Feature expansion of single dimensional time series data for machine learning classification

Daeun Jung, Jungjin Lee and Hyunggon Park
Multiagent Communications and Networking Laboratory
Ewha Womans University, Seoul, Republic of Korea
{daeun.jung, jungjin.lee}@ewhain.net, hyunggon.park@ewha.ac.kr

2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN) | 978-1-7281-6476-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICUFN49451.2021.9528690

***Abstract*—In this paper, we propose a feature expansion approach for the lowest one-dimension (1-D) time series data classification problems, where the expanded features include temporal, frequency, and statistical characteristics. We show that the proposed feature expansion can improve the classification accuracy compared to conventional machine learning algorithms for data classification. This is because the expanded features enable classifiers to consider multiple dimensions which are not feasible for low dimension data. Experiment results show that the proposed feature expansion method can improve the classification performance compared to conventional machine learning algorithms for 1-D actual biosensor data.**

Keywords*—*machine learning; feature expansion; time series data; biosensor data

I. Introduction

The advent of IoT and the development of the healthcare industry have accelerated in processing bio-data for different purposes using machine learning algorithms, where machine learning algorithms have showed the effectiveness and been developed for practical applications. For example, a prediction of aggression in youth with Autism Spectrum Disorder with wearable biosensor data is discussed in [1], where aggression to other people can be predicted one minute before with high accuracy based on a logistic regression algorithm. In clinical data analysis, a hybrid machine learning algorithm is proposed to find important features that may lead to improved accuracy in predicting cardiovascular disease [2].

It is however shown that studies with biosensor data have several limitations [3]. Reliable generalizability has been a problem to achieve in machine learning algorithm design as individual person has its own patterns. While there is an attempt to facilitate an available signal to achieve better performance, an unreliable and noisy signal may hinder the generalization of the algorithms. To overcome this limitation, a statistical feature-based method is presented in the field of network analysis. An illustrative example is in [4], where statistically driven entropy based features for network flow data is proposed to distinguish benign and anomalous flows.

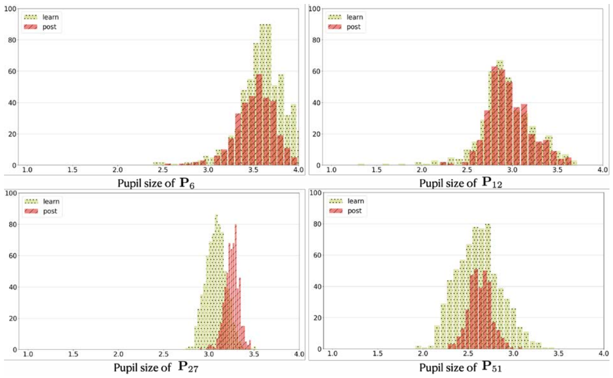

Fig. 1. Histogram of average pupil size for participants $p = 6, 12, 27, 51$

In this paper, our goal is to infer learning states of participants based on their pupil size data. Based on the fact that different human cognition process can lead to the changes of pupil size and complex learning tasks can increase pupil diameter [5], we use the pupil size data to analyze and interpret the cognitive and psychological response of participants. However, it is challenging for inference of the cognitive state as each person has different response and biosensor signal may be corrupted by noise. As shown in Fig. 1, the distribution of pupil sizes shows distinctive patterns based on learning state. Since it is difficult to identify the tendency of each person in each state with a single feature, better performance can be achieved by modifying the time series data. Hence, we propose a *feature expansion* method that transforms the 1-D time series data into a vector of expanded features with statistically driven features, temporal features and frequency features represented by coefficients of transform measures. The transformed multiple features enable machine learning algorithms to adaptively use multiple features, leading to improved classification performance. The performance improvement based on the proposed feature expansion approach is confirmed by the actual pupil size data.

