Logistic Regression Multinomial for Arrhythmia Detection

Omar Behadada
Biomedical Engineering Laboratory,
Faculty of technology,
University of Tlemcen, Algeria
Email: o_behadada@mail.univ-tlemcen.dz

Marcello Trovati
School of Computing
Edge Hill University
Ormskirk, UK
Email: Marcello.Trovati@edgehill.ac.uk

Chikh MA
Biomedical Engineering Laboratory,
Faculty of technology,
University of Tlemcen, Algeria
Email: mea_chikh@mail.univ-tlemcen.dz

Nik Bessis
School of Computing
Edge Hill University
Ormskirk, UK
Email: Nik.Bessis@edgehill.ac.uk

Yannis Korkontzelos
School of Computing
Edge Hill University
Ormskirk, UK
Email: Yannis.Korkontzelos@edgehill.ac.uk

***Abstract*—In this paper, we introduce a method based on logistics Regression multi-class as a classifier to provide a powerful and accurate insight into cardiac arrhythmia. As suggested by our evaluation, this provide a robust, scalable, and accurate system, which can successfully tackle the challenges posed by the utilization of big data in the medical sector.**

***Index Terms*—Logistic Regression Multinomial, Knowledge Extraction, Big Data.**

I. Introduction

In the last World Health Report 2013, cardiovascular diseases are confirmed to be one of most worrying health issues and the largest cause of mortality in the world [1]. Therefore, the creation of low-cost and high-quality cardiac assessments is indeed an important and topical challenge. Furthermore, the availability of a huge amount of information produced by the continuous development of big data methods and techniques provides new challenges as well as real opportunities in this field.

In particular, the detection of cardiac arrhythmia is a very promising area, because this is closely correlated with premature ventricular contraction (PVC), which is an effective predictor of sudden death. Many classification methods have been used in this field such as Support Vector Machine (SVM), Sparsity based Orthogonal Matching Pursuit (OMP), Independent Component Discriminant Analysis (ICDA), etc [1]. In [2], an algorithm based on Multinomial Logistic Regression (MLR) is introduced, which aims to find the posterior class probability which is aided by a semi supervised segmentation [3], [4].
Classification methods with good classification rates tend to have a low degree of interpretability, which plays a crucial role in any knowledge-based system as it facilitates an effective decision-making progress by providing interpretable knowledge [5]. In [6], the authors introduce a method to define semi-automatically fuzzy partition rules based on a text mining approach from textual sources such as PubMed. The information extracted is combined with expert knowledge and experimental data, to provide a robust, scalable and accurate system.

In this paper we propose logistic Regression Multinomial (MLR) as a classifier of cardiac arrhythmia, which learns the posterior probability distributions of each class, in order to create a robust, scalable and accurate knowledge-based system, which provides a crucial insight into arrhythmia detections from big data information sources [6]. There is much research on the comparison between data mining algorithms, which shows the potential of MLR [7].

The dataset analyzed in the article is part of ongoing research, which is integrating a variety of information from potentially huge unstructured sources, including Doppler and Mir images. See [6] for more details.
The rest of this section focuses on the relevant medical background. In Section II, data preparation and correlation studies are discussed, and in Section III an overview of Multinomial Logistic Regression (MLR) based classification is presented, which is subsequently investigated and assessed in Section IV. Finally, Section V concludes the paper and discusses future research directions.

A. Medical Context

Electrocardiogram (ECG) is a type of non-invasive medical assessment to monitor the activity of the central blood circulatory system. An ECG signal can provide useful information on the normal and pathological physiology of heart activity, as depicted in Figure 1. Therefore, ECG is an important clinical tool for the identification and diagnosis of heart conditions [6]. Early and a prompt detection and classification of ECG arrhythmia is crucial, which particularly affects the treatment of patients in the intensive care units [1]. For more than four decades, computer-aided diagnostic (CAD) systems have been

