Optimal Feature Selection for EMG-Based Finger Force Estimation Using LightGBM Model
Yuhang Ye, Chao Liu, Nabil Zemiti, Chenguang Yang

To cite this version:

HAL Id: lirmm-02315613
https://hal-lirmm.ccsd.cnrs.fr/lirmm-02315613
Submitted on 14 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract—Electromyogram (EMG) signal has been long used in human-robot interface in literature, especially in the area of rehabilitation. Recent rapid development in artificial intelligence (AI) has provided powerful machine learning tools to better explore the rich information embedded in EMG signals. For our specific application task in this work, i.e. estimate human finger force based on EMG signal, a LightGBM (Gradient Boosting Machine) model has been used. The main contribution of this study is the development of an objective and automatic optimal feature selection algorithm that can minimize the number of features used in the LightGBM model in order to simplify implementation complexity, reduce computation burden and maintain comparable estimation performance to the one with full features. The performance of the LightGBM model with selected optimal features is compared with 4 other popular machine learning models based on a dataset including 45 subjects in order to show the effectiveness of the developed feature selection method.

I. INTRODUCTION

Human-robot interaction (HRI) has been an active research topic during past few decades as the advances in robot technology and artificial intelligence (AI) have enabled the robotic systems to work side-by-side or in a collaborative fashion with humans. The applications have extended beyond traditional industrial domain to new areas such as service, health care, etc [1].

Efficiency and safety are of paramount importance in practical HRI applications. As indicated in [2], one principle of efficient HRI is to “Directly Manipulate the World”, meaning interfaces should allow the task to be done without drawing attention to the robot and the interface per se. Conventional human-robot interfaces (such as cameras, force sensors) are not sufficient due to their limitations in working condition requirements, cost, flexibility, level of intelligence, etc. Myoelectric or Electromyogram (EMG) signal has been used as a complementary source of information for HRI since it originates from the eletrophysiological and mechanical activation of muscle fibers in vivo and hence can be considered as a natural pathway to detect human motion intention from the nervous system without constraining human motion. This signal has been especially and almost exclusively used in the domain of robotic rehabilitation in order to control artificial limb prosthesis and exoskeleton system since early literature [3], [4]. In rehabilitation literature, the EMG signal has been mainly used for motion and posture control of artificial limbs prostheses and hands. Much less but recent research efforts have also been made for force control of artificial hands using EMG signal [5].

During the past few years, along with the booming development of artificial intelligence, new research results indicate that plain and traditional EMG signals can be further explored in a better way by improving the signal analysis methods. Except conventional motion and posture control of upper limb prostheses and hands, recent research works have been able to derive new biomechanical information such as arm stiffness and to study finer scale such as hand fingers with the aide of new machine learning methods [6]. And the applications of EMG is no longer limited to rehabilitation devices, and they also find wider potential applications in HRI such as teleoperation of robots, haptic devices, and so on. In these new HRI applications, human-robot interaction force is of increasing importance in successful completion of task, with the haptic teleoperation as a good example [7]. An increasing number of researchers have been attracted to work on improving the performance of surface EMG (sEMG)-based force estimation from various aspects. In literature, numerous models have been proposed to map the EMG-force relationship, including the Hill model [8], polynomial fitting model [9], fast orthogonal search (FOS) [10], and parallel cascade identification (PCI) [11]. As state-of-the-art, models of two main groups are widely used: the ones based on neural network (NN) [12] and the ones based on decision tree [13]. Both kinds of methods can be effective for different estimation tasks, and both have their advantages and limitations.

