Paper Break Prediction Based on Classification in Multivariate Time Series

Shenghan Guo and Changxi Wang

Department of Industrial and Systems Engineering, Rutgers University,

Piscataway, NJ 08854, USA

Abstract

Multivariate time series (MTS) classification has been a challenging problem in recent years. MTS arise when multiple interconnected sensors record observations over time. They are commonly found in manufacturing processes that have several interconnected sensors collecting the data over time. The challenge lies in the following two aspects: First, there are correlations between different features. Second, there is large amounts of irrelevant data and noise. To address the problem, the second derivatives of the predictors are added to the original raw features to predict sudden changes in response. Two alternative procedures based on bag-of-words representation are proposed. The first procedure further selects the features by using chi-square test and then applies logistic regression to the selected features. The second procedure adopts SAX-VSM algorithm that classify time series directly without the feature selection step. A comparison between the two is conducted using the "paper manufacturing" data set. Experimental results from a pulp-and-paper mill show that our proposed methods can achieve effective early classification on MTS.

Keywords: Bags-Of-Patterns, Symbolic Aggregate Approximation, logistic regression, multivariate time series classification

I. Introduction

Multivariate time series (MTS) arise when multiple interconnected streams of data are recorded over time. They are widely used in many areas such as speech recognition, anomaly detection of EFG/ECG signal, smart homes (Gulisano *et al.*, 2016), machine surveillance (Mutschler *et al.*, 2013), energy forecasting (smart grids) (Hobbs *et al.*, 1999, Lew *et al.*, 2011), position tracking and others. MTS are challenging to deal with because an MTS sample contains multiple observations in a single time instant. MTS classification is an important problem because of its multiple features, different modalities of different sensors, the interplay between multiple features and others. To address the problem a wide variety of models are developed (Hüsken *et al.*, 2003, Kim, 2013, Li *et al.*, 2013). For example, pattern mining (Batal *et al.*, 2012, Batal *et al.*, 2009, Kadous *et al.*, 2005), classification methods (Chandrakala *et al.*, 2010, Nguyen *et al.*, 2011, Orsenigo *et al.*, 2010), similarity measurements (Chen *et al.*, 2013, Yang *et al.*, 2005, Yoon *et al.*, 2005). Meanwhile, the early detection of MTS attracts lots of attention. For example, 1-nearest neighbor (1NN) classification is proposed as an efficient approach to make early prediction (Xing *et al.*, 2009).

However, the early classification of MTS is still an open but useful problem. For example, analyzing the MTS generated by sensors monitoring the pulp-and-paper process may identify the abnormalities as early as possible and could offer workers an emergency alarm before paper break happens. Another example is that analyzing the MTS generated by patient monitoring and identifying its anomalies could offer doctors an emergency alarm. Up to the moment, the research on early classification on MTS data is sparse except (Ghalwash *et al.*, 2012). Ghalwash *et al.* (2012) proposed a multivariate shapelet, which is consisted of multiple segments. All the segments are extracted in the same sliding time window at the same time. However, as this model have different intervals for each variable, the shapelets are incapable of including distinctive patterns of all variables unless their lengths are sufficiently long. If a shapelet is too long, it can hardly classify the data as soon as possible.

In this report, we provide two approaches for MTS classification based on Bag-Of-Patterns. The selected features are then used to classify the time series with logistic regression.

II. Methodology

In this section, we describe two alternative approaches for MTS classification. Both of them are based on Bag-Of-Patterns (BOP) representation. Bags of words are generated using Symbolic Aggregate Approximation (SAX) (Schäfer *et al.*, 2017). Taking the bags of words as input, a chi-squared test is conducted for feature selection. Then we use logistic regression to classify the time series (Schäfer *et al.*, 2012); the second approach adopts the SAX-VSM algorithm (Schäfer *et al.*, 2017) that performs a weighting scheme based on term frequency (tf) and inverse document frequency (idf).