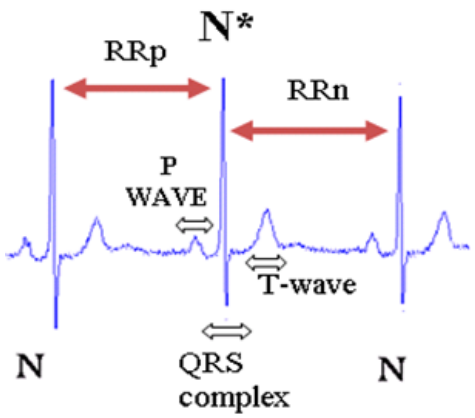

Fig. 1. Standard ECG beat

utilized in the classification of the ECG. This has created a wealth of techniques and methods in this field, which include multivariate statistics, decision trees, fuzzy logic, expert systems and hybrid approaches [1]. In designing of CAD system, the most crucial and relevant step is the integration of a suitable feature extractor and pattern classifier such that they can operate in coordination to create an effective and efficient CAD system [6].

Figure 1 depicts a standard ECG beat. More specifically, the normal heart rhythm has four specific parameters: a P wave, which represents atrial depolarization, a QRS complex associated with ventricular depolarization, a T wave, which represents ventricular repolarization, and finally a U wave connected to papillary muscle repolarization.

II. Data Preparation

The patients who have been considered in the experiments, taken from MIT-BIH [6], are shown in Figures 2 and 3. More specifically, we considered 29 patients, which are referred to as $101, 103, 104, 105, \ldots 209$, as depicted in Figure 2. For example, patient 101 is associated with the following parameters

- Number of normal (N) heart beat: 1860
- Number of arterial contraction beats (A) heart beat: 3
- Number of junctional contraction beats (J): 0
- Number of ventricular contraction beats (V): 0

Figure 3 summarizes all the heart beat taken from MIT-BIH data base.

The R peaks of the ECG signals were identified by utilizing the Tompkins algorithm [1], which is an on-line real time QRS detection algorithm, and it efficiently identifies QRS complex using slop, amplitude, and width information. Furthermore, by automatically adjusting the thresholds and parameters periodically to the standard 24h MIT-BIH arrhythmia database, this algorithm correctly detects 99.3% of QRS complex. From patients with cardiac arrhythmia, taken from MIT-BIT database,

Record	N	A	J	V
101	1860	3	-	-
103	2082	2	-	-
104	163	-	-	2
105	2526	-	-	41
106	1507	-	-	520
107	-	-	-	59
108	1739	4	-	17
109	-	-	-	38
111	-	-	-	1
112	2537	2	-	-
113	1789	-	-	-
114	1820	10	2	43
115	1953	-	-	-
116	2302	1	-	109
117	1534	1	-	-
118	-	96	-	16
119	1543	-	-	444
121	1861	1	-	1
122	2476	-	-	-
123	1515	-	-	3
124	-	2	29	47
200	1743	30	-	826
201	1625	30	1	198
202	2061	36	-	19
203	2529	-	-	444
205	2571	3	-	71
207	-	107	-	105
208	1586	-	-	992
209	2621	383	-	1

Fig. 2. Evaluation data taken from the MIT-BIH database, where a "Record" refers to a patient.

Class	Normal	PVC	PAC	PJC
Number of samples	60190	6709	2130	83

Fig. 3. Details of the dataset from the MIT-BIH database.

we chose only patients with three conditions, namely premature ventricular contraction beats (PVC) premature arterial contraction beats (PAC) and premature junctional contraction beats (PJC), since they provide the best quality of records, and more specifically PVC is a predictive element of the CA sudden death.

TABLE I
THE VARIOUS DESCRIPTORS AS DISCUSSED IN SECTION II-A.