In this work, the specific application is to estimate in real time the finger force of intact subject based on forearm surface EMG signal measurement. Although not new, this serves as the first step towards our study on the effects of surgeons’ arm and hand muscle activation on their skill performances in both robot-assisted surgical training and operation, where the force and motion generated by surgeon’s hand finger and wrist are the decisive factors. To achieve accurate and efficient finger force estimation, the proper estimation model is to be first decided. Considering the requirement of generalisability due to large number of potential users, NN-based methods may not be the best choice as neural networks are usually hard to train and tune due to the subject-dependant cost for training and optimizing.
A LightGBM (Gradient Boosting Machine) model under the general class of Gradient Boosting Decision Tree (GBDT) methods has been chosen for this estimation task with its wide generalisability, robustness against signal noises and less proneness to overfitting. However, it is noticed that the estimation/regression performance of LightGBM model depends on the features extracted from the sample datasets. The feature(s) selection is often empirical and hence cannot guarantee the best performance. Generally, the performance is better with more features used. On the other hand, more feature extraction implies more computation cost and therefore longer computation time, which is critical for real time application.

Therefore, in this study, we develop a method to automatically select the optimal features among various time-domain and frequency-domain features to be used in the LightGBM model for the specific aforementioned application scenario. The number of selected optimal features should be minimal in order to save the calculation burden but do not cause noticeable performance deterioration at the same time. The feature selection is performance-based and therefore is not affected by the designer’s experience. The comparisons of the finger force estimation performance obtained from the LightGBM with optimal features are made to the one of LightGBM with full features and also to the estimation performances of 4 other popular models (linear regression (LR), support vector regression (SVR), convolutional neural network (CNN), and Multilayer Perceptron (MLP)) with all features available in order to confirm the effectiveness of the selected optimal features and the automatic selection method.

II. MATERIALS AND METHODS

A. Experimental Setup

In this work, we use the online open dataset "putEMG-Force" provided by the Biomedical Engineering and Biocybernetics Team of Poznan University of Technology, Poland, to serve as the basis for evaluation of the method we propose. The putEMG Dataset is a database of surface electromyographic activity recorded from forearm which allows the development and evaluation of algorithms for both gesture recognition and hand finger force estimation. The essential information on the experiment platform setup and data collection is provided in following, and more details can be found on the putEMG website and [16].

The experiment platform is illustrated in Fig. 1. The system is dedicated to sEMG signal acquisition of forearm muscle activity for a single subject. In this work, we use the EMG signals recorded from 24 electrodes fixed around subject right forearm and the finger forces measured by tensometer sensors as shown in Fig. 2. The dataset includes experiment data collected from 45 healthy, fully-abled subjects (8 females, 37 males) aged 19 to 37 years old. It should be noted that the value of measured force from the tensometer has no unit of physical meaning, since the conversion has not been done yet according to the dataset provider.

![putEMG experiment platform](image1)

**Fig. 1:** putEMG experiment platform

![EMG electrodes placement and finger force sensors](image2)

(a) EMG band placement  
(b) Hand dynamometer  
(c) Sensor numeration

**Fig. 2:** EMG electrodes placement and finger force sensors

B. Background

EMG signal is a non-stationary signal containing a variety of noises and artifacts. These noises will deteriorate the models in performance if the raw EMG data is used as the input for classification or regression. Therefore, features extraction is very important in EMG signal processing. The widely used features can be summarized into three categories: time domain (TD) features, frequency domain (FD) features and time-frequency domain (TFD) features.

Time-domain statistical analysis is the most commonly used method for EMG signal processing because they are faster and smaller than other features, and this is very important for real-time systems. In 1993, Hudgins et al. proposed five TD feature of sEMG, including mean absolute value (MAV), mean absolute value slope (MAVSLP), slope sign changes (SSC), waveform lengths (WL) and zero crossings (ZC) [18], [19], [20]. Zardoshi et al. extracted the features of integral of absolute value (IVA), zero crossing (ZC), variance (VAR), Willison amplitude (WAMP) to control artificial limb by EMG signal [17]. Ahmad and Chappel adopted the skewness (Skew), Kurtosis (Kurt) and moving approximate entropy (mApEn) as the TD feature for the first time [21].

In order to investigate the stability of time-domain EMG features on a task of low and high forces classification by EMG, Dennis Tkach et al. [22] used 11 TD features in

---

their experiment and they found the classifier performance could be improved by the use of at least four combined EMG features and the multi-dimensional features, such as autoregression coefficients (AR) and cepstrum coefficients (CC), could get the best performance.

