Paper:

Patient classification based on sEMG signals using extreme gradient boosting algorithm

Juan Zhao [1,2], Jinhua She [3]*, Dianhong Wang [1], and Feng Wang [1,2]

[1] School of Automation, China University of Geosciences, Wuhan 430074, China;
[2] Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China;
[3] School of Engineering, Tokyo University of Technology, Hachioji, 192-0982, Japan;
zhaojuan0859@cug.edu.cn; she@stf.teu.ac.jp; {wangdh, wangfeng}@cug.edu.cn;

Abstract: Studying surface electromyography(sEMG) signals of the muscles near the knee joints in patients with gonarthritis is helpful for the diagnosis of knee joint inflammation. If the influence of different sEMG signals and their weights on knee inflammation can be analyzed through machine learning methods, it will greatly improve diagnostic accuracy. The extreme gradient boosting (XGBoost) algorithm is an excellent machine learning algorithm. Inspired by this algorithm, we presented a signal classification method based on the XGBoost algorithm to distinguish between patients with gonarthritis and healthy subjects. The sEMG signals collected from four muscles around the knee are extracted as features, which are used as the input variables to the classification model. The XGBoost algorithm determines the output by improving the objective function based on sample proportion and weight. The experimental results show that the XGBoost algorithm has higher accuracy and better classification performance when compared with the support vector machine (SVM) and the deep neural networks (DNN) algorithms. This indicates that the advantage of the XGBoost algorithm on classifying patients with gonarthritis based on sEMG signals.

Keywords: Surface electromyography(sEMG); Machine learning; Extreme gradient boosting (XGBoost); Patients with gonarthritis

1. Introduction

Knee joint inflammation is a disease based on degenerative pathological changes. Middle-aged and elderly people mostly suffer from this disease. Knee joint inflammation always occultly grows in the body of the patients, which is the main reason for leg pain. It may lead to joint deformities, disabilities, and other symptoms if it is not treated in time. Current clinical diagnosis mostly uses X-ray [1], arthroscopy [2], and other medical imaging diagnostic techniques. X-ray has low accuracy in the diagnosis of early rheumatoid arthritis, synovial hyperplasia, joint effusion, and bone erosion [3]. And long-term X-ray examination has a low radiation effect on bodies and is detrimental to patients who need regularly to see a doctor for knee pain due to seasonal variations. An arthroscope may cause different degrees of sequelae, including infection, knee adhesion, and localized skin scarring. Then these methods are not suitable for continuous and long-term diagnosis of knee joint inflammation. There is a huge market demand for a non-radiative, non-invasive, long-term diagnostic method for diagnosing patients with gonarthritis.

Surface electromyography(sEMG) is an electrical signal that detects neuromuscular activity by using electrodes on the skin surface. It can quantitatively record the electrophysiological characteristics of neuromuscular activity, muscle fatigue degree, and nerve conduction velocity. The detection method through sEMG is relatively simple and non-invasive. The sEMG signals around the knee joint reflect the state and coordination level of neuromuscular in real-time, which provides important evidence for the occurrence, development, and rehabilitation process of motor injury of the knee joint. So, it is highly valued in the fields of human movement science, ergonomics, and rehabilitation medicine and has a wide range of research values [4].

The sEMG signals collected from the same muscles vary greatly for different patients with knee motion disorders, and even for the same patient at different stages of rehabilitation [5]. Because the sEMG signals have the non-stationary characteristics, it is feasible to use machine learning methods to classify sEMG signals under different conditions. Support vector machine (SVM) is a kind of machine learning method of dichotomy model that is a linear classifier with the maximum interval defined in the feature space. The learning strategy of SVM is interval maximization, and the learning algorithm is to find the optimal solution of convex quadratic programming. Lucas. e.t. used the SVM to supervise and classify sEMG signals so as to realize the control of the sEMG prosthesis [6]. However, the probability density function of sEMG obeys the Gaussian distribution with the mean value of 0 and is not strictly linear fractal data. Using standard SVM algorithm to classify sEMG signals may result in low classification accuracy. Liu. e.t. proposed a method based on deep neural networks (DNN) for sEMG signal classification and has high accuracy [7]. But, the DNN easily falls into the local optimal values. The activation of different

initial values may lead to different local optimal solutions [8]. Thus, the DNN-based method may not be suitable for the classification of knee arthritis.