The rest of the paper is structured as follows. In Section II, we formulate the problem with description of data set and the proposed approach feature expansion method that transforms 1-D time series data into an expanded feature vector. In Section III, the experiment design and result are illustrated in detail. Lastly, conclusions and suggestions for future work are presented in Section IV.

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00024, Supervised Agile Machine Learning Techniques for Network Automation based on Network Data Analytics Function, No. 2021-0-00739, Development of Distributed/Cooperative AI based 5G+ Network Data Analytics Functions and Control Technology) and supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(No. NRF-2020R1A2B5B01002528).

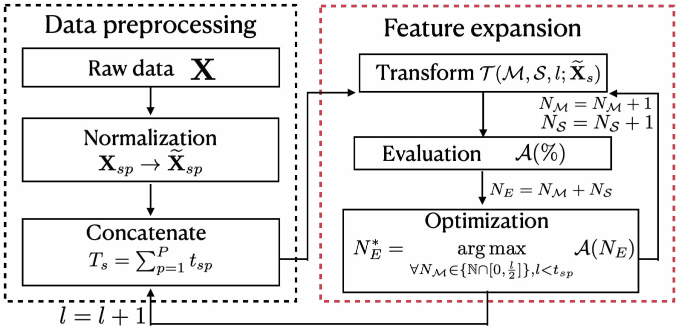
Data preprocessing
Raw data $\mathbf{X}$
Normalization
$\mathbf{X}_{sp} \rightarrow \widetilde{\mathbf{X}}_{sp}$
Concatenate
$T_s = \sum_{p=1}^{P} t_{sp}$
Feature expansion
Transform $\mathcal{T}(\mathcal{M}, \mathcal{S}, l; \widetilde{\mathbf{X}}_s)$
$N_\mathcal{M} = N_\mathcal{M} + 1$
$N_\mathcal{S} = N_\mathcal{S} + 1$
Evaluation $\mathcal{A}(\%)$
$N_E = N_\mathcal{M} + N_\mathcal{S}$
Optimization
$N_E^* = \underset{\forall N_\mathcal{M} \in \{\mathbb{N} \cap [0, \frac{l}{2}]\}, l < t_{sp}}{\arg\max} \mathcal{A}(N_E)$
$l = l + 1$

Fig. 2. A diagram for the proposed feature expansion approach.

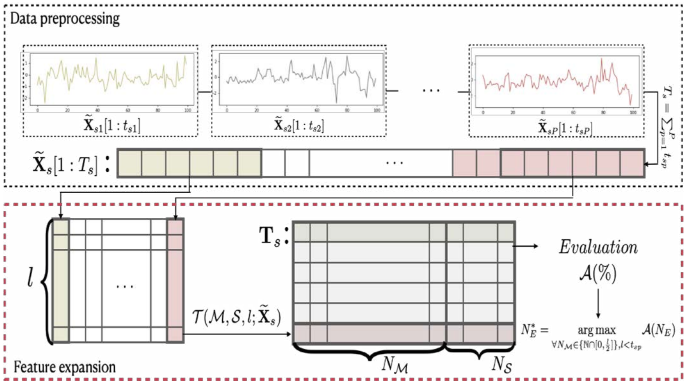
Data preprocessing
$\widetilde{\mathbf{X}}_{s1}[1:t_{s1}]$
$\widetilde{\mathbf{X}}_{s2}[1:t_{s2}]$
$\cdots$
$\widetilde{\mathbf{X}}_{sP}[1:t_{sP}]$
$T_s = \sum_{p=1}^{P} t_{sp}$
$\widetilde{\mathbf{X}}_s[1:T_s]$:
$\cdots$
l
$\cdots$
$\mathcal{T}(\mathcal{M}, \mathcal{S}, l; \widetilde{\mathbf{X}}_s)$
$\mathbf{T}_s$:
Evaluation
$\mathcal{A}(\%)$
$N_E^* = \underset{\forall N_\mathcal{M} \in \{\mathbb{N} \cap [0, \frac{l}{2}]\}, l < t_{sp}}{\arg\max} \mathcal{A}(N_E)$
$N_\mathcal{M}$
$N_\mathcal{S}$
Feature expansion

Fig. 3. The proposed feature expansion algorithm.

