MULTISAB: A Web Platform for Analysis of Multivariate Heterogeneous Biomedical Time-series

Jovic A.[1], Kukolja D.[1], Friganovic K.[1], Jozic K.[2] and Cifrek M.[1]

[1]University of Zagreb Faculty of Electrical Engineering and Computing, Unska 3, HR-10000 Zagreb, Croatia
[2]INA – Industrija Nafte d.d., Avenija Veceslava Holjevca 10, p.p. 555, HR-10002 Zagreb, Croatia
`alan.jovic@fer.hr; davor.kukolja@fer.hr; kresimir.friganovic@fer.hr`

Abstract. There is a growing need for efficient and accurate biomedical software in healthcare community. In this paper, we present MULTISAB, a web platform whose goal is to provide users with detailed analysis capabilities for heterogeneous biomedical time series. We describe the system architecture, including its subprojects: frontend, backend and processing. Emphasis is placed on the processing subproject, implemented in Java, which incorporates data analysis methods. The subproject is divided into several frameworks: record input handling, preprocessing, signal visualization, general time series features extraction, specific (domain) time series features extraction, expert system recommendations, data mining, and reporting. Common signal features extraction framework includes a great number of features in time (both linear and nonlinear), frequency and time-frequency domain. Currently, domain specific frameworks for heart rate variability, ECG and EEG feature extraction are supported, which also include preprocessing techniques for noise reduction and detection methods for characteristic waveforms (like QRS complexes, P and T waves in ECG). Parallelization is implemented for feature extraction to increase performance. It is realized using multithreading on several levels: for multiple records, traces, and segments. Expert system is implemented, which provides automatic recommendation of the set of significant expert features that should be extracted from the analyzed signals, depending on the analysis scenario. The expert system, apart from the role in recommending features, can also participate in automatic diagnosis, after the features are extracted. Current expert system prototype contains diagnostic rules for acute myocardial ischemia, based on medical guidelines. Data mining framework contains dimensionality reduction methods and machine learning classifiers used to construct accurate and interpretable disorder models. A report is produced at the end of the process using openly available libraries. The platform includes best practices from medicine, biomedical engineering, and computer science in order to deliver detailed biomedical time series analysis services to its users.

Keywords: biomedical time series, web platform, analysis

1 Introduction

In the era of expanding big data, the need for healthcare community to develop tools for analysis and inspection of their data is inevitable. Clinicians and researches are mainly restricted to use off-line solutions, like MATLAB (or Octave) based scripts or frameworks [1], or other specialized software, e.g. for heart rate variability (HRV) analysis [2].

In this paper, we present a web-based platform for analysis of biomedical time series, developed under the ongoing MULTISAB (MULtivariate TIme Series Analysis in Biomedicine) project. The platform is intended for researchers and clinicians to perform detailed analysis and set diagnosis based on imported biomedical time series data, as well as to visualize and inspect signals data. The access to MULTISAB can be achieved through web browsers, and is therefore available from any computer, without requisite installations. A major benefit of using the platform is that a user doesn't need excessive computing power to do the analysis, as the computations are done remotely. The project is currently in the process of module integration and extensive testing. We plan to provide access in a few months.

The MULTISAB project is divided into three subprojects: *backend*, *frontend*, and *processing*. Detailed description of the architectural overview of the platform can be found in [3]. Due to lack of space, in continuation of this paper, we inspect in some detail only the *processing* subproject, which covers the majority of platform's functionality. Architecture and frameworks implemented in the subproject are described in Section 2, while in Section 3 we introduce an example of analysis in MULTISAB platform from a user's perspective. Conclusion is given in section 4.

2 The *processing* subproject architecture

The *processing* subproject, located on the server side, is responsible for data analysis. Within it, key frameworks for processing and analysis of biomedical time series are implemented, namely: record input handling, preprocessing, signal visualization, general time series features extraction, specific (domain) time series features extraction, expert system recommendations, data mining, and reporting frameworks.

To reduce computation time needed to perform an analysis, multithreaded parallelization is implemented under the *analysis* package for feature extraction. We plan to add it also for data mining methods. The details about parallelization will be presented in another forthcoming paper. All the *processing* frameworks are written in Java 9 and run on a host computer. They are described in the following subsection.