Extreme gradient boosting (XGBoost) algorithm is a new machine learning algorithm presented in 2017 [9]. It has been proved to have a high accuracy rate of 92.09% in the safety evaluation of underground comprehensive pipe gallery [10] and an accuracy rate of 98.4% in the recognition of hip shape [11]. Therefore, this paper is inspired by the XGBoost algorithm and uses it to classify sEMG signals in patients with gonarthritis. We studied the sEMG signals of 11 patients with previously diagnosed knee joint inflammation and 11 healthy subjects. These sEMG signals were classified by establishing an individual risk classification model for knee lesions to identify patients and healthy subjects. The results show that the XGBoost algorithm provides an auxiliary method for the diagnosis of knee joint inflammation from the intersection of medicine, data science, and computer science.

2. Data and Method

We utilized some sEMG signals of an open dataset to carry on the experiments. Then we present an XGBoost-based method to discriminate patients with gonarthritis from healthy subjects.

2.1. Data

The data source is from an open dataset of Batalla de Sanidad with the support of the Mueva Granada Military University in July 2012. The data were collected from 22 participants: 11 patients diagnosed previously with gonarthritis and 11 healthy volunteers that had no history of neurological disorders. The participants had various genders, ages, heights, and other demographic information. Each participant was instructed to collect sEMG signals from the rectus femoris muscle, biceps femoris, medius femoris, and semitendinosus muscle while sitting with the legs stretched and bent. This collection was repeated. A total of 60 cases were collected in this dataset.

In this study, the sEMG signals from four different muscles were selected as the features for training models and testing models.

2.2. Preprocessing

Because the data were obtained from actual scenes, there were some non-standard data that can not be directly applied to the experiment. Thus, data preprocessing is needed to solve this problem.

First, some data are incomplete, possibly due to incorrect acquisition or missing storage. Such data may interfere with the training model and are discarded directly. Second, some data lack labels that show whether they are from patients or healthy people, which should not be included in the training samples and needed to be excluded. Finally, the ranges of data collected from the same muscle vary from person to person, which has a great impact on the optimization rate and performance of the training model. Therefore, the input data are standardized in this study as follows:

$$x_i = \frac{x_i' - \mathrm{Mean}(R_i)}{\mathrm{Std}(R_i)} \quad (1)$$

Where, x_i' is the value of the ith feature, $\mathrm{Mean}(R_i)$ is the expected value of the ith feature of all data within the reference range, and $\mathrm{Std}(R_i)$ is the standard deviation of all data within the reference range.

After the above preprocessing, 22 cases with valid data remained, and the length of each feature is 15,300. We use these data as the training samples and testing samples for our method.

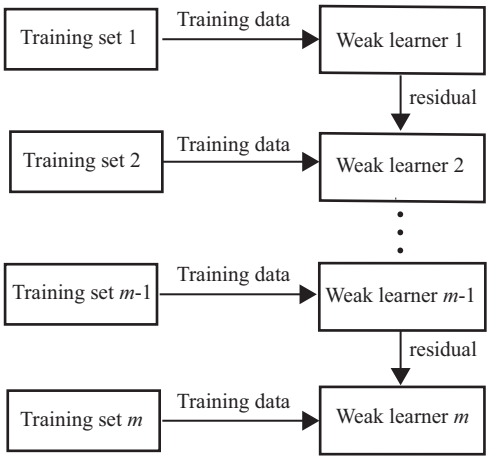

Fig. 1. Flow Chart of training weak learners

2.3. XGBoost

The method aims to predict the samples according to the features and to determine whether the subject is a patient with gonarthritis or not. This is a typical classification, which is to find a relationship between an input about some samples, $X = \{x_1, x_2, \cdots, x_n\}$, and an output Y. We use the XGBoost algorithm to classify the input samples in order to construct an intelligent detection system with high accuracy more quickly.