2.1 BOP by SAX

BOP extracts subsequences of a time series and discretizes these real-valued subsequences into a word, which is a string of symbols over a predefined alphabet (Schäfer *et al.*, 2017). Classification models can be built upon BOP representation by constructing a feature vector using word counts and then apply classifiers on the selected features (Schäfer *et al.*, 2017). To transform time series into bags of words, two major discretization functions are frequently used in recent studies, i.e. SAX (Senin *et al.*, 2013) and Symbolic Fourier Approximation (SFA) (Schäfer *et al.*, 2012). The difference between the two is that SAX relies on the discretization of means while SFA is based on the discretization of coefficients of the Fourier transforms. SFA is considered to have better data adaptivity over SAX (Schäfer *et al.*, 2017). In this project, we adopt SAX for its wider adoption (Ashouri *et al.*, 2018, Georgoulas *et al.*, 2015, Sun *et al.*, 2014, Yin *et al.*, 2019, Zhang *et al.*, 2019) and more straightforward implementation.

To apply SAX on real data, we first define L as the size of a sliding window that extracts subsequences of time series. Overlapping windows are adopted here, i.e. the window is moved forward for one time point every time. We further define a word size, w, and an alphabet size, a, that configure the SAX function (Senin *et al.*, 2013). For MTS, each univariate time series it contains is put into SAX and converted to words. Suppose the MTS is denoted as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n], \boldsymbol{x}_i = [x_{1i}, x_{2i}, \dots, x_{ti}]^T, i = 1, 2, \dots, n.$ For the tth window, observations $x_{t1}, x_{(t+1)1}, \dots, x_{(t+L-1)1}$ are extracted from univariate time series, $\boldsymbol{x}_i$, that is converted into a string of word with w letters. For example, "acc" if $w = 3$. The SAX calculation is repeated for each univariate time series. The obtained BOP representation for each univariate time series are then grouped together, so for each window we will have an observation vector of words (each word has w letters) with length n, denoted as $\boldsymbol{Z}_t = [z_{t1}, z_{t2}, \dots, z_{tn}]$. By sliding the window forward, a design matrix of words is constructed for classification. To match the design matrix with response, $\boldsymbol{Y} = [y_1, y_2, \dots, y_t, \dots]^T$, we connect $\boldsymbol{Z}_{t-1}$ with y_t so we have one-observation-ahead prediction.

2.2 Classification

After transforming time series subsequences into BOP representation, we proceed to classification. We propose two different types of classifiers, i.e. logistic regression (Sec. 2.2.1) and VSM algorithm (Senin *et al.*, 2013) (Sec. 2.2.2). Both classifiers intake the same bags of words from SAX computation.

2.2.1 Logistic Regression

Inspired by the work of Schäfer and Leser (2019) (Schäfer *et al.*, 2017), we adopt a logistic regression scheme for MTS classification. This procedure exactly follows their WEASEL+MUSE method, except that the SFA function is replaced by SAX. After BOP representation, we have words instead of numerical values for each predictor. These words now become the feature. A Chi-square Test is conducted on these words for feature selection. We group all the words appeared in the design matrix into bag B, all the words associated with an observed negative response (class 0) into bag B_0, and those associated with observed positive response (class 1) into bag B_1. This categorization of words enables us to identify the required information in chi-square score calculation for each feature: the total number of positive instances containing the feature (A), the total number of instances containing the feature (M), the total number of positive instance (P) and the total number of instance (N). The chi-square score is then calculated as follows (Meesad *et al.*, 2011):

$$\chi^2 = \frac{N(AN - MP)^2}{PM(N-P)(N-M)} \tag{1}$$

As a statistical test for independence, Chi-square Test would accept the null hypothesis that two events are independent when the chi-square score is small. In other words, the larger χ^2 is, the more dependent the two events are, i.e. a stronger correlation between the feature and the response. Therefore, we calculate the chi-square score for each feature and select those with the largest scores for the following logistic classification. The size of the feature set (the number of words it contains), S, is a user-defined value.