2.1 Implemented frameworks

Record input handling framework supports a majority of data formats used for biomedical time series storage: *.ann*, *.csv*, *.edf/edf+c* and *.txt.* file formats. Additionally, metadata and signal parameter data are supported. Signal visualization is entirely done through the web browser. Signals are uploaded and, based on the selected record data, the image of a record or its segment is generated. The images are generated in the *processing* subproject and sent to the *backend* subproject in order to be delivered to the user on *frontend*.

In the preprocessing framework, functions for noise filtering and morphological mathematical operations (erosion, dilation, opening and closing) are implemented. Commonly used features in signal processing, with emphasis on biomedical applications, are extracted under the general time series features framework. In Table 1, a list of those features is given. EEG, ECG and HRV specific time series features extraction are implemented under eegAnalysis, ecgAnalysis and hrvAnalysis classes, respectively, and are also shown in Table 1.

Table 1. General time series features implemented in the framework

Domain	Features
Time	auto correlation coefficient, Fano factor, mean, standard deviation, variance, minimum, maximum, PCA
Frequency	Fast Fourier transform PSD estimate, Burg autoregressive PSD estimate, Lomb-Scargle PSD estimate
Nonlinear	correlation dimension, space fill index (SFI), central tendency measure (CTM), largest Lyapunov exponent, recurrence plot, nonlinear forecasting, DFA, Higuchi dimension, Hurst exponent, 1/f power law alpha exponent, corrected conditional Shannon entropy, alphabet entropy, approximal entropy (ApEn), one dimensional Carnap entropy, correlation entropy, Renyi entropy, fuzzy ApEn, Sample entropy, Shannon entropy, cross recurrence, mutual dimension, synchronization likelihood, Allan factor, Lempel-Ziv complexity
ECG specific	R wave amplitude, QRS complex duration, P wave amplitude, P wave absence, P wave duration, T wave amplitude, T wave duration, ST segment duration, ST segment slope, QT interval duration, PR interval duration, J point amplitude, R/S ratio, Q wave amplitude, R wave duration, S wave duration, atrial fibrillatory rate
EEG specific	mean of first differences, mean of second differences, mean of first differences normalized, mean of second differences normalized, total PWR, delta band power, theta band power, alpha band power, beta band power, gamma band power, spectral entropy
HRV specific	AVNN, SDNN, RMSSD, pNNX, NNX, SDSD, SDANN, HRV triangular index, TINN, total PWR, ULF, VLF, LF, HF, LF/HF ratio, spectral entropy, standard deviation ratio

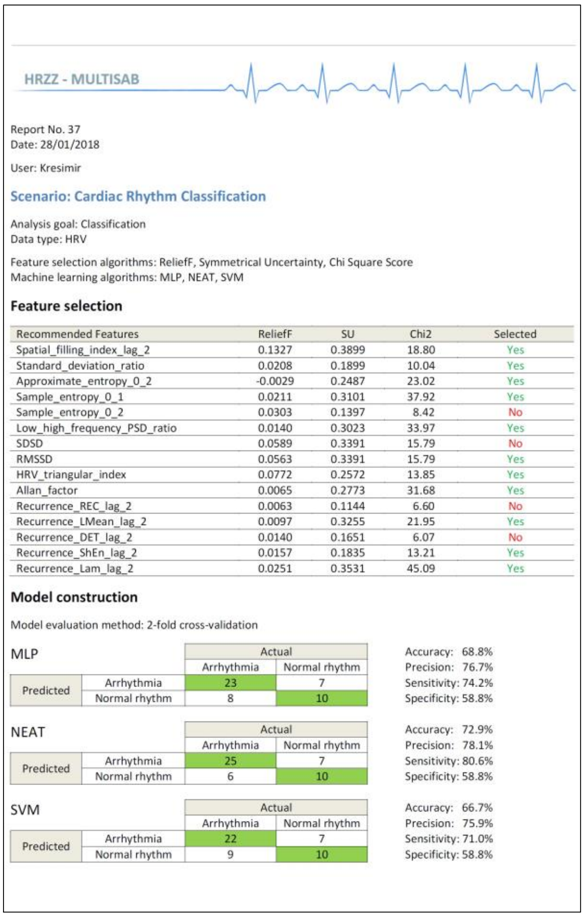
Two phases of an expert system are implemented under the MULTISAB platform. The first phase is designed to recommend a list of specific features intended for feature extraction (some cardiac diagnoses are currently supported), based on medical knowledge (guidelines and other scientific sources) about a disease. A user can add or remove the features from the list, which enables flexibility in particular cases. The second phase consists of a set of rules for reaching a specific diagnosis, for example for acute heart ischemia, which is done after the features are extracted. For feature selection, we currently support ReliefF, Symmetrical Uncertainty (SU) and Chi Square Score (Chi2) [4]. Multilayer Perceptron (MLP), RBF, NEAT and PNN neural networks classification methods are supported via Encog library [5], while SVM is supported via LIBSVM library [6]. Class-Attribute Contingency Coefficient is used for discretization of numerical features. Additional methods are planned for implementation (e.g. random forest, RIPPER). For evaluation, *k*-fold cross-validation and train:test:validation dataset separation are supported.