II. Problem Setup and Proposed Feature Expansion Approach

In this paper, we consider a classical classification problem, where a state needs to be inferred given a raw data set $\mathbf{X}$ that consists of 1-D vectors $\mathbf{X}_{sp}$ for participant $p \in \mathbf{P} = \{1, 2, \ldots, P\}$ in stats $s \in \mathbf{S} = \{1, 2, \ldots, S\}$ as shown in Fig. 2. The data for participant p at state s with its data length t_{sp} is denoted by $\mathbf{X}_{sp}[1 : t_{sp}]$. For data classification, several machine learning algorithms can be deployed. However, blind adoption of machine learning algorithms for data classification on the raw data $\mathbf{X}$ may result in low performance, in particular, for 1-D data, which also depends on individual participants.

In order to overcome these limitations, we first normalize the data $\mathbf{X}_{sp}$ as

$$\widetilde{\mathbf{X}}_{sp}[1 : t_{sp}] = \{\tilde{x}_p^{(1)}, \cdots, \tilde{x}_p^{(t_{sp})}\}^T$$

which can reduce the variations in the raw data $\mathbf{X}$ caused by individual participants. The normalized data $\widetilde{\mathbf{X}}_{sp}[1 : t_{sp}]$ is further processed as

$$\begin{aligned}\widetilde{\mathbf{X}}_s[1 : T_s] &= \{\tilde{x}_1^{(1)}, \cdots, \tilde{x}_1^{(t_{s1})}, \cdots, \tilde{x}_P^{(1)}, \cdots, \tilde{x}_P^{(t_{sP})}\}^T \\ &= \{\tilde{x}_1^{(1)}, \cdots, \tilde{x}_P^{(T_s)}\}^T\end{aligned}$$

where $T_s = \sum_{p=1}^{P} t_{sp}$ represents the total number of data points in state s. These processes can anonymize the data by removing the dependency in each participant, so that the classification can only be made for the state.

We then convert the normalized and anonymized 1-D data set $\widetilde{\mathbf{X}}_s[1 : T_s]$ into multi-dimensional data based on the feature expansion function $\mathcal{T}$: $\mathbb{R}^{l \times 1} \rightarrow \mathbb{R}^{1 \times N_E}$, defined as

$$\mathcal{T}(\mathcal{M}, \mathcal{S}, l; \mathbf{x}) = \mathbf{t} \tag{1}$$

where l denotes a window size that is a truncation unit for the original 1-D data. The feature expansion function $\mathcal{T}$ maps $\mathbf{x}$ with length l into expanded feature vector $\mathbf{t}$ that consists of N_E number of features containing the coefficients induced from measure $\mathcal{M}$ and statistical method $\mathcal{S}$. By definition in (1), $\widetilde{\mathbf{X}}_s[1 : T_s]$ is transformed to $\mathbf{T}_s$ such that

$$\mathcal{T}(\mathcal{M}, \mathcal{S}, l; \widetilde{\mathbf{X}}_s) = \mathbf{T}_s$$

where $\mathbf{t}_s^{(u)} \in \mathbf{T}_s$ for $u = 1, \cdots, \left\lceil \frac{T_s}{l} \right\rceil$. Each element $\mathbf{t}_s^{(u)} = \{t_1^{(u)}, \cdots, t_{N_E}^{(u)}\}$ is a vector of N_E features.