To do logistic regression, we build a feature vector in the form of a histogram by counting the appearance of each (selected) feature in an observation, which further transform the design matrix of words into a matrix of discrete, counting values. Then we apply the conventional logistic regression technique on the design matrix and the response for time series classification.

2.2.2 SAX-VSM

An alternative method for classification is SAX-VSM introduced by Senin and Malinchik (2013) (Senin *et al.*, 2013). The BOP representation from SAX is fed into a weighting scheme that assign a *tf*idf* weight for term t, which is the product of term frequency (*tf*) and inverse document frequency (*idf*). By following the SAX-VSM scheme, feature selection is avoided. Words in B are all taken as features. A weight vector covering all its words will be assigned to bag B_0 and B_1, respectively, based on each word's frequency of appearance. Note that words only appear in B that do not show in B_0 (or B_1) is equivalent to having a frequency of 0 in B_0 (or B_1). An unlabeled observation is classified to the class with the larger inner product of *tf*idf* weights and frequency.

III. Results and Discussion

This section provides the results from applying procedures in part II. We will display the results clearly in table and compare the performance with respect to different parameter values/procedures by visualization.

3.1 Data Inspection

Before applying classification procedures, we conduct a primary data inspection to better understand the characteristics of the "paper manufacturing" data set. First, we perform a correlation analysis among the raw variables. Figure 1 is the correlogram for the inter-variable correlation:

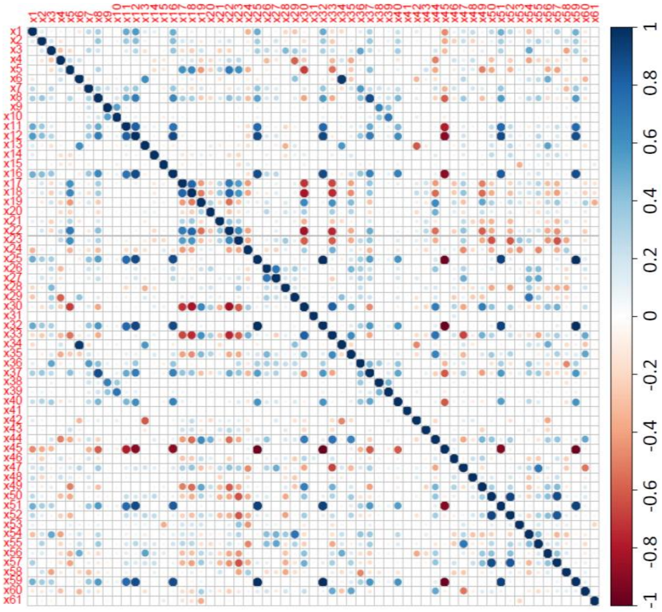


Figure 1. Inter-variable correlogram

The scattered dark-colored dots indicate that correlations widely exist among many pairs of variables, so the variables in this MTS must be analyzed as a whole instead of taken as univariate time series.

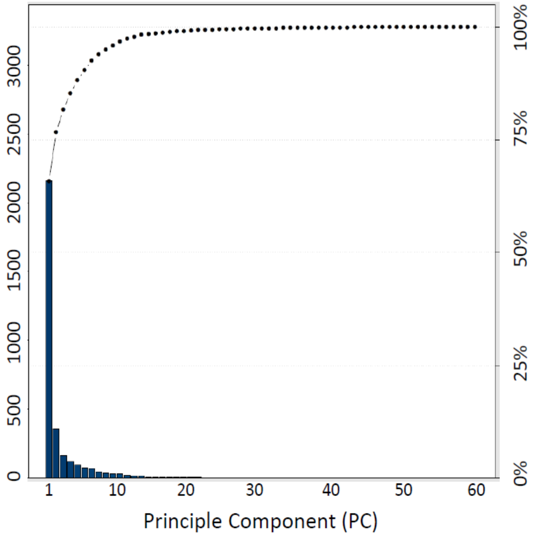


Figure 2. Variance explained by individual PCs

A straightforward approach for correlated data analysis is Principal Component Analysis (PCA). We apply PCA on the entire data and obtain the principal components (PCs), whose portions of variance explained are visualized in Figure 2.