Attributes	Meaning
RR precedent: $RR0$	Distance between the peak of the current beat R and the previous one
RR next : RRn	Distance between the peak of the present R and the next beat
QRS complex	Beginning of the Q-wave and the end of the S wave
Comp	The ratio $RR0/RRs$
PP	Peak to peak of the R wave of the QRS complex
Energy	Energy of the QRS complex

A. Feature Selection

The feature vector x, which is used for the identification of heart beats, has been selected as follows:

- The R-R interval of the beat RRp, which is calculated as the difference between the QRS peak of the present and previous beat,
- The ratio $r = RR1$-to-RRo. RRn is calculated as the difference between the QRS peak of the present and following beat see Figure 1,
- The QRS width w, which is according to the Tompkins [6].

In this way, each beat is stored as $3-$element vector. In fact, although there are three derivations ($D1, D2$, and $D3$) in the ECG, in our study we have considered just one derivation, namely $D3$. Furthermore, from this derivation we chose six attributes: $RR0$, RRn, QRS, COMP, PP, and Energy. Table I provides the most relevant parameters used in this type of investigation.

B. Data Visualisation

Classifying four classes with only one feature, namely $RR0$, RRS, QRS, etc. is, in general, a complex task due to the lack of threshold for each class. Figures 4, 5, and 6 depict such difficulty in distinguishing between classes according of each features separately. Note that in Figure 5, it is possible to separate between V class and N, which is a promising aspect exploited in our experiment.

C. Correlation study

The aim of this section is to assess the existence of any relationship between the different features. Figure 7 shows the distribution of samples.

In order to evaluate and assess the relationships between all features and the output, we have investigated the corresponding correlations properties. In particular, the calculation of the R parameters is analyzed individually, as well as within the corresponding classes. This has enabled to deduce that the mean feature maximizing the relationships within the corresponding classes are QRS and COMP. This confirms that these two parameters are very important in arrhythmias detection, also suggesting the presence of correlation between features. This is clearly to be expected, as most of them are linked to the cardiac rate and temporal evolution of the signal ECG.

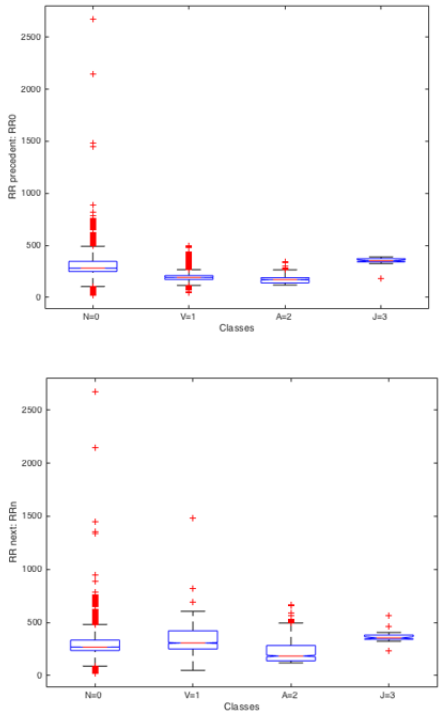

Fig. 4. The distribution of features $RR0$ and RRn in different classes.

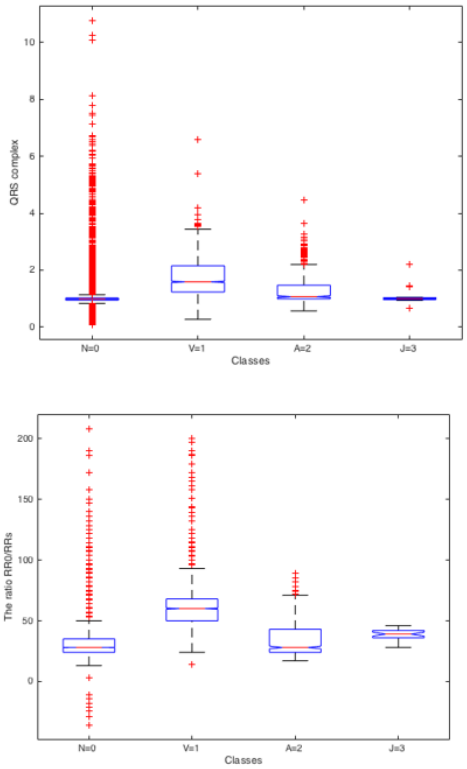

Fig. 5. The distribution of features QRS and COMP in different classes.

