Spain
epuerto@ainguraiiot.com
[2] Universidad Politécnica de Madrid, Madrid, Spain

Abstract. Much attention has been given to supervised feature subset selection methodologies in data streams. However, less attention has been given to data streams produced by sensors in industrial environments, where labels are difficult to obtain. Feature subset selection is critical in online analysis since it can accelerate and improve the performance of any model inference and reduce data storage issues especially when no cloud is available. In this work we propose an online feature subset selection methodology based on hidden Markov models (HMM) for unsupervised data streams of ball-bearings in order to determine which fundamental and harmonic frequencies are relevant during operation. A validation of the proposed methodology is done with synthetic data and ball-bearing real data in a controlled data stream ambient.

Keywords: Hidden Markov models · Feature selection · Ball-bearings · Frequency analysis · Data stream

1 Introduction

In recent years, the advances in electronics, data processing and storage have enabled industries to perform continuous surveillance of their assets using sensors. In the data processing phase, several mechanical, thermal, electrical and other type of variables are measured. However, some of them may not be always relevant for the underlying dynamic process; therefore, it is crucial to have a model or algorithm capable of determining the relevant variables depending on the current state of the asset. As it will be explained below, ball-bearings can be characterized by their fundamental frequencies able to describe their behavior [1]. Thus, a straightforward question arises: which frequencies and harmonics are relevant and when do they begin to be important?

In spite that several works can be found in feature subset selection (FSS) in data streams in supervised problems, many fewer appear in unsupervised problems [2]. As instance, [3] developed an FSS methodology for data streams based

H. Sanjurjo González et al. (Eds.): SOCO 2021, AISC 1401, pp. 546–555, 2022.
https://doi.org/10.1007/978-3-030-87869-6_52

on matrix sketching[1] from the up-to-time data, and a ridge regression[2]. The coefficients of the regression are used as a relevancy score. [4] assumed that several datasets could arrive simultaneously and the goal was to determine jointly for each dataset their relevant features. A penalized non-negative matrix factorization was carried out for each dataset coming from each view to extract the feature relevancy vectors. [5] proposed a dynamic FSS algorithm that could be used together with any model based clustering. Their idea was to perform a cluster feature selection once a buffer of data was filled. The selected features were used to update a relevancy score vector which indicated the pertinent features. The features which overpassed a threshold were used in the cluster model. In all the mentioned methodologies, the FSS is performed whenever a new instance or chunk of data arrives.

Feature saliency methodologies can be viewed as embedded FSS techniques capable of determining the relevancy of certain variables during the learning phase of a clustering model. In particular, [6,7] and [8] developed variants of hidden Markov model (HMM), where a set of feature saliencies were used to determine which variables were relevant to describe the data. In [6] and [8], a variational Bayesian method was used to maximize the log-likelihood of the model and learn the parameters, whereas in [7] a maximum a posteriori approach was used to learn the model parameters. The previous models were developed only for offline analysis.

In this work the unsupervised model of [7], that we denote as FS-HMM, is applied as the cornerstone for a data stream unsupervised FSS methodology with an application to ball-bearings surveillance. This model is chosen since it has a simple formulation, interpretation and it is easy to implement. In this application, ball-bearing frequencies are recorded and the goal is to determine dynamically their relevancy. As it will be shown, the proposed methodology updates the relevant features when needed and not whenever a new instance or chunk of data arrives as in the previous reviewed articles. The article is structured as follows: Sect. 2 introduces the main concepts of hidden Markov models to understand the methodology. Section 3 explains the proposed methodology. Later, Sect. 4 describes the synthetic and real data used for validation. Then, Sect. 5 shows the obtained results from the real and synthetic data. Finally, Sect. 6 rounds off the article with the relevant conclusions.

2 Theoretical Framework

2.1 Hidden Markov Models

An HMM can be seen as a double-chain stochastic model, where one chain is observed, namely $\boldsymbol{X}^{0:T} = (\boldsymbol{X}^0, ..., \boldsymbol{X}^T)$, where $\boldsymbol{X}^t = (X_1^t, ..., X_M^t) \in \mathcal{R}^M$ and the other chain is hidden, namely $\boldsymbol{Q}^{0:T} = (Q^0, ..., Q^T)$. Here, $T + 1$ is the length

[1] Using singular value decomposition (SVD), a lower rank matrix is extracted from a matrix.