The core idea is to integrate many weak learners to form a strong learner. Each weak learner is a tree model generated by the classification and regression trees (CART). The XGBoost algorithm keeps splitting features to grow a new tree, which is to learn a new function to fit the residual of the previous prediction. Supposing m is the number of weak learners, and $k \in [1, m]$, the XGBoost algorithm trains the kth weak learner according to the residuals of the previous $(k-1)$ weak leaners and the training data of the kth weak learner. The flow chart of training weak learners is shown in Fig. 1.

we randomly select a series of preprocessed data as the training data and construct the weak learners. For the kth weak learner, if $\hat{y}_i^{(k)}$ is the prediction of the true value $y_i^{(k)}$, its prediction is the sum of the leaf weights on all previous

weak learners.

$$\widehat{y}_i^{(k)} = \sum_{t=1}^{k} f_t(x_i), \quad \ldots \quad (2)$$

where $f_t(x_i)$ is the function of the kth weak learner. Considering the prediction of the $(k-1)$th weak learner, $\widehat{y}_i^{(k)}$ is further expressed as

$$\widehat{y}_i^{(k)} = \widehat{y}_i^{(k-1)} + f_k(x_i). \quad \ldots \quad (3)$$

Then we use the loss function to measure the accuracy of our prediction model, which is expressed as

$$L^{(k)} = \sum_{i=1}^{s} l(y_i^{(k)}, \widehat{y}_i^{(k)}), \quad \ldots \quad (4)$$

where s is the number of samples on all leaf nodes when we grow k trees, and $l(y_i, \widehat{y}_i) = |(y_i - \widehat{y}_i)|$ is the residual of $\widehat{y}_i$ and y_i. The least loss function means the most accurate prediction model.

Note that the XGBoost algorithm is an excellent algorithm that achieves the balance between model accuracy and computational speed. It introduces a regularization item to measure the model complexity, and thus, it measures the computational efficiency. The regularization item is expressed as

$$\Omega(f_t) = \gamma m + \frac{1}{2}\lambda \sum_{t=1}^{m} f_t^2, \quad \ldots \quad (5)$$

where, γ is the penalty parameter of L1 norm, and λ is the penalty parameter of L2 norm. The regularization item is determined by the number of leaf nodes of the generated tree and the L2 norm of the corresponding value vector of leaf nodes. So it enhances the generalization ability of the model.

We add the regularization item to the loss function, which forms an objective function:

$$\mathrm{O}^{(\mathrm{k})} = \sum_{i=1}^{s} l(y_i^{(k)}, \widehat{y}_i^{(k)}) + \sum_{t=1}^{k} \Omega(f_t). \quad \ldots \quad (6)$$

The first part is to measure the fitting ability of the model for the samples, and the second part is to control the complexity of the model and avoid over-fitting.

Considering the prediction of the $(k-1)$th weak learner, there is

$$\sum_{t=1}^{k} \Omega(f_t) = \sum_{t=1}^{k-1} \Omega(f_t) + \Omega(f_k). \quad \ldots \quad (7)$$

Since the regularization items of the previous weak learners are known, $\sum_{t=1}^{k-1} \Omega(f_t)$ can be regarded as a constant C.

Then substituting (3) and (7) into (6), the objective function is expressed as

$$\mathrm{O}^{(k)} = \sum_{i=1}^{s} l(y_i^{(k)}, \widehat{y}_i^{(k-1)} + f_k(x_i)) + \Omega(f_k) + C. \quad (8)$$

If we find the optimal solution of this objective function, we get a suitable model. In this case, the constant can be ignored in the solving process.

We expand the loss function, $l(y_i^{(k)}, \widehat{y}_i^{(k-1)} + f_k(x_i))$, in (8) in a second order Taylor series:

$$l(y_i^{(k)}, \widehat{y}_i^{(k-1)}) + g_i f_k(x_i) + \frac{1}{2} h_i f_k^2(x_i). \quad \ldots \quad (9)$$

Where, g_i is the first derivative of $l(y_i^{(k)}, \widehat{y}_i^{(k-1)})$, and h_i is the second derivative of $l(y_i^{(k)}, \widehat{y}_i^{(k-1)})$. Because $\widehat{y}_i^{(k-1)}$ is a known value in the kth step, $l(y_i^{(k)}, \widehat{y}_i^{(k-1)})$ is a constant and has no effect on the function optimization. Thus, the objective function of (8) is nearly

$$\mathrm{O}^{(k)} \approx \sum_{i=1}^{s} [g_i f_k(x_i) + \frac{1}{2} h_i f_k^2(x_i)] + \Omega(f_k). \quad \ldots \quad (10)$$