Since PC1 explains more than 60% data variance, we visualize the PC1-transformed "paper manufacturing" data and see the matching between it and the rare event, "paper break" (Figure 3). Albeit powerful, PCA cannot effectively reveal the patterns in data. There is no clear matching established between the PC1-transformed data and the "paper break". It is possible that the time-varying nature of time series complicates the data with a nonstationary variance that is not well modeled by PCA. However, this attempt does give us an implication that the pattern may not be phenomenal and can be easily masked by the noise, providing a clue for the root cause analysis.

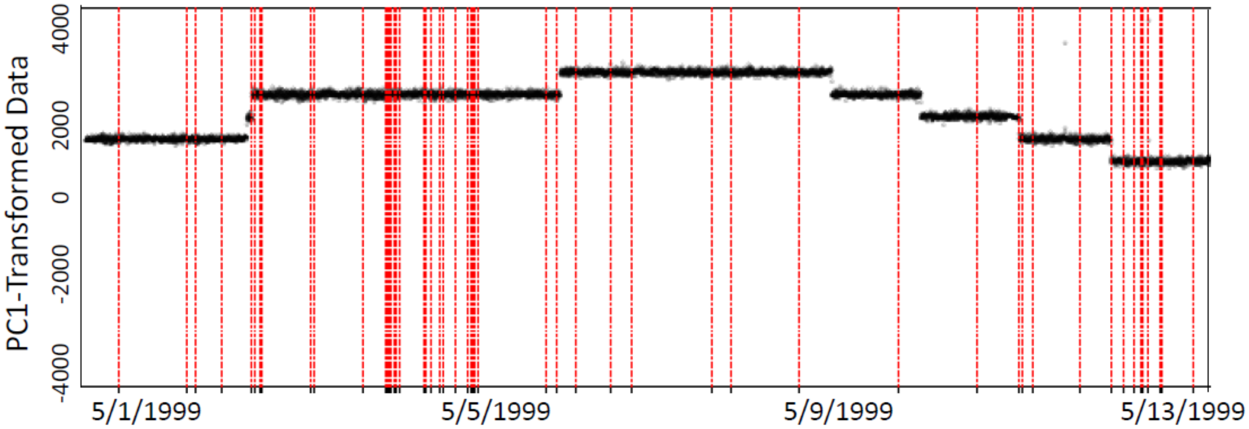


Figure 3. PC1-transformed data visualization. Vertical red dash lines represent "paper break"

3.2 BOP Classification Results

This subsection applies the BOP classification procedures on the data set. We let the window size be $L = [100, 500, 1000]$. For feature selection – logistic regression, we let the size of feature set be $S = [20, 200, 2000]$, the cutoff value of a positive event be 0.5 per the convention; for SAX-VSM, we let the word size be $w = 10$ and the alphabet size be $a = 20$. The first 8000 observations of the “paper manufacturing” data set is reserved as training resources and the rest are used for testing purpose. The BOP representation calculated with SAX is done on both the raw variables and their derivatives, i.e. $x'_{ti} = \left|x_{ti} - x_{(t-1)i}\right|, i = 1, 2, ...$ Hence, we have 122 variables as the input for SAX.