Although time-domain features are easy to extract, a lot of studies have shown that when the muscle contraction force changes slightly, the frequency-domain description of EMG signals is relatively more stable than the time-domain features [23]. The frequency domain analysis of EMG mainly applies power spectrum analysis methods. As early as the late 1980s, Christensen used Fourier transform to analyze the power spectrum of surface EMG signals [24]. The difference of power spectrum between normal individuals and patients with neuromuscular diseases was revealed by the amplitude ratio of high frequency to low frequency. The features of power spectrum extraction mainly include peak frequency (PKF), mean power frequency (MNP), Mean frequency, median frequency (MDF) and so on, which have been successfully used as muscle force, fatigue and geometry indices [25]. To further improved the EMG performance, Al-Timemy et al. [26] considered the time-frequency characteristics of EMG and proposed a set of TD-PSD feature for hand movement classification with three force levels.

Traditional Fourier transform can better depict the signal characteristics in global frequency but can hardly provide the frequency information in the time domain, which is insufficient for the analysis of the non-stationary EMG signal. The time-frequency analysis methods combine both the time and the frequency information for EMG signal analysis and thus received wide research attention. At present, the main time-frequency analysis methods used in EMG analysis are short-time Fourier transform (STFT), Wigner-Ville transform (WVD), Choi-Williams distribution (CWD) and wavelet transform (WT). Veer et al. used Short-time Fourier transform and wavelet transform for the recognition of arm movements and their result showed that wavelet transform performs better [27].

Benefiting from the improved computation power and the development of big data, deep learning algorithms have been studied and applied in the fields of image recognition, speech recognition, and intelligent robots. It has a strong capability to automatically extract features by end-to-end learning. In recent years, many researchers have tried to apply deep learning to the field of sEMG signal recognition. Atzori et al. [28] used the proposed CNN architecture for hand movement recognition based on the EMG signal. After the preprocessing of sub-sample and low-pass filter, EMG data are fed into the CNN model which is composed of convolutional, pooling, and softmax layers. They compared the CNN method with other classical classification methods and got the conclusion that the classification accuracy of CNN’s is higher than the average results of classical methods but lower than the best reference methods using random forests algorithm with the features of discrete wavelet transform (DWT), histogram (HIST), waveform length (WL) and root mean square(RMS). Therefore, CNN is not always superior in feature extraction, and how to extract the well-performed features and find optimal feature combination is very important in EMG recognition field, which is also the focus in this work.

C. Optimal Features Selection Strategy

1) Data preprocessing and feature extraction: In order to extract the effective signal of EMG and reduce noise, a bandpass filter of 10-350Hz is used on the 24-channels EMG data while a 50Hz low pass filter is adopted on the force data. Taking the accuracy and real-time requirements of force estimation into consideration synthetically, we use a sliding window with the length of 256 sample points (50 ms) and move across 32 sample points (6.25 ms) for each step to segment the EMG and the force signals. The segmenting process is illustrated in Fig. 3.

![Fig. 3: EMG segmentation with 50ms sliding window](image)

To evaluate the feature importance and find the best feature combination, we consider the most commonly used features both in time and frequency domain and most of them have been introduced in subsection II-B. We do not select the time-frequency domain feature because our features were extracted in a 50ms window so the Fourier transform can be seen as a Short-time Fourier transform which contains the time and frequency information. The features we consider is shown in Table I.

After data filtering and feature extraction, EMG raw data of one sample is transformed into a feature vector \( \{x_1, x_2, ..., x_d\} \). Because the features are calculated by different methods as shown in Table I, their scale are different which will affect the convergence speed and precision of the algorithm. A min-max scaled normalization method was adopted here to convert the features to the same scale,

\[
\hat{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (1)
\]

where \( \hat{x}_i \) is the normalized features, \( x_{\min} \) and \( x_{\max} \) are the minimum and maximum value of this feature. After that, all the features will be normalized to a range from 0 to 1.