If we define $I_j = \{i | q(x_i) = j\}$ is the sample set on all leaf nodes for the jth weak learner, the function of this weak learner can be expressed in terms of the sum of the weights of all leaf nodes, that is, $f_j = w_j = \sum_{i \in I_j} w_i$. Then let $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$, substituting (5) into (10) obtains

$$\mathrm{O}^{(k)} = \sum_{j=1}^{m} [G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma m. \quad \ldots \quad (11)$$

When we know the samples of each leaf node, G_j and H_j are known. Then we calculate the minimum value of (11) and get the optimal weights:

$$w_j^* = -\frac{G_j}{H_j + \lambda}. \quad \ldots \quad (12)$$

Substituting the optimal weight into the objective function again gets the result:

$$\mathrm{O}^{(k)} = -\frac{1}{2} \sum_{j=1}^{m} \frac{G_j^2}{H_j + \lambda} + \gamma m. \quad \ldots \quad (13)$$

Thus, we weigh the output of all weak learners and get the output of a final strong learner.

2.4. Algorithm

The flow chart of the XGBoost algorithm is shown in Figure 2.

Step 1: Preprocess the data of the training sets. Carrying out cleaning and standardized processing selects the suitable data sets, which are used as the input data set of the model.

Step 2: Train the model. The algorithm first initializes the weights of the leaf nodes. Then it calculates the objective function and determines whether the value is minimum. If the value is not minimum, the weights are updated. Otherwise, using the weights gets the weak learners and establishes the corresponding strong learner. Thus, we obtain the classification model.

Step 3: Verify with the testing set. The testing set enters the classifier, and then the classification effect is obtained from the classifier.

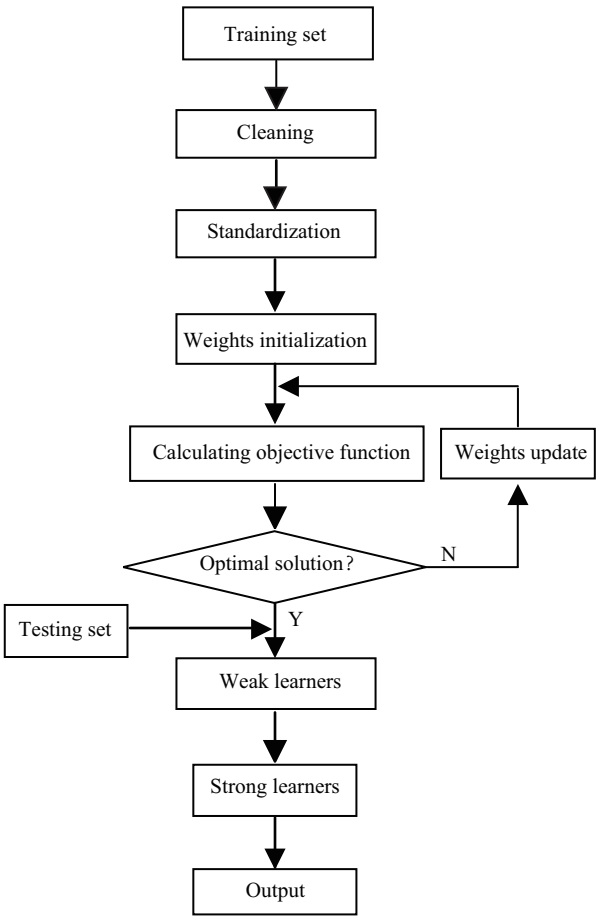

Fig. 2. Flow Chart of XGBoost-based method

3. Experimental results and analysis

This study used the SVM, DNN, and XGBoost algorithms to process the same preprocessed samples in order to compare the classification effect. The selected data set is divided into a training set and a testing set in a ratio of 0.8:0.2. And the random seeds are set to be the same to ensure that the extraction methods of training set and testing set are consistent for different models.

