OPTIMIZING FACTORS INFLUENCING ON ACCURACY OF BIOMETRICAL CARDIOMETRY

Marat R. Bogdanov[1,2], Aleksander A. Dumchikov[2], Vadim M. Kartak[1,2], Aigul I. Fabarisova[2]

[1] Ufa State Aviation Technical University, Ufa, Russia,
redfoxufa@gmail.com
[2] M.Akmullah named after Bashkir State Pedagogical University, Ufa, Russia,
kvmail@mail.ru

Abstract. The paper is about some aspects concerning person biometric identification based on using of electrocardiograms. Signal preprocessing routing is considered in the paper. Classification was carried out with support vector machines algorithm. Tuning of hyper parameters of classification is considering.

Keywords: Biometric person identification, electrocardiogram, machine learning. Support vector machines, hyper parameter tuning

1 Introduction

Various biometric methods of person identification are getting more popular. Fingerprinting, face, voice and retina recognition are widely used in various security systems. The vulnerabilities of traditional methods of biometric identification were revealed over time. Researchers are increasingly turning their attention to such person biometric features as electrocardiograms, electroencephalograms and DNA [1]. In this paper, we would like to discuss some practical aspects of person identification using ECG.

2 Motivation and Aim

The problem of person biometric identification concerns classification problems. To solve it, we have to consider algorithms from some finite set and choose an algorithm that gives the least error of the forecast [2]. Let's introduce some notation.

Let us suppose X is a space of objects.

Y is a set of answers.

$$X^l = (x_i, y_i)_{i=1}^l \tag{1}$$

is a training set, l is a sample size.

$$y_i = y^*(x_i), \tag{2}$$

$$A_t = \{a : X \to Y\} \tag{3}$$

are a models of algorithms, $t \in T$, T is a number of algorithms under consideration.

$$\mu_t : (X \times Y)^l \to A_t \tag{4}$$

are learning methods. It is required to find a method μ_t with the best generalizing power.

When finding a method μ_t, we often have to solve the following subtasks:

- Choice of the best model A_t (model selection).
- Choice of learning method μ_t for a given model A_t (in particular, optimization of hyperparameters).
- Features selection:

$$F = \{f_j : X \to D_j : j = 1, ..., n\} \tag{5}$$

is a set of features. The method of learning μ_j uses only features $J \subseteq F$.

To assessment the quality of learning by precedents it's used:

$L(a, x)$ is a cost function of algorithm a on the object x.

$$Q(a, X^l) = \frac{1}{l} \sum_{i=1}^{l} L(a, x_i) \tag{6}$$

is a functional of accuracy a on X. In this case we consider an internal quality criterion that is measured on the training set X^l:

$$Q\mu(X^l) = Q(\mu(X^l), X^l) \tag{7}$$

and an external criterion evaluating the quality of learning on hold-out set X^k[2]:

$$Q\mu(X^l, X^k) = Q(\mu(X^l), X^k). \tag{8}$$

In the paper presented we will consider such aspects of person biometric identification as feature selection, model selection, choice of methods (tuning of hyperparameters), assessment of the quality of learning.

3 Feature Selection

We used the MGH/MF Waveform Database hosted at `physionet.org` resource [3], [4]. The Massachusetts General Hospital/Marquette Foundation (MGH/MF) Waveform Database is a comprehensive collection of electronic recordings of hemodynamic and electrocardiographic waveforms of stable and unstable patients in critical care units, operating rooms, and cardiac catheterization laboratories. It is the result of a collaboration between physicians, biomedical engineers and nurses at the Massachusetts General Hospital. The database consists of recordings from 250 patients and represents a broad spectrum of physiologic and pathophysiologic states. Individual recordings vary in length from 12 to 86 minutes, and in most cases are about an hour long [3], [4].

The typical recording includes three ECG leads, arterial pressure, pulmonary arterial pressure, central venous pressure, respiratory impedance, and airway CO_2 waveforms. The raw sampling rate of 1440 samples per second per signal was reduced by a factor of two to yield an effective rate of 360 samples per second per signal relative to real time [3], [4].

When preprocessing stage we used a biopsy python library by John Reid. The package enables the development of Pattern Recognition and Machine Learning work flows for the analysis of biosignals including ECG [5]. Using *biopsy* we extracted first lead from electrocardiogram and performed a low pass filter for reducing of redundancy. After applying of low-pass filter R-peaks was extracted from ECG-signal using a *wfdb* python library by Chen Xie and Julien Dubiel [6]. The software allow extract peaks and QRS-cycles from electrocardiograms. We choice amplitude and temporal features of Q,R and S-peaks (Q_x, Q_y, R_x, R_y, S_x, S_y). In total we had 6 features. Feature table together label class vector were randomly splitted into training set and testing set in the ratio of 75:25 for further cross validation. We learned a classifier on training set and performed measuring of classifying accuracy on testing set.