2) Measure of the Feature importance: Gradient Boosting Decision Tree (GBDT) [37] is an ensemble learning algorithm that utilizes decision or the regression trees as weak classifiers. It continuously adds new regression trees by the method of gradient descent, so that the total model residuals are continuously reduced.

Supposed that we have a training set with \( n \) samples of EMG data after normalization represented as \( \{x_1, x_2, ..., x_n\} \),
and each \( x_i \in \mathbb{R}^d \) is a feature vector with \( d \) dimensions extracted using the methods in Table I. \( \{g_1, g_2, \ldots, g_n\} \) are the negative gradients with respect to the output of the GBDT model which are calculated to select the most informative features for the tree node split. The most informative features can split samples in the largest information gain so that the tree can converge faster. The information gain of GBDT is usually measured by the variance after splits. In each iteration, we calculate the information gain of each segmentation point \((j,s)\) by traversing each feature \(j\) and the split value \(s\) in turn. The information gain is formulated in the following form:

\[
V(j, s) = \frac{1}{n} \left( \frac{\sum_{x_{ij} \leq s} g_i}{n_l^j(s)} + \frac{\sum_{x_{ij} > s} g_i}{n_r^j(s)} \right)^2
\]

where \( n \) is the number of the data in training set. \( x_{ij} \) denotes the feature \( j \) of the \( i^{th} \) sample. \( n_l^j(s) \) and \( n_r^j(s) \) are the numbers of the samples split into left and right child node, so \( n_l^j(s) = \sum I[x_{ij} \leq s] \) and \( n_r^j(s) = \sum I[x_{ij} > s] \) and \( I[\cdot] \) is the indicator function. For each feature \( j \), the GBDT algorithm selects the best split value by \( s_j^* = \arg \max_s V(j, s) \) and calculates the largest gain \( V(j, s_j^*) \). Then the algorithm will select the best feature and split value for the data split.

Here, we make full use of the advantages of the GBDT algorithm in feature selection and measure the importance of a feature by calculating the number of times this feature is used for data split. Obviously, those features that can achieve greater information gain often provide better discrimination because they always put similar samples together and separating different samples in data split.

3) **Features selection:** From the above subsection, the importance of each feature can be easily calculated by counting the number of times the features used in data split for the tree growth. To select the optimal feature combination, the root mean square error (RMSE) is used here to measure the regression performance, which is defined as below:

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]

where \( n \) is the number of the EMG samples, \( y_i \) is the label of the force signal and \( \hat{y}_i \) is the predict output of \( x_i \). Firstly, the result of RMSE using all the features by GBDT model is set as a baseline performance. Then, an iterative algorithm is used and drop the last 10% or 20% features according to the feature importance in each time. Performance of the remaining features is tested in GBDT model and the feature importance is re-evaluated. If the performance loss (compared with the baseline performance) is less than certain threshold after deleting the feature, the algorithm will continue to drop the last features based on the new feature importance until the performance loss exceeds the threshold or the number of features is less than the preset minimum, and the last feature combination will be the optimal feature. It should be noted that for a specific application the feature selection is expected to be conducted in advance to accelerate the online EMG signal processing, thus the calculation time for the offline feature selection phase is not demanding. The algorithm is outlined in Algorithm 1.

### III. EXPERIMENTAL EVALUATION

In robot-assisted surgical operation, surgeons usually use their thumb and index finger to control the joystick, so the muscle activation related to the index finger shows more potential in surgeons’ skill performances analysis. Therefore, force 4 (channel 4 in the force sensor shown in Fig.2c) estimation of index finger is used for experimental evaluation in this work. For the putEMG database, there are five kinds of trajectories consisting of varying combinations of pressure magnitude, duration and shape, and the trajectories of “repeats_long”. Each trial of “repeats_long” contains 5 action blocks (each block for 70s and 350 seconds in total) where thumb, index, middle, ring+small and all fingers take turns. For each block, press action repeated 11 times with varying parameters. The press action of the index finger is mainly applied between 70 and 140 seconds in a trial, which was selected for our experiment. For each trial, we divided the data into three parts: the last 20% data for testing, the rest 80% data for validation, and the rest for training. Force prediction for multiple channels and subject-independent estimation are our future works and thus not considered in this paper.