Table 1. Performance metrics and the definition

Performance Metric	**Calculation**
FPR	FP/(FP+TN)
FNR	FN/(FN+TP)
Accuracy	(TP+TN)/(TP+TN+FP+FN)
Recall	TP/(TP+FN)
Precision	TP/(TP+FP)
F1-score	2TP/(2TP+FP+FN)

To measure the classification performance, we adopt a series of metrics. Denote the false positive by FP, false negative by FN, true positive by TP and true negative by TN. Based on the definitions as shown in Table 1, we calculate the performance metrics for the classification results. For real data implementation, it is possible that TP and FN are both zero or TP, FP and FN are all zero, in which case the corresponding metrics would have an “NA” instead of a numerical value. Table 2 shows the performance using feature selection + logistic regression for both training and testing phase:

Table 2. Classification performance of feature selection – logistic regression

		Training			Testing		
Feature set size	**Window size**	**100**	**500**	**1000**	**100**	**500**	**1000**
20	**FPR**	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000
	FNR	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Accuracy	0.9937	0.9935	0.9933	0.9929	0.9932	0.9929
	Recall	NA	NA	NA	0.0000	NA	NA
	Precision	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	F1 score	NA	NA	NA	NA	NA	NA
200	**FPR**	0.0001	0.0001	0.0026	0.0062	0.0303	0.0519
	FNR	0.9000	0.9796	0.9149	0.9859	0.9552	0.9697
	Accuracy	0.9942	0.9935	0.9913	0.9870	0.9634	0.9416
	Recall	0.8333	0.5000	0.1818	0.0156	0.0101	0.0042
	Precision	0.1000	0.0204	0.0851	0.0141	0.0448	0.0303
	F1 score	0.1786	0.0392	0.1159	0.0148	0.0164	0.0073
2000	**FPR**	0.0000	0.0003	0.0000	0.0325	0.3868	0.1663
	FNR	0.0000	0.0000	0.2128	0.9718	0.5821	0.8333
	Accuracy	1.0000	0.9997	0.9986	0.9610	0.6119	0.8290
	Recall	1.0000	0.9608	1.0000	0.0060	0.0074	0.0071
	Precision	1.0000	1.0000	0.7872	0.0282	0.4179	0.1667
	F1 score	1.0000	0.9800	0.8810	0.0100	0.0145	0.0136

The values shown in Table 2 seem to have a trend with respect to the parameter values, which will be discussed in detail in Sec. 3.2. We now show the metrics for SAX-VSM procedure in Table 3:

Table 3. Classification performance of SAX-VSM

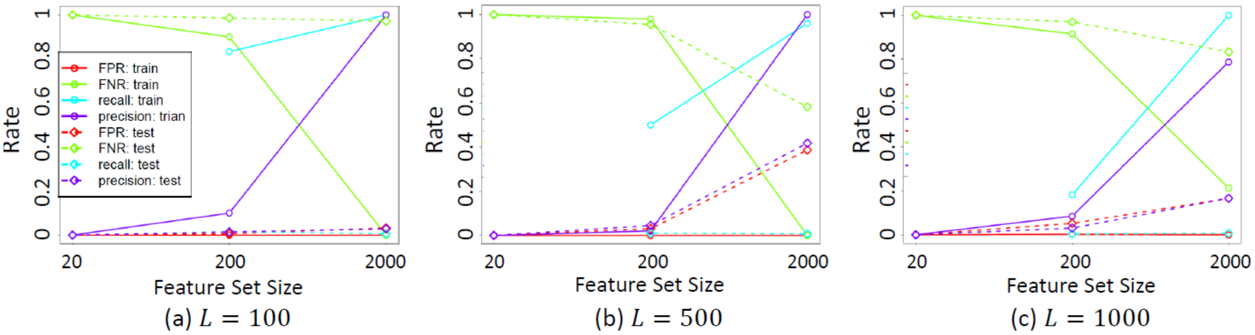
	Training			Testing		
Window size	**100**	**500**	**1000**	**100**	**500**	**1000**
FPR	0.0089	0.0082	0.0972	0	0	0
FNR	0.9200	1.0000	0.8085	1	1	1
Accuracy	0.9853	0.9853	0.8980	0.993039	0.993163	0.992902
Recall	0.0541	0.0000	0.0131	NA	NA	NA
Precision	0.0800	0.0000	0.1915	0	0	0
F1-score	0.0645	NA	0.0246	NA	NA	NA

More discussion of the results will be provided shortly in the next section.