4 Model Selection

We used a Support Vector Machines (SVM) algorithm for classification. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. SVM is primarily a classier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables [7].

To construct an optimal hyperplane, SVM employs an iterative training algorithm, which is used to minimize an error function. According to the form of the error function, SVM models can be classified into four distinct groups:

- Classification SVM Type 1 (also known as C-SVM classification)
- Classification SVM Type 2 (also known as nu-SVM classification)
- Regression SVM Type 1 (also known as epsilon-SVM regression)
- Regression SVM Type 2 (also known as nu-SVM regression)

We used a Classification SVM Type 1 (also known as C-SVM classification) model.

5 Classification SVM Type 1

For this type of SVM, training involves the minimization of the error function:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \tag{9}$$

subject to the constraints:

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i \geq 0, i = 1, ..., N, \tag{10}$$

where C is the capacity constant, w is the vector of coefficients, b is a constant, and ξ_i represents parameters for handling nonseparable data (inputs). The index i labels the N training cases. Note that $y \in \pm 1$ represents the class labels and x_i represents the independent variables. The kernel ϕ is used to transform data from the input (independent) to the feature space. It should be noted that the larger the C, the more the error is penalized. Thus, C should be chosen with care to avoid overfitting.

6 Kernel Functions

$$K(X_i, X_j) = \left\{ \begin{array}{cc} X_i \cdot X_j & Linear \\ (\gamma X_i \cdot X_j + C)^d & Polynomial \\ exp(-\gamma |X_i - X_j|^2) & RBF \\ tanh(\gamma X_i \cdot X_j + C) & Sigmoid \end{array} \right\}, \tag{11}$$

where $K(X_i, X_j) = \phi(X_i) \cdot \phi(X_j)$ that is, the kernel function, represents a dot product of input data points mapped into the higher dimensional feature space by transformation ϕ.

7 Gamma Is an Adjustable Parameter of Certain Kernel Functions

The RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis [8].

Support vector machine classifier supported by sklearn python library uses as default following hyper parameters: C=1.0, kernel='rbf', gamma='auto'. When using of default parameters while performing of classification of electrocardiograms we had accuracy score equal to 0.93. We tuned hyper parameters of classification with Grid Search procedure varying C parameter in range of [1, 10, 100, 1000], kernel in range of ['linear, 'rbf"], and gamma in range of [1e-3, 1e-4]. After performing of tuning we had the following best parameters set: 'kernel': 'rbf', 'C': 10, 'gamma': 0.001. Using these parameters we had accuracy score equal to 0.99.

8 Results and Discussion

During the preprocessing of electrocardiograms we extracted first leads of signal and performed low-pass filter for reducing redundancy. Then we extracted cardiac cycles from the leads and extracted Q, R and S peaks from cardiac cycles.

Using amplitude and temporal features of peaks we composed a feature table containing 6 features (Q_x, Q_y, R_x, R_y, S_x, S_y) and class labels vector y. Further we randomly splitted a feature table and class labels vector into training set and testing set on ration of 75:25 for further cross-validation. Training set was used for learning a classifier and testing set was used for assessment of quality of learning. SVM classifier supported by sklearn python library using default options show accuracy score equal to 0.93. We found the best hyper parameters are following: 'kernel': 'rbf', 'C': 10, 'gamma': 0.001. Using these parameters we could improved accuracy score up to 0.99.

References

1. Abdulmonam, O., Ahlal, H., Fawzia, E.: Vulnerabilities of Biometric Authentication. Threats and Countermeasures. IJICT. 4(11), 947-958 (2014)
2. Machine Learning course by K.V. Voroncov: `http://www.MachineLearning.ru/wiki` [On-line; accessed 12-April-2011]
3. Welch, J., Ford, P., Teplick, R., Rubsamen, R.: The Massachusetts General Hospital-Marquette Foundation Hemodynamic and Electrocardiographic Database Comprehensive collection of critical care waveforms. J Clinical Monitoring 7(1), 96-97 (1991)
4. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mietus, J., Moody, G., Peng, C., Stanley, H.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 101(23), 1-6 (2000)
5. The biosppy Toolbox: `http://biosppy.readthedocs.io/en/stable/` [On-line; accessed 24-March-2016]
6. The WFDB Python Toolbox: `https://pypi.python.org/pypi/wfdb` [On-line; accessed 11-July-2015]
7. Support Vector Machines (SVM) Introductory Overview: `http://www.statsoft.com/Textbook/Support-Vector-Machines` [On-line; accessed 26-May-2012]
8. Ting-Fan, W., Chih-Jen, L., Weng, R., Singer., Y. (eds.): Probability Estimates for Multi-class Classification by Pairwise Coupling. Journal of Machine Learning Research. 5, 975-1005 (2004)