In the experiment, the proposed optimal features selection strategy is tested in six randomly selected subjects from the all 45 subjects. The GBDT model we choose is the LightGBM (LGB), which performs extremely well in both

---

**TABLE I: List of Features**

<table>
<thead>
<tr>
<th>Features</th>
<th>Abbreviation</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Integral Absolute Value</td>
<td>IAV</td>
</tr>
<tr>
<td>2</td>
<td>Average Amplitude Change</td>
<td>AAC</td>
</tr>
<tr>
<td>3</td>
<td>Auto-Regressive Coefficients</td>
<td>AR</td>
</tr>
<tr>
<td>4</td>
<td>Cepstral Coefficients</td>
<td>CC</td>
</tr>
<tr>
<td>5</td>
<td>Kurtosis</td>
<td>Kurt</td>
</tr>
<tr>
<td>6</td>
<td>LOG Detector</td>
<td>LOG</td>
</tr>
<tr>
<td>7</td>
<td>Maximum Amplitude Slope</td>
<td>MAX</td>
</tr>
<tr>
<td>8</td>
<td>Mean Absolute Value</td>
<td>MAV</td>
</tr>
<tr>
<td>9</td>
<td>Mean Absolute Value Slope</td>
<td>MAVSLP</td>
</tr>
<tr>
<td>10</td>
<td>Multiple Hamming Windows</td>
<td>MHW</td>
</tr>
<tr>
<td>11</td>
<td>Myopulse Percentage Rate</td>
<td>MYOP</td>
</tr>
<tr>
<td>12</td>
<td>Skewness</td>
<td>Skew</td>
</tr>
<tr>
<td>13</td>
<td>Slope Sign Change</td>
<td>SSC</td>
</tr>
<tr>
<td>14</td>
<td>Absolute Temporal Moment</td>
<td>TM</td>
</tr>
<tr>
<td>15</td>
<td>Variance</td>
<td>VAR</td>
</tr>
<tr>
<td>16</td>
<td>Willison Amplitude</td>
<td>WAMP</td>
</tr>
<tr>
<td>17</td>
<td>Waveform Length</td>
<td>WL</td>
</tr>
<tr>
<td>18</td>
<td>Zero Crossing</td>
<td>ZC</td>
</tr>
<tr>
<td>19</td>
<td>Mean Frequency</td>
<td>MF</td>
</tr>
<tr>
<td>20</td>
<td>Median Frequency</td>
<td>MF</td>
</tr>
<tr>
<td>21</td>
<td>Mean Power frequency</td>
<td>MNP</td>
</tr>
<tr>
<td>22</td>
<td>Frequency Ratio</td>
<td>FR</td>
</tr>
<tr>
<td>23</td>
<td>Variance of Central Frequency</td>
<td>VCF</td>
</tr>
<tr>
<td>24</td>
<td>Power Spectrum Ratio</td>
<td>PSR</td>
</tr>
<tr>
<td>25</td>
<td>Signal-to-Noise Ratio</td>
<td>SNR</td>
</tr>
<tr>
<td>26</td>
<td>Maximum-to-minimum Drop in Power Density Ratio</td>
<td>DFR</td>
</tr>
</tbody>
</table>
Algorithm 1 Automatic Algorithm for Optimal Feature Selection

**Input:** Training data with high dimension features \( x_i \); Label of the training data \( y_i \); The percentage of features retained for each iteration \( P \% \); Performance loss threshold \( \theta \); Minimum feature numbers \( k \).

**Output:** The optimal features combination best.

**Initial:** The number of iteration \( t = 0 \); The features selected are all the features \( Feature = all \); The RMSE performance of the all the features in validation set \( R_0 \).