[2] The regression target is the right matrix of the SVD of the sketch matrix.

of the data. It is assumed that the range R of the hidden variable is finite, i.e., $R(Q^t) = \{1, 2, ..., N\}$ for $t = 0, 1, ..., T$. The usual approach for HMMs [9] is to assume that the hidden process has the first-order Markovian property, that is, $P(Q^t|\boldsymbol{Q}^{0:t-1}) = P(Q^t|Q^{t-1})$. It is assumed that the observable process depends on the hidden process, more specifically $b_{Q^t}(\boldsymbol{X}^t) := P(\boldsymbol{X}^t|\boldsymbol{X}^{0:t-1}, \boldsymbol{Q}^{0:t}) = P(\boldsymbol{X}^t|Q^t)$.

All the previous HMM specifications can be summarized with the parameter $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}) \in \Omega$, where Ω denotes the space of all possible parameters, $\mathbf{A}$ is a matrix representing the transition probabilities between hidden states, $\mathbf{B}$ is a vector representing the emission probability of the observations given the hidden state and $\boldsymbol{\pi}$ is the initial probability distribution of the hidden states.

2.2 Feature Relevancy Model

The idea behind any feature saliency model is to declare a set of binary variables, say $\{Z_m\}_{m=1}^{M}$ which will indicate the feature relevancy. Each Z_m variable follows a Bernoulli distribution with a parameter ρ_m which is called the feature saliency of variable X_m. If $\rho_m = 1$, it implies that the feature is always relevant, whereas if $\rho_m = 0$, the variable is never relevant. These variables are added into the model to determine if an observed variable is relevant or not. For example, in the case of HMMs [7], the feature saliency parameters are added to the emission probabilities: $b_{Q^t}(\boldsymbol{X}^t) = \prod_{m=1}^{M} \rho_m \mathcal{N}(X_m^t|\mu_{Q^t m}, \sigma^2_{Q^t m}) + (1-\rho_m)\mathcal{N}(X_m^t|\epsilon_m, \tau_m^2)$, where $\mathcal{N}$ is the probability density function (pdf) of a normal distribution with mean μ and variance σ^2. Observe that when $\rho_m = 1$, the pdf used for variable X_m depends on the hidden state Q^t. However, if $\rho_m = 0$ the pdf does not depend on the hidden state Q^t and variable X_m is considered as drifted Gaussian noise with mean ϵ and variance τ^2. As $\rho_m \in [0, 1]$, it is plausible that some variables relevancies lie in gray levels. In this article we assume that a feature is relevant if $\rho > 0.9$. Further details on optimization of parameters and inference for these type of models can be found in [7].

2.3 Novel Trend Detection

Since the aim of this article is to work with data flows, it is expected that the underlying parameters that drive the data evolve with time. Given a model, when a change in the relevancy of the underlying parameters has occurred it is not expected that the model will predict and explain correctly the new data. These changes in the intrinsic data parameters are called a gradual or abrupt 'concept drift'. A novel concept drift can appear gradually or rapidly in data. Any model which tries to describe and predict the data stream must be capable to adapt when necessary. Traditionally, three strategies had been used to determine the appearance of novel concept drifts. The first one consists of control charts, as in [10]. The second one of sequential analysis, as in [11]. The third one are hypotheses tests about distribution similarities, as in [12]. In practice, sequential tests are deemed to be robust and fast enough to be used in data stream environments.

The use of sequential tests can be used with Chernoff bounds to detect concept drifts with more confidence [13]. The Chernoff bounds are used to give a threshold in the amount of drifting data in the sequential test before a novel concept drift is detected. Once a new concept drift is found, the current model is updated to fit the new trends in the data. The updating process can vary depending on the application. However, in this work, the updating process consists of retraining the model with the up-to-time data. This is necessary since the model requires to discriminate between Gaussian noise and time-changing variables. If there are memory limitations, a large-size buffer can store the data and retrain the model when a novel concept drift is detected. However, in this work we assume that there are no memory problems.

2.4 Ball-Bearing Fundamental Frequencies

Ball-bearings are an important component inside several mechanical-tool machines. These mechanical components are under force and thermal charges which may degrade their structural integrity and therefore cause failures inside larger mechanical-tool machines. The ball-bearings consist of four parts: inner ring, outer ring, balls and cage train. Depending on the geometry of the ball-bearing, each part can be characterized with a frequential component, say: ball pass frequency outer (BPFO) related to the ball-bearing outer ring, ball pass frequency inner (BPFI) related to the ball-bearing inner ring, ball spin frequency (BSF) related to the ball-bearing balls and fundamental train frequency (FTF) related to the ball-bearing train. Using a spectral feature extraction (SFE) algorithm based on spectral kurtosis and envelope demodulation [14], the previous frequencies and their harmonics are extracted from the ball-bearing acceleration signal and used as inputs for the FSS algorithm.