3.2 Discussion

A straightforward comparison between Table 2 and 3 reveals that feature selection plus logistic regression gives a better performance than the SAX-VSM procedure with the selected parameter values. In Table 2, with the increasing of feature set size, both training and testing phase obtain improved performance in general that dominate the metric values in Table 3. Therefore, our method-wise comparison suggest that the procedure based on logistic regression has a better fitness for the "paper manufacturing" data set.

Let's take a further look into Table 2. Figure 4 visualizes the performance metrics against the feature set size, S:

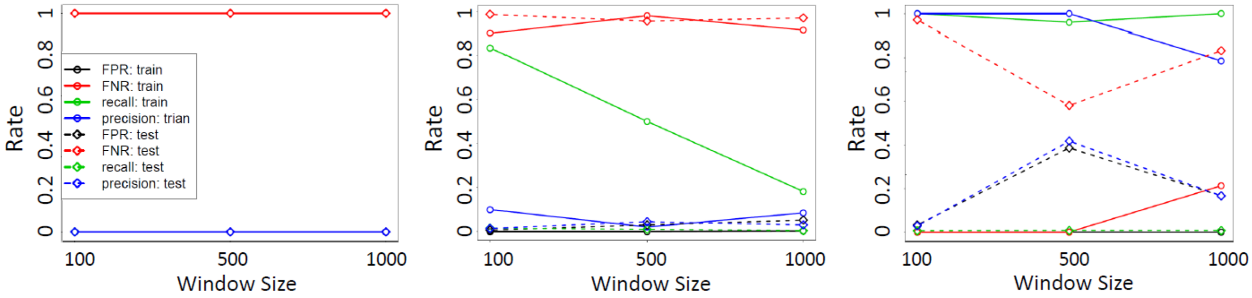


(a) $L = 100$ (b) $L = 500$ (c) $L = 1000$

Figure 2. Visualization of FPR, FNR, recall and precision against feature set size for both training and testing sets

For each window size, the training phase classification performance shows a clear tendency of improvement regarding the feature set size. Starting from a precision of zero and FNR of one, for all three window sizes, the four visualized metrics almost reach their best extreme for the training data and much enhanced for the testing data as the feature set size becomes sufficiently large. Indeed, the testing phase classification has a less desirable performance than the training phase, which may be further improved by using a larger S. The phenomenon revealed by Figure 4 is easily understood, because the largest obtainable size for the feature set is the one before chi-square test feature selection, which preserves all the information from the data and thus the best classification power. Nonetheless, the purpose of feature selection is to reduce computational complexity by decreasing the feature set size, so an extremely large S should not be used. In effect, with the experimented feature set size, the classification performance is already favorable, which demonstrates the superiority of the procedure (Sec. 2.2.1) in addressing rare event prediction for MTS.

Figure 4 reveals the effect of S, but does not show much distinction among the three subplots, so we cannot reach a definite conclusion about the effect of window size. To explore the performance sensitivity against L, we visualize FPR, FNR, recall and precision against the window size, as shown in Figure 5:



(a) $S = 20$ (b) $S = 200$ (c) $S = 2000$

Figure 3. Visualization of FPR, FNR, recall and precision against window size for both training and testing sets

The three subplots in Figure 5 have clear distinctions. From left to right, as the feature set size increases, the performance generally become way better, which is already found from Figure 4. In terms of the window size effect, however, we cannot conclude that a larger window size is absolutely good. In fact, as the window size increase, the recall can be less desirable, as in Figure 5(b); the FNR and precision can also deteriorate, as shown by Figure 5(c). It is likely that an optimal window size will lead to the best performance for the given data, which needs parameter tuning to find out. Due to the time limit, we do not expand on this direction but save it for future work.