3 Analysis Scenario

In this chapter, we will introduce the MULTISAB platform with one analysis scenario, from user's perspective. For the use case, we have selected cardiac rhythm classification based on MIT-Arrhythmia Database records [7]. After selecting a new analysis, the user can select a predefined scenario for cardiac rhythm classification (for more detail, see [8]). This predefined scenario contains all the necessary preprocessing steps and recommended expert features that need to be extracted from the signals. The cardiac rhythm classification scenario is based on the authors previous research [9] and on consultations with cardiologists. The goal type of the analysis in this scenario is classification. The data type is HRV. To complete the specification of the scenario, the user has selected 30 seconds segment length; ReliefF, SU and Chi2 feature selection algorithms; MLP, NEAT, and SVM classifiers. After the specification of the scenario, the user needs to upload the data for analysis.

In this use case, the 48 .ann records from MIT-Arrhythmia are uploaded into the platform. In the next step, the user can inspect and visualize the uploaded data. In the following preprocessing step, the user can select various procedures for signal filtering, characteristic waveform detection, and data transformations. In the case of HRV analysis for cardiac rhythm classification, the user needs to select the appropriate transform to obtain spectrum for frequency domain feature extraction. The expert system by default selects Lomb Scargle [10]. Feature extraction is the central step in the analysis of biomedical time series. In this use case, the user will already be provided with a set of recommended expert features. The list of expert features is not fixed, it depends on the chosen segment length and preprocessing procedures. Regardless of the scenario, the user can modity the list of features in the feature extraction phase. The user can also specify parameters for calculation of a particular feature.

In this use case, the user has selected one segment length of 30 seconds, starting at 0 seconds. This gives one sample (feature vector) per record. In the feature extraction analysis tab, the user defines classification label/disease for each segment or record. Although .ann files from MIT database have reference annotations, in this use case, the user has manually annotated each record with arrhythmia or normal rhythm. Alternatively, the user could select automatically the definition based on segment annotations, but here, that would result in detection of 12 diseases plus normal rhythm, with too few examples for each class for classifiers to learn from. After feature extraction, the user can download extracted features in .arff or .csv file format and continue independent analysis locally. In feature selection step, the user can see the results for feature selection measures. Based on these results, the user can omit certain features from further analysis. To complete the analysis, the user has to, in model construction step, select the model evaluation method(s). At the end of the analysis, the user can see the generated report, based on JasperReports library [11]. In Fig. 1, we can see the report created for this scenario. At the beginning of the report, we can see the report number, analysis end date and the user who conducted the analysis. In the feature selection part, the list of all extracted feature is given, together with the information about feature selection measures. We also see which features were used for model construction (green "Yes" in the "Selected" column). In model construction part, we see the results for arrhythmia classification for each machine learning method. In this scenario, model evaluation method was 2-fold cross-validation. The results are given in the form of confusion matrix and related accuracy, precision, sensitivity and specificity values. For the final version of MULTISAB platform, we plan to develop a more detailed report, which will contain the information about all the analysis steps.



HRZZ - MULTISAB

Report No. 37
Date: 28/01/2018

User: Kresimir

Scenario: Cardiac Rhythm Classification

Analysis goal: Classification
Data type: HRV

Feature selection algorithms: ReliefF, Symmetrical Uncertainty, Chi Square Score
Machine learning algorithms: MLP, NEAT, SVM