3 Proposed Methodology

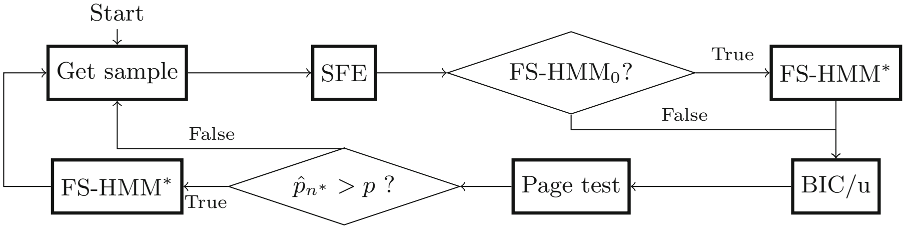


Fig. 1. Flow diagram of the proposed methodology. The idea is to update the model and the relevant features whenever a novel trend is detected.

Here we expose the proposed data-stream procedure to update the FS-HMM, such that the set of relevant features are changed only when needed. Figure 1 shows a flow diagram of the proposed methodology. The first step is to obtain ball-bearing frequential data from sensors using the SFE algorithm (*Get Sample* and *SFE*). Then, the first FS-HMM* is created if no previous FS-HMM model is available (first iteration). Later, the Bayesian information criterion per unit of observation (BIC/u) is computed using the current FS-HMM. The BIC/u is the BIC score [15] divided by the length of the data. These steps are done with the boxes FS-HMM* and BIC/u^t. Whereas, the decision node FS-HMM$_0$? asks whether there exists a FS-HMM or not. Next, the BIC/u scores are used in the Page sequential test to indicate whether the current observations are outliers or not. If the percentage of outliers overpasses a percentage of permissible outliers in a window given by the Chernoff bounds, then, a re-training process is performed and the model is updated with the up to time collected data. The Page sequential test is done with the box *Page test* and with the question node $\hat{p}_{n^*} > p$? it is determined if updating the model is required or not. The updating process is done in the box FS-HMM*. Finally, the methodology goes back to the captured data to continue the process of model learning and updating.

4 Experimental Set up

4.1 Synthetic Data Description

It is assumed that there are eight variables which may change their dynamical behavior over time. The goal is to determine when a variable is totally irrelevant or it changes to be relevant. Also, it is assumed that the process is described by normal Gaussian distributions $\mathcal{N}(\mu, \sigma^2)$. The parameters used for this experiment are provided in Table 1.

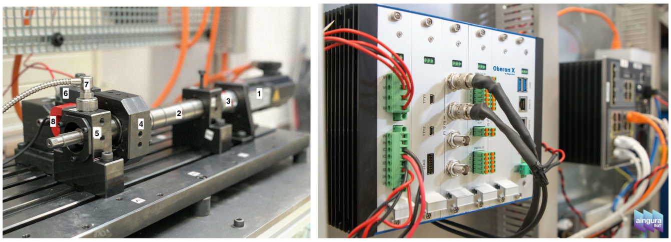
Table 1. Parameters used to generate the synthetic data

X_1			X_2			X_3			X_4		
μ	σ	t	μ	σ	t	μ	σ	t	μ	σ	t
–3	1.5	[0, 1500]	–2	2.5	[0, 1500]	2.3	2.1	[0, 1500]	0.0	1.2	[0, 3000]
4	1.2	(1500, 3000]	6	1.2	(1500, 3000]	2.8	1.1	(1500, 8000]	–2.6	2.5	(3000, 6000]
10	2.1	(3000, 6000]	4	1.2	(3000, 6000]				3.1	0.8	(6000, 8000]
–3	1.5	(6000, 8000]	–2	2.5	(6000, 8000]						
X_5			X_6			X_7			X_8		
μ	σ	t	μ	σ	t	μ	σ	t	μ	σ	t
–3.2	1.5	[0, 8000]	2.7	2.5	[0, 8000]	–5.3	1.4	[0, 8000]	1.6	1.2	[0, 1500]
									7.8	1.0	(1500, 3000]
									–6.5	2.0	(3000, 6000]
									1.6	1.2	(6000, 8000]

The parameters are chosen such that variables X_1, X_2 and X_8 are relevant variables. On the other hand, variables X_5, X_6 and X_7 are irrelevant since they are Gaussian noise. Variable X_4 is at first sight irrelevant but becomes relevant with the progression of periods. Variable X_3 becomes irrelevant over time.