We can wrap up this discussion with several discovers. First, the feature selection – logistic regression scheme outperforms the SAX-VSM algorithm for the “paper manufacturing” data. As a brave guess, we may consider the former candidate method is more suitable for rare event prediction for MTS. Second, the feature selection – logistic regression scheme is sensitive to the feature set size. A moderate value for this parameter suffices to guarantee a favorable training phase performance. Enlarging this parameter can improve the testing phase performance also. Third, the scheme is likely to have an optimal window size that can be found from numerical experiments. This optimal window size will lead to a balance among different performance metrics.

IV. Conclusions

In this report, we propose two approaches for MTS classification based on Bag-Of-Patterns. Taking the bags-of-words (BOP) representation as input, in the first approach, a chi-squared test is conducted for feature selection. Then the logistic regression is used to classify the time series. The second approach adopts the SAX-VSM algorithm that performs a weighting scheme based on term frequency (*tf*) and inverse document frequency (*idf*). The results show that feature selection plus logistic regression is suitable for rare event prediction for MTS. We also investigate the influences of window size and feature set size for this scheme. It shows that the method is performance-sensitive to the feature set size. A larger feature set size improves the performance in both training and testing phase classification, yet a moderate feature set size suffices to generate satisfactory results. An optimal value is likely to exist for the window size, which leads to a balance among the examined performance metrics. This optimal value can be found from numerical experiments.

This study has provided a preliminary analysis to the “paper manufacturing” data set, which can be extended to other multivariate time series data. For future work, several directions are worthy further exploration. First, how the training set size impacts the performance is worth intensive study. Our experiment is based on a fixed training/testing set partition. A fixed training set size, 8000, is used throughout our analysis. Yet, it is likely that modifying the training data size will affect the classification performance. Nowadays, manufacturing involves low-volume, highly-customized production frequently that results in limited training resources. Hence, understanding the influence of training set size is of a broad impact. The second future direction is a nature consequence of the first one, which is how to improve the classification performance when there are limited training samples. In such an extreme scenario, transfer learning may be a feasible method that borrows “similar” information from previous study or analysis to current application in order to improve the quality of classification model fitting and subsequently the prediction power.

References

ASHOURI, A., HU, Y., NEWSHAM, G. R. and SHEN, W. Energy Performance Based Anomaly Detection in Non-Residential Buildings Using Symbolic Aggregate Approximation. *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*, 2018. IEEE, 1400-1405.

BATAL, I., FRADKIN, D., HARRISON, J., MOERCHEN, F. and HAUSKRECHT, M. Mining recent temporal patterns for event detection in multivariate time series data. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012. ACM, 280-288.

BATAL, I., SACCHI, L., BELLAZZI, R. and HAUSKRECHT, M. Multivariate time series classification with temporal abstractions. *Twenty-Second International FLAIRS Conference*, 2009.

CHANDRAKALA, S., CHANDRA SEKHAR, C. J. I. J. O. D. M., MODELLING and MANAGEMENT 2010. Classification of varying length multivariate time series using Gaussian mixture models and support vector machines. 2**,** 268-287.

CHEN, Y., HU, B., KEOGH, E. and BATISTA, G. E. DTW-D: time series semi-supervised learning from a single example. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013. ACM, 383-391.

GEORGOULAS, G., KARVELIS, P., LOUTAS, T. and STYLIOS, C. D. 2015. Rolling element bearings diagnostics using the Symbolic Aggregate approXimation. *Mechanical Systems and Signal Processing,* 60**,** 229-242.

GHALWASH, M. F. and OBRADOVIC, Z. J. B. B. 2012. Early classification of multivariate temporal observations by extraction of interpretable shapelets. 13**,** 195.