Feature selection

Recommended Features	ReliefF	SU	Chi2	Selected
Spatial_filling_index_lag_2	0.1327	0.3899	18.80	Yes
Standard_deviation_ratio	0.0208	0.1899	10.04	Yes
Approximate_entropy_0_2	-0.0029	0.2487	23.02	Yes
Sample_entropy_0_1	0.0211	0.3101	37.92	Yes
Sample_entropy_0_2	0.0303	0.1397	8.42	No
Low_high_frequency_PSD_ratio	0.0140	0.3023	33.97	Yes
SDSD	0.0589	0.3391	15.79	No
RMSSD	0.0563	0.3391	15.79	Yes
HRV_triangular_index	0.0772	0.2572	13.85	Yes
Allan_factor	0.0065	0.2773	31.68	Yes
Recurrence_REC_lag_2	0.0063	0.1144	6.60	No
Recurrence_LMean_lag_2	0.0097	0.3255	21.95	Yes
Recurrence_DET_lag_2	0.0140	0.1651	6.07	No
Recurrence_ShEn_lag_2	0.0157	0.1835	13.21	Yes
Recurrence_Lam_lag_2	0.0251	0.3531	45.09	Yes

Model construction

Model evaluation method: 2-fold cross-validation

MLP

		Actual	
		Arrhythmia	Normal rhythm
Predicted	Arrhythmia	23	7
	Normal rhythm	8	10

Accuracy: 68.8%
Precision: 76.7%
Sensitivity: 74.2%
Specificity: 58.8%

NEAT

		Actual	
		Arrhythmia	Normal rhythm
Predicted	Arrhythmia	25	7
	Normal rhythm	6	10

Accuracy: 72.9%
Precision: 78.1%
Sensitivity: 80.6%
Specificity: 58.8%

SVM

		Actual	
		Arrhythmia	Normal rhythm
Predicted	Arrhythmia	22	7
	Normal rhythm	9	10

Accuracy: 66.7%
Precision: 75.9%
Sensitivity: 71.0%
Specificity: 58.8%

Fig 1. The report created for cardiac rhythm classification on MIT Arrhythmia Database

4 Conclusion

In this work, we have provided a brief view into the currently developing MULTISAB platform, with the focus on data analysis (*processing*) part. The progress made will soon enable on-the-web analysis of multiple biomedical time series for interested medical and biomedical engineering professionals. For near future work, we plan to add more support for heterogeneous biomedical time series analysis and expand the selection of currently implemented data mining methods.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work has been fully supported by the Croatian Science Foundation under the project number UIP-2014-09-6889.

References

1. Robbins, K.A.: EEGVIS: A MATLAB Toolbox for Browsing, Exploring, and Viewing Large Datasets. Front Neuroinform. 6, 17 (2012).
2. Tarvainen, M.P., Niskanen, J.P., Lipponen, J.K., Ranta-aho, P.O., Karjalainen, P.A.: Kubios HRV – Heart rate variability analysis software. Comput Methods Programs Biomed 113(1), 210-220 (2014).
3. Friganovic, K., Jovic, A., Jozic, K., Kukolja, D., Cifrek, M.: MULTISAB project: a web platform based on specialized frameworks for heterogeneous biomedical time series analysis - an architectural overview. In: Proc. Int. Conf. on Med & Biol Eng (CMBEBiH 2017), pp. 9-15. Springer Nature, Sarajevo, Bosnia and Herzegovina (2017).
4. Tang, J., Alelyani, S., Liu, H.: Feature Selection for Classification: A Review. In: Aggarwal, C. (ed.). Data Classification: Algorithms and Applications. CRC Press. (2014)
5. Heaton, J.: Encog: Library of Interchangeable Machine Learning Models for Java and C#. *JMLR* 16, 1243-1247. (2015).
6. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. *ACM TIST* 2(27), 1-27. (2011).
7. Moody, G.B., Mark. R.G.: The impact of the MIT-BIH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine 20(3), 45-50 (2001).
8. Jovic, A., Kukolja, D., Friganovic, K., Jozic, K., Car, S.: Biomedical Time Series Preprocessing and Expert-System Based Feature Extraction in MULTISAB Platform. In: *Proc. Int. Conf. MIPRO 2017*, pp. 349-354. IEEE, Opatija, Croatia (2017).
9. Jovic, A., Jovic, F.: Classification of cardiac arrhythmias based on alphabet entropy of heart rate variability time series. Biomedical Signal Processing and Control 31(2017), 217-230 (2017).
10. Narin, A., Isler, Y., Ozer, M.: Investigating the Performance Improvement of HRV Indices in CHF using Feature Selection Methods based on Backward Elimination and Statistical Significance. Computers in Biology and Medicine 45, 72-79 (2014).
11. Tibco Software, Inc. 2018. JasperReports® Library, Open Source Java Reporting Library. https://community.jaspersoft.com/project/jasperreports-library (last accessed: 2018/01/21).