4.2 Real Data Description

The purpose of this ball-bearing testing set up is to monitor vibrations over time during ball-bearing useful life. The experimental set up is shown in Fig. 2 (a). This testbed has a Bosch IndraDyn MS2N synchronous servomotor $\boxed{1}$ that guarantees the required speed during testing, a shaft $\boxed{2}$ with different ball-bearing clamping positions and an elastic coupling $\boxed{3}$ to the servomotor. The testbed has three ball-bearing supports, one for the axial force actuator $\boxed{4}$ and two for support. The system is designed to affect the outside ball-bearing $\boxed{5}$ useful life when axial force is applied. A screw-based force actuator system together with a load sensor $\boxed{6}$. An IMI 607A61 accelerometer $\boxed{7}$ is used to measure vibrations and a thermocouple $\boxed{8}$ for monitoring temperature.



(a) (b)

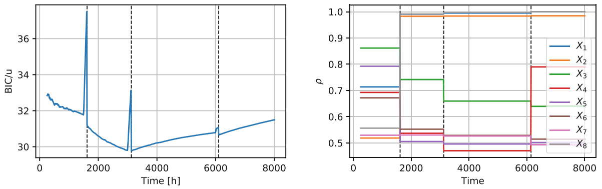
Fig. 2. Experimental testbed: (a) Ball-bearing testbed details, (b) Aingura Insights computing module

Figure 2(b) shows the data acquisition and pre-processing computing module called Aingura Insights (CMAI). The CMAI has an embedded system with a Zynq® Ultrascale+$^{\text{TM}}$ MPSoC, four channels for accelerometer readings up to 19.5 kHz. Inside the CMAI, the SFE procedure is performed and the fundamental frequency amplitudes and up to four harmonics are computed and stored (20 variables). An Eco 6004-2RS ball-bearing is tested at a radial force of 2.41 kN at 3180 RPM (53 Hz). The ball-bearing runs during T = 400 h or equivalently 2.5 times its theoretical operation life.

5 Results

The number of hidden states of the FS-HMM must be fixed beforehand. For synthetic data, from the construction of the data, the number of hidden states is three. In the case of real data this is unknown and models with three, four and five hidden states were tried. However, the best temporal BIC/u was obtained using three hidden states, i.e., lower maximum and minimum BIC/u scores were attained with three hidden states. Additionally, the length of the first window for training data was 256; later, the window was increased by 10 data points. From the Chernoff bounds, the window to determine anomalies was of size $n^* = 87$ and $p = 10\%$. These parameters were selected after a sensibility study.

5.1 Results from Synthetic Data



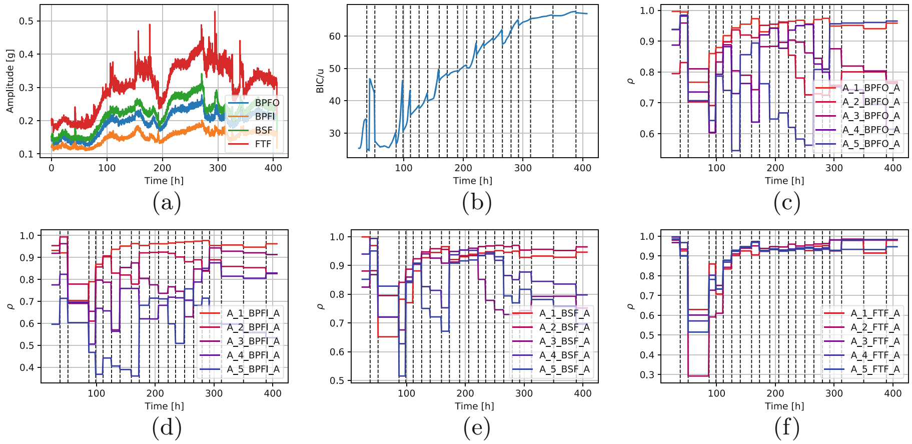
(a) (b)

Fig. 3. Evolution of the relevancy level for the different variables. Whenever a dotted vertical line is observed, it implies that an updating model procedure was done.