3.1. Metrics

Three classical metrics of the model, accuracy (Acc), sensitivity (Sen), and specificity (Spe), are used for model evaluation:

$$\mathrm{Acc} = \frac{\mathrm{Tp}+\mathrm{Tn}}{\mathrm{Tp}+\mathrm{Tn}+\mathrm{Fp}+\mathrm{Fn}}, \quad (14)$$

$$\mathrm{Sen} = \frac{\mathrm{Tp}}{\mathrm{Tp}+\mathrm{Fp}}, \quad (15)$$

$$\mathrm{Spe} = \frac{\mathrm{Tn}}{\mathrm{Tn}+\mathrm{Fp}}. \quad (16)$$

Where, Tp is the abbreviation of "true and positive" and represents the number of samples that the labels of input samples are 1 (confirmed patients) and the prediction results of the model are also 1, Tn means "true and negative" and represents the number of samples that the labels are 0 (confirmed healthy subjects) and the prediction results of the model are also 0; Fp means "false and positive" and represents the number of samples that the labels are 0 while the prediction results are 1; Fn is "false and negative" and presents the number of samples that the labels are 1 while the prediction results are 0.

Accuracy refers to the proportion of correctly predicted samples to the total samples, with a range of [0,1]. The larger the value is, the better prediction ability the model has. When the sensitivity of the model is higher, the recognition ability of the model for patients with gonarthritis is stronger, that is, the model has a lower misdiagnosis rate. For the specificity, the larger value means that the model has a stronger discrimination ability for ordinary patients and the misdiagnosis rate is lower.

Table 1. Comparison of performance evaluation for samples of 1 patient with gonarthritis and 1 healthy subject.

Methods	Acc [%]	Sen [%]	Spe [%]
SVM	99.70	99.12	93.81
DNN	98.91	97.61	99.56
XGBoost	99.70	97.88	99.47

Table 2. Comparison of performance evaluation for samples of 4 patients with gonarthritis and 4 healthy subjects.

Methods	Acc [%]	Sen [%]	Spe [%]
SVM	85.09	76.55	79.67
DNN	85.50	84.32	85.49
XGBoost	89.02	85.08	89.45

3.2. Results

We randomly selected a set of data from one of the 11 patients with gonarthritis and a set of data from one of the 11 healthy subjects. The data was classified by using the SVM, DNN, and XGBoost algorithms to determine whether the output is a patient or not. The three metrics were calculated and the experimental results are shown in Table 1.

Similarly, we randomly selected the samples of 4 patients and 4 healthy subjects and classified them with three algorithms. The results of the indexes are shown in Table 2. Finally, we chose all samples to be classified and calculated the evaluation indexes, which are shown in Table 3.

The experimental results show that most of the values of the XGBoost algorithm are significantly higher than those of other algorithms in three cases. This proves the superior performance of the XGBoost-based method on classifying patients with gonarthritis.

Table 3. Comparison of performance evaluation for samples of 11 patients with gonarthritis and 11 healthy subjects.

Methods	Acc [%]	Sen [%]	Spe [%]
SVM	78.90	87.37	80.74
DNN	82.41	75.46	82.62
XGBoost	87.70	94.33	87.61

3.3. Comparison and Analysis

From Table 1, it is seen that the accuracy of the XGBoost algorithm is 99.70%, while those of the SVM and DNN algorithms are 98.91% and 99.70%, separately. This means that the XGBoost and SVM algorithms have good prediction ability and a strong fitting effect when classifying 1 patient with gonarthritis. For the sensitivity, the XGBoost algorithm has the value of 97.88%, which is higher than the DNN algorithm but lower than the SVM algorithm. This shows that the SVM algorithm has the best diagnosis performance in this case. In terms of specificity, the XGBoost algorithm obtains the value of 99.47%, which is outstanding when compared with other methods. In this table, it can be seen that most of the values obtained by the XGBoost algorithm are larger than those obtained by other algorithms. This illustrates that the XGBoost algorithm has the best classification effect. And the SVM algorithm gets some values slightly larger than the DNN algorithm, which means that the SVM algorithm has advantages over the DNN method when distinguishing 1 patient between 1 healthy subject.

From Tables 2 and 3, it can be seen that the values obtained by the XGBoost algorithm are higher than those obtained by the SVM and DNN algorithms for the same evaluation metrics. The sensitivity of the SVM algorithm is larger than that of the DNN algorithm in Table 3. Except of this, the other values of the SVM algorithm are smaller than the corresponding values of the DNN algorithm. These tables also indicate that the XGBoost algorithm has the best classification effect on determining more patients with gonarthritis and healthy persons while the DNN algorithm is superior to the SVM algorithm in most cases.