**Repeat:**

1. \( t = t + 1 \).
2. The best feature combination \( best = Feature \).
3. Train the GBDT model by gradient boosting. Use the variance as shown in eq.2 for the informative gain measure and split the data by the best feature and best split value.
4. Calculate the feature importance by computing the number of time for each features using in data split. And sort the feature by feature importance.
5. Use the RMSE to evaluate the performance in validation set and get the result of \( R \).
6. Select the top \( P \% \) feature \( tp \) according to the feature importance, \( Feature = tp \). And len(\( tp \)) is the number of the feature in top \( P \% \). Update the training set.
7. Until : \( R - R_0 > \theta \) or \( len(tp) < k \).

From Table II, it is seen that AAC and TM both appear as the best feature depending on different subject, and we can therefore conservatively draw a conclusion that AAC and TM features may be considered as the optimal feature combination in the total 26 features.

![Fig. 4: The force estimation performance with the selected features and all features for subject 35](image)

**TABLE II: The selected feature of the six subjects**

<table>
<thead>
<tr>
<th>Subject</th>
<th>s12</th>
<th>s14</th>
<th>s15</th>
<th>s35</th>
<th>s37</th>
<th>s39</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected feature</td>
<td>AAC</td>
<td>AAC</td>
<td>AAC</td>
<td>TM</td>
<td>TM</td>
<td>AAC</td>
</tr>
</tbody>
</table>

From Table II, it is seen that AAC and TM both appear as the best feature depending on different subject, and we can therefore conservatively draw a conclusion that AAC and TM features may be considered as the optimal feature combination in the total 26 features.

To further evaluate whether the selected features contain enough information for force estimation with small performance loss, the results of selected optimal features (AAC and TM) in lightGBM model are compared with 4 widely used models using all the features. The 4 models are linear regression (LR), support vector regression (SVR), convolutional neural network (CNN), and Multilayer Perceptron (MLP). The CNN model contains 6 convolution layers, 3 pooling layers, and 1 fully-connected layer. The activation function is rectified scaled exponential linear units (SELU), batch normalization, and dropout are also used in CNN models. There are 3 hidden layers in MLP model and the activation function is rectified.
Table III: Comparison result of the five models evaluated by RMSE. Models of LR, SVR, MLP, and CNN use all the features while LGB are tested on all and the selected two feature. In the column of LGB, on the left side of "/'" are the all features and on the right side the selected features.

<table>
<thead>
<tr>
<th>Subject</th>
<th>LR</th>
<th>SVR</th>
<th>MLP</th>
<th>CNN</th>
<th>LGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>s12</td>
<td>45.89</td>
<td>49.16</td>
<td>43.69</td>
<td>43.03</td>
<td>38.54/38.33</td>
</tr>
<tr>
<td>s14</td>
<td>53.83</td>
<td>44.25</td>
<td>43.82</td>
<td>29.00</td>
<td>37.91/40.34</td>
</tr>
<tr>
<td>s15</td>
<td>38.29</td>
<td>30.69</td>
<td>26.42</td>
<td>22.75</td>
<td>25.04/25.32</td>
</tr>
<tr>
<td>s35</td>
<td>45.77</td>
<td>43.68</td>
<td>44.85</td>
<td>37.88</td>
<td>36.31/36.64</td>
</tr>
<tr>
<td>s39</td>
<td>46.79</td>
<td>43.69</td>
<td>38.20</td>
<td>40.23</td>
<td>39.15/38.49</td>
</tr>
<tr>
<td>Average</td>
<td>42.01</td>
<td>39.62</td>
<td>36.77</td>
<td>32.93</td>
<td>33.06/33.43</td>
</tr>
</tbody>
</table>

ACKNOWLEDGMENT

This work was partially supported by the LabEx NUMEV incorporated into the I-Site MUSE [Grant AAP-Exploratoire 1830]; the French National Center for Scientific Research [Grant PRC2014]; the National Nature Science Foundation (NSFC) [Grant 61811530281]; the French ANR within the Investissements d’Avenir Program (Labex CAMI, ANR-11-LABX0004, Labex NUMEV, ANR-10-LABX-20, and the Equipex ROBOTEX, ANR-10-EQPX-44-01).

REFERENCES