In Fig. 3 the results obtained from the synthetic data are drawn. Vertical dotted lines are used to indicate novel concept drift detection. In (a) the BIC/u evolution is shown. Note that, after an important growth in BIC/u, the model is re-trained and the BIC/u trend decreases and the model fitness becomes stable. In particular, three updating processes are performed at times close to $t = 1500, 3000, 6000$ related to the variable definitions in Table 1. The delay in the updating process is caused by the Chernoff bounds; however, these Chernoff bounds are necessary to prevent unnecessary model updating due to outliers. In (b) the evolution of relevancies of the variables X_1 to X_8 are drawn. Before the first novel detection, only one feature has a relevancy ρ greater than 0.9, this is reasonable since before $t = 1500$ all the features behaved like noise. After $t = 1500$ variables X_1, X_2 and X_8 have a high relevancy level (close to 1); however variables X_3, X_4, X_5, X_6 and X_7 have a low relevancy value. Nevertheless, the variable X_4 obtains more relevancy with each re-updating process since its dynamical behavior becomes more clear. On the other hand, variables X_5, X_6 and X_7 which were built as noise, keep a low level of relevancy after all the re-updating processes. From this experiment, it can be determined that the model can discriminate between noise and relevant features dynamically, and the model is updated only when needed.

5.2 Results from Real Data

Figure 4(a) shows the evolution of the amplitude of the fundamental frequencies. Note that, the FTF amplitude is the frequency which shows the greater values and changes. In comparison, the BPFI and BPFO amplitudes show the lowest values and the dynamical changes are less evident.



(a) (b) (c)

(d) (e) (f)

Fig. 4. (a) recorded fundamental frequencies amplitudes. (b) evolution of BIC/u. (c), (d), (e) and (f) show the progressions of the relevancy levels for different features. Dotted vertical lines imply that an updating model procedure was done.

Observe in Fig. 4(b) that the BIC/u score is always going up in spite of the several novel concept drifts detected. However, at the end of the process, the data is stable and fewer re-updates were needed.

We write $A\_i\_f\_A$ as the $(i-1)$-harmonic of the f frequency (being the 0-harmonic the fundamental f frequency). (c) corresponds to the progression of the relevancy for the BPFO and its harmonics. It is noticeable that the level of relevancy of the fundamental frequency is high for the whole process, whereas its harmonics can be less relevant. In particular, the fourth harmonic shows the greatest decay in relevancy to become irrelevant. As for the BPFO case, the fundamental frequency for BPFI in (d) is always relevant, whereas its harmonics are less relevant. In particular, the first harmonic showed low to intermediate levels of relevancy, but at later periods it increased its relevancy. In the case of BSF in (e), the fundamental frequency and its first harmonic are the most relevant features. It is also remarkable that the remaining harmonics have times of high or low relevancy. In (f), the harmonics of FTF sometimes are more relevant than the fundamental frequency during the ball-bearing evolution.

Since the FTF was the fundamental frequency with the more evident dynamic changes, a deeper analysis of this feature is carried out. In Fig. 5(a) the FTF amplitude progression is segmented by the most relevant harmonic. In particular, during the first important change at time $t = 60$, the fundamental frequency is the most relevant feature; however, after that, it does not appear again. On the other hand, the first harmonic is the most relevant feature for a large time period at $t \in [180, 360]$. In (b), the evolution of the least relevant harmonic is shown; in particular, the first harmonic, is the least relevant at the early stages of the data stream to later become the most relevant harmonic. Also, it is remarkable that there are times where the fundamental frequency is the

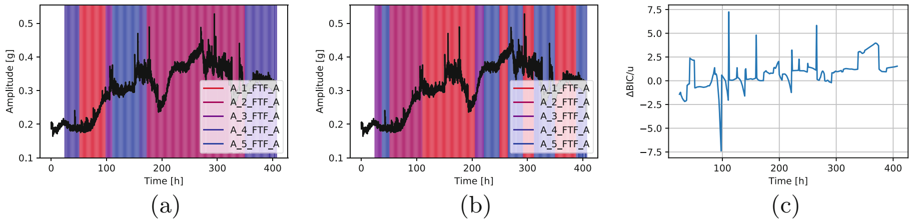


Fig. 5. FTF amplitude segmented by (a) the most relevant feature and (b) the least relevant feature. (c) shows the temporal difference ΔBIC/u

least relevant; which implies that information would be lost if the harmonics are omitted from the analysis. Therefore, some harmonics play a relevant role in describing the dynamical data and the role changes over time. Accordingly, monitoring frequency/harmonics is of crucial interest.