GULISANO, V., JERZAK, Z., VOULGARIS, S. and ZIEKOW, H. The DEBS 2016 grand challenge. *Proceedings of the 10th ACM International Conference on Distributed and Event-based Systems*, 2016. ACM, 289-292.

HOBBS, B. F., JITPRAPAIKULSARN, S., KONDA, S., CHANKONG, V., LOPARO, K. A. and MARATUKULAM, D. J. J. I. T. O. P. S. 1999. Analysis of the value for unit commitment of improved load forecasts. 14**,** 1342-1348.

HÜSKEN, M. and STAGGE, P. 2003. Recurrent neural networks for time series classification. *Neurocomputing,* 50**,** 223-235.

KADOUS, M. W. and SAMMUT, C. J. M. L. 2005. Classification of multivariate time series and structured data using constructive induction. 58**,** 179-216.

KIM, M. 2013. Semi-supervised learning of hidden conditional random fields for time-series classification. *Neurocomputing,* 119**,** 339-349.

LEW, D., MILLIGAN, M., JORDAN, G. and PIWKO, R. 2011. Value of Wind Power Forecasting. National Renewable Energy Lab.(NREL), Golden, CO (United States).

LI, C., CHIANG, T.-W. and YEH, L.-C. 2013. A novel self-organizing complex neuro-fuzzy approach to the problem of time series forecasting. *Neurocomputing,* 99**,** 467-476.

MEESAD, P., BOONRAWD, P. and NUIPIAN, V. A chi-square-test for word importance differentiation in text classification. *Proceedings of International Conference on Information and Electronics Engineering*, 2011. 110-114.

MUTSCHLER, C., ZIEKOW, H. and JERZAK, Z. The DEBS 2013 grand challenge. *Proceedings of the 7th ACM international conference on Distributed event-based systems*, 2013. ACM, 289-294.

NGUYEN, M. N., LI, X.-L. and NG, S.-K. Positive unlabeled learning for time series classification. *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

ORSENIGO, C. and VERCELLIS, C. 2010. Combining discrete SVM and fixed cardinality warping distances for multivariate time series classification. *Pattern Recognition,* 43**,** 3787-3794.

SCHÄFER, P. and HÖGQVIST, M. SFA: a symbolic fourier approximation and index for similarity search in high dimensional datasets. *Proceedings of the 15th International Conference on Extending Database Technology*, 2012. ACM, 516-527.

SCHÄFER, P. and LESER, U. 2017. Multivariate time series classification with WEASEL+ MUSE. *arXiv preprint arXiv:1711.11343*.

SENIN, P. and MALINCHIK, S. Sax-vsm: Interpretable time series classification using sax and vector space model. *2013 IEEE 13th international conference on data mining*, 2013. IEEE, 1175-1180.

SUN, Y., LI, J., LIU, J., SUN, B. and CHOW, C. 2014. An improvement of symbolic aggregate approximation distance measure for time series. *Neurocomputing,* 138**,** 189-198.

XING, Z., PEI, J. and PHILIP, S. Y. Early prediction on time series: a nearest neighbor approach. *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

YANG, K., YOON, H. and SHAHABI, C. CLeVer: a feature subset selection technique for multivariate time series. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2005. Springer, 516-522.

YIN, J., XU, M. and ZHENG, H. 2019. Fault diagnosis of bearing based on Symbolic Aggregate approXimation and Lempel-Ziv. *Measurement*.

YOON, H., YANG, K. and SHAHABI, C. 2005. Feature subset selection and feature ranking for multivariate time series. *IEEE Transactions on Knowledge and Data Engineering,* 17**,** 1186-1198.

ZHANG, Y., DUAN, L. and DUAN, M. 2019. A new feature extraction approach using improved symbolic aggregate approximation for machinery intelligent diagnosis. *Measurement,* 133**,** 468-478.