In addition, for the SVM algorithm, the accuracy does not fluctuate very much and is around 80%, and the sensitivity is around 85%, while the specificity is around 80%. This shows that the SVM algorithm is not sensitive to the number of input data. But for the DNN algorithm, the values of the indexes are very different in three cases: the values are the largest for samples of 1 patient with gonarthritis and 1 healthy person, while the values are the smallest for samples of 11 patients with gonarthritis and 11 healthy people. This indicates that the DNN algorithm is sensitive to the number of input data. When the method is used to process small amounts of data, it has good classification performance. Otherwise, the classification effect is not ideal.

The experimental results show the advantages of the XGBoost algorithm on sEMG signal classification. This is related to its algorithm process. The XGBoost algorithm implies a computationally efficient variant of the gradient boosting algorithm. It uses not only the first derivative but also the second derivative. And it introduces the regular term to avoid over-fitting in the search for the optimal solution, which makes the loss function more precise. Moreover, because the probability density function of sEMG signals has the Gaussian distribution with the mean value of 0, the XGBoost algorithm is an appropriate processing approach.

4. Conclusion

Accurate classification of diseases is helpful to efficient diagnosis and treatment. This paper presents an XGBoost-based method to differentiate patients with gonarthritis from healthy subjects, which is characterized by sEMG signals from four muscles. The method improves the objective function based on the proportion and weight of samples, which improves the classification and prediction ability. The experimental results show that, compared with the SVM and DNN algorithms, the XGBoost-based algorithm achieves high classification accuracy and has good classification and prediction performance. This indicates significant potential for classifying patients with gonarthritis based on sEMG signals.

References:

[1] Y. Li, N. Xu, Q. Lyu, Construction of a knee osteoarthritis diagnostic system based on X-ray image processing, Cluster Computing. 2019, 22(6): 15533-15540.

[2] N. A. Segal, M. C. Nevitt, J. A. Lynch, et al., Diagnostic performance of 3D standing CT imaging for detection of knee osteoarthritis features, The Physician and Sportsmedicine. 2015, 43(3):1-8.

[3] M. Avci, N. Kozaci, Comparison of X-Ray imaging and computed tomography scan in the evaluation of knee trauma, Medicinal. 2019, 55(10): 623-633.

[4] X. Liu, Z. A. Tipu, G. B. Peter, Intraoperative monitoring of motor symptoms using surface electromyography during stereotactic surgery for movement disorders, Journal of Clinical Neurophysiology. 2005, 22(3): 183-191.

[5] J. Fu, L. Xiong, X. Song, et al., Identification of finger movements from forearm surface EMG using an augmented probabilistic neural network, 2017 IEEE/SICE International Symposium on System Integration (SII). 2019, 55(10): 623-633.

[6] M. F. Lucas, A. Gaufriau, S. Pascual, et al., Multichannel surface EMG classification using support vector machines and signal-based wavelet optimization, Biomedical Signal Processing & Control. 2008, 3(2):169-174.

[7] G. Liu, L. Zhang, B. Han, et al., sEMG-Based continuous estimation of knee joint angle using deep learning with convolutional neural network, 2019 IEEE 15th International Conference on Automation Science and Engineering(CASE). 2019: 140-145.

[8] Y. Bian, X. Xie, Generative chemistry: drug discovery with deep learning generative models, Journal of Molecular Modeling. 2021, 27(3): 71:1-10.

[9] Y. Liu, H. Wang, Y. Fei, et al., Research on the prediction of green plum acidity based on improved XGBoost, Sensors. 2021,21(3): 930:1-8.

[10] B. Snider, E. A. Mcbean, Improving time-to-failure predictions for water distribution systems using gradient boosting algorithm, WDSA\CCWI Joint Conference Proceedings. 2018,7:1-8.

[11] D. Molinaro, I. Kang, J. Camargo, et al., Biological hip torque estimation using a robotic hip exoskeleton, The 8th IEEE RAS\EMBS International Conference on Biomedical Robotics & Biomechatronics. 2020, 11: 150-155.