The proposed feature expansion approach can adopt any measure $\mathcal{M}$ to extract the characteristics in time series data, where the characteristics of the data is represented by the number of induced coefficients $N_\mathcal{M}$ from $\mathcal{M}$. Moreover, statistically driven $N_\mathcal{S}$ features are included in the proposed feature expansion approach. In summary, the feature expansion $\mathcal{T}$ induces $N_E = N_\mathcal{M} + N_\mathcal{S}$ coefficients for $\widetilde{\mathbf{X}}_s[1 : T_s]$. The classification performance with N_E coefficients can be measured by

$$\mathcal{A}(N_E) = \frac{TP + TN}{TP + TN + FP + FN} \times 100(\%) \tag{2}$$

where TP, TF, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

Finally, the number of expanded features is optimally determined as N_E^* such that it can maximize the accuracy $\mathcal{A}(N_E)$ defined in (2), i.e.,

$$N_E^* = \underset{\forall N_\mathcal{M} \in \{\mathbb{N} \cap [0, \frac{l}{2}]\}, l < t_{sp}}{\arg\max} \mathcal{A}(N_E). \tag{3}$$

In this paper, all possible values of N_E are evaluated by machine learning algorithms to determine N_E^*,

The proposed feature expansion procedure is shown in Fig. 3.

III. Experiment

A. Experiment Setup

In this paper, we consider the 1-D pupil size data collected from left and right eyes of 35 participants (i.e., $\mathbf{P} = \{1, 2, \ldots, 35\}$) measured at 30Hz. The data sample in $\widetilde{\mathbf{X}}$ for pupil sizes is generated by considering the deviation from the mean pupil size in the base response for each participant p and the average of them per second [6]. Each participant $p \in \mathbf{P}$ had 2 states $s \in \{1, 2\} = \mathbf{S}$, which represents learning ($s = 1$) state and post learning ($s = 2$) state, respectively.

We consider two measures $\{M_F, M_A\} \in \mathcal{M}$ for FFT (Fast Fourier Transform) and ARIMA (Auto Regressive Integrated Moving Average) that return the number of coefficients $\{N_F, N_A\} \in N_\mathcal{M}$, respectively. FFT is a mathematical transform representing time domain data into the frequency domain by decomposing different frequency sinusoidal waves. The

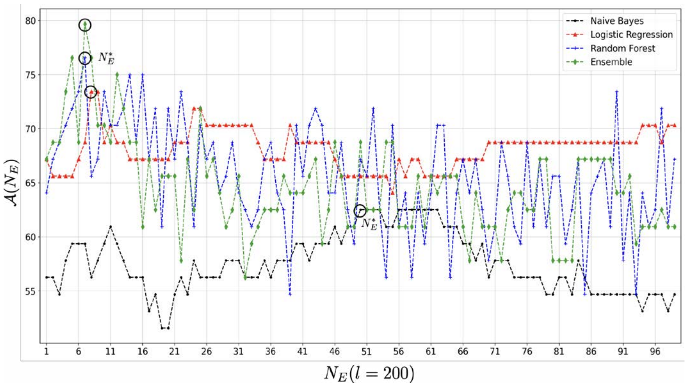

Fig. 4. Performance of the proposed feature expansion over N_E.

number of positive coefficients N_F is obtained by FFT for the data with length l. The other measure is ARIMA, a widely used method to analyze time series data based on its past values, which can capture the existing patterns in non-seasonal series. The ARIMA model is characterized by three parameters, p, d, q, where p is the number of lag observations, d is the number of differences needed for stationarity, and q represents the number of lagged value for errors [7]. We determined the number of parameters p, q as N_A, and $d = 0$ since $\widetilde{\mathbf{X}}_{sp}$ is assumed to be stationary. For $\mathcal{S}$, we use mean, standard deviation, variation, skewness, kurtosis, median absolute deviation, interquartile range (IQR) and standard error.

The optimal number of expanded features N_E^* in (3) is determined by evaluating the performance of model with $\mathbf{T}_s$ driven by different set of (N_F, N_A, l). The performance is evaluated by the machine learning based classifiers, which are Ensemble, Random Forest, Logistic Regression, and Naïve Bayes [8]–[12].