Finally, in (c) we compute $\Delta\text{BIC/u} = (\text{BIC/u})_{\text{ï}} - (\text{BIC/u})_{\text{FS}}$ where $(\text{BIC/u})_{\text{FS}}$ is the BIC/u by using a FS-HMM model and $(\text{BIC/u})_{\text{ï}}$ using a naïve-HMM [16]. The mean of (c) is $\overline{\Delta\text{BIC/u}} = 0.74 > 0$, i.e., $\overline{(\text{BIC/u})_{\text{ï}}} > \overline{(\text{BIC/u})_{\text{FS}}}$ which suggests that for this study, the use of FS-HMM improves the performance.

6 Conclusion

The approach presented in this paper showcases a key enabler towards Industrial Internet of Things/Industry 4.0, where communications infrastructure and deployed machine learning applications can benefit in terms of reduction of data traffic and improved performance respectively due to the feature subset selection. Specifically, this paper adapts an offline unsupervised machine learning model to an online dynamic feature subset selection methodology based on novel concept drift detection in data stream. To demonstrate its applicability within real-world scenarios, the approach has been used to determine relevant frequency amplitudes of a ball-bearing. The methodology was capable of changing the subset of relevant features when needed in both synthetic and real data instead of computing the relevancy whenever a new instance or chunk of data arrived as it is done in previous methodologies [3–5]. Moreover, the algorithm could capture the evolution of relevancy for fundamental frequencies and their harmonics. For real data and depending on the ball-bearing part, for its corresponding frequency amplitude, some harmonics may be more or equally relevant as the fundamental frequency; also, this relevancy could change over time.

Acknowledgements. This work has been partially supported by the Spanish Ministry of Science and Innovation through the PID2019-109247GB-I00 and DSTREAMS project and by the H2020 IoTwins projects (Distributed digital twins for industrial SMEs: A big-data platform) funded by the EU under the call ICT-11-2018-2019, Grant Agreement No. 857191.

References

1. Jáuregui-Correa, J.C., Lozano-Guzman, A.: Mechanical Vibrations and Condition Monitoring. Academic Press, Cambridge (2020)
2. Villa-Blanco, C., Bielza, C., Larrañaga, P.: Feature subset selection for data and feature streams: a review. In: Submitted to: Artificial Intelligence Review Manuscript, pp. 1–50 (2021)
3. Huang, H., Yoo, S., Kasiviswanathan, S.P.: Unsupervised feature selection on data streams. In: CIKM '15, pp. 1031–1040. Association for Computing Machinery (2015)
4. Shao, W., He, L., Lu, C., Wei, X., Yu, P.S.: Online unsupervised multi-view feature selection. In: 2016 IEEE 16th International Conference on Data Mining, pp. 1203–1208 (2016)
5. Fahy, C., Yang, S.: Dynamic feature selection for clustering high dimensional data streams. IEEE Access **7**(127), 128–127, 140 (2019)
6. Zhu, H., He, Z., Leung, H.: Simultaneous feature and model selection for continuous hidden Markov models. IEEE Signal Process. Lett. **19**(5), 279–282 (2012)
7. Adams, S., Beling, P.A., Cogill, R.: Feature selection for hidden Markov models and hidden semi-Markov models. IEEE Access **4**, 1642–1657 (2016)
8. Zheng, Y., Jeon, B., Sun, L., Zhang, J., Zhang, H.: Student's t-hidden Markov model for unsupervised learning using localized feature selection. IEEE Trans. Circ. Syst. Video Technol **28**(10), 2586–2598 (2018)
9. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. In: Readings in Speech Recognition, pp. 267–296. Morgan Kaufmann (1990)
10. Lowry, C.A., Montgomery, D.C.: A review of multivariate control charts. IIE Trans. **27**(6), 800–810 (1995)
11. Page, E.S.: Continuous inspection schemes. Biometrika **41**(1/2), 100–115 (1954)
12. Kifer, D., Ben-David, S., Gehrke, J.: Detecting change in data streams. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases, pp. 180–191. Morgan Kaufmann (2004)
13. Diaz-Rozo, J., Bielza, C., Larrañaga, P.: Machine-tool condition monitoring with Gaussian mixture models-based dynamic probabilistic clustering. Eng. Appl. Artif. Intell **89**, 103434 (2020)
14. Wang, Y., Liang, M.: An adaptive SK technique and its application for fault detection of rolling element bearings. Mech. Syst. Signal Process. **25**, 1750–1764 (2010)
15. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
16. Martinez, M., Sucar, L.: Learning dynamic naive Bayesian classifiers. In: Proceedings of the 21th International Florida Artificial Intelligence Research Society Conference, pp. 655–659 (2008)