B. Experiment Results

The classification performance based on the proposed feature expansion is shown in Fig. 4. For performance comparison, we consider prior work [13] that also solves the classification problem for pupil size data. Unlike the proposed feature expansion approach, the work in [13] does not expand the features. Rather, it exploits the generation of new data points by grouping the original data points. This approach cannot consider the characteristics of time series data.

Fig. 4 shows the accuracy achieved by several machine algorithms for classification with N_E and $l = 200$. It is clearly observed that the accuracy significantly depends on the number of expanded features N_E and each algorithm has different N_E^* that achieves the best performance.

Table I shows the classification accuracy of the proposed feature expansion. The proposed approach shows 76%-81% accuracy while kernel SVM (Support Vector Machine) and kNN (k-Nearest Neighbors) used in [13] only achieves 64%-76% accuracy, depending on the number of grouped data points from 1 to 60 [13]. For the proposed approach, parameters are optimized, i.e., 23-37 features are determined as N_E^* for each algorithm with l. The experiment results confirm that the frequency and temporal characteristics features in data should be explicitly considered besides statistical features in order to achieve better performance.

TABLE I. THE COMPARISON OF CLASSIFICATION PERFORMANCE

Experiment	[13]		Proposed method			
Algorithm	kernel SVM	kNN	Ensemble	Random Forest	Logistic Regression	Naïve Bayes
$\mathcal{A}(\%)$	64-75	64-75	79.2453	81.2498	79.3103	76.4706
N_E^*, l	-	-	37, 240	35, 201	23, 220	23, 248

IV. Conclusion

In this paper, we propose a feature expansion approach that transforms 1-D time series data into multi-dimensional data, leading to improved classification performance. Since blind adoption of machine learning algorithms on 1-D data classification may show limited performance, we propose to expand the features of 1-D data to multiple dimensional data. The proposed approach can be generalized by including more feature extraction measures and statistical methods. We confirm that the proposed feature expansion is effective for actual biosensor data, i.e., pupil size data, as it improves the classification accuracy.

References

[1] M. S. Goodwin, C. A. Mazefsky, S. Ioannidis, D. Erdogmus, and M. Siegel, "Predicting aggression to others in youth with autism using a wearable biosensor," *Autism research*, vol. 12, no. 8, pp. 1286–1296, 2019.

[2] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81 542–81 554, 2019.

[3] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC medicine*, vol. 17, no. 1, pp. 1–9, 2019.

[4] R. Sharma, R. Singla, and A. Guleria, "A new labeled flow-based dns dataset for anomaly detection: PUF dataset," *Procedia Computer Science*, vol. 132, pp. 1458–1466, 2018, international Conference on Computational Intelligence and Data Science.

[5] S. Jainta and T. Baccino, "Analyzing the pupil response due to increased cognitive demand: An independent component analysis study," *International Journal of Psychophysiology*, vol. 77, no. 1, pp. 1–7, 2010.

[6] D. Huh, J. Kim, and I. Jo, "A novel method to monitoring changes in cognitive load in video-based learning," *Journal of Computer Assisted Learning*, vol. 35, no. 6, pp. 721–730, 2019.

[7] J. D. Hamilton, *Time series analysis*. Princeton university press, 1994.

[8] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 41–53, 2016.

[9] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," *Journal of Biomedical Science and Engineering*, vol. 6, no. 05, p. 551, 2013.

[10] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of Biomedical Informatics*, vol. 35, no. 5-6, pp. 352–359, 2002.

[11] I. Rish, "An empirical study of the Naïve Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[12] M. Aly, "Survey on multiclass classification methods," *Neural Network*, vol. 19, pp. 1–9, 2005.

[13] J. Lee, I. Jo, and H. Park, "Data reconfiguration algorithm for efficient learning state classifications based on pupil sizes," *The Journal of Korean Institute of Communications and Information Sciences*, vol. 45, no. 10, pp. 1756–1766, 2020.