III. LOGISTIC REGRESSION

The MLR based supervised learning algorithm aims to design a classifier which is capable of distinguishing K

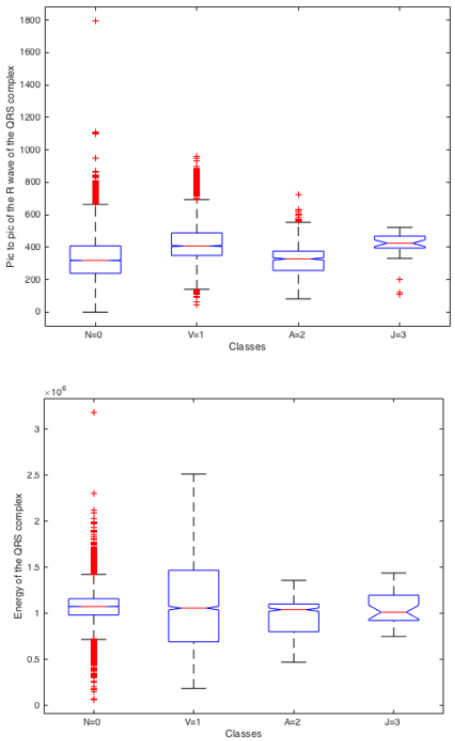

Fig. 6. Distribution of features PP and ENERGY of QRS complex in different classes .

	RR0	RRn	QRS	COMP	PP	ENERGY	CLASS
RR0	1	0.49728	-0.43924	-0.28433	-0.03675	-0.09099	-0.36555
RRN	0.49728	1	0.41501	0.01896	0.05735	-0.03505	0.04703
QRS	-0.43924	0.41501	1	0.32951	0.07565	0.06068	0.41057
COMP	-0.28433	0.01896	0.32951	1	0.17278	0.05531	0.49441
PP	-0.03675	0.05735	0.07565	0.17278	1	0.18565	0.13435
ENERGY	-0.09099	-0.03505	0.06068	0.05531	0.18565	1	0.00306
CLASS	-0.36555	0.04703	0.41057	0.49441	0.13435	0.00306	1

Fig. 7. The correlation coefficients.

classes, using the L labeled training samples, when feature vectors are given as the input for classification [4]. The algorithm involves a training phase and a testing phase. L training samples with known class labels are indicated as $D_L = \{(X_1, Y_1), \ldots, (X_L, Y_L)\}$, which is called the training set. Subsequently, the posterior class distribution using MLR model is computed for MAP estimation of regressors w. The general MLR model is given as

$$P(y_1 = k|x_i, w) = \frac{\exp\left(w^{(k)} x_i\right)}{\sum_{k=1}^{K} \exp\left(w^{(k)} x_i\right)}, \tag{1}$$

where

- W^k is the set of logistic regressors for class k
- w is defined as $(w^{(1)T}, \ldots, w^{(K-1)T})$, where the value of $w^{(K)}$ is generally set to zero. This is due to the fact that the $K-$th conditional probability is found by subtracting the sum of estimated regressors of $(K-1)$ classes from unity.
- Finally, $x = (x_1, \ldots, x_i)$ represents the feature vectors selected for training the model.

In this paper, a Gaussian Radial Basis Function (RBF), which is defined as

$$K(x_i, x_j) = \exp\left(\frac{-||x_i - x_j||^2}{2\sigma^2}\right), \tag{2}$$

represents the training vectors and offers improved data separability in the transformed space [4]. The posterior probability density of w with Y_L (set of labels) and X_L (set of feature vectors in the given labeled training samples) is represented as

$$P(w|Y_L, X_L) \alpha p(Y_L|X_L, w) p(w|X_L) \tag{3}$$

Using Expectation Maximization (EM), the expressions for the MAP estimation of w, which maximizes the conditional log data likelihood, is defined as

$$\hat{w} = \arg max(l(w) + \log p(w|X_L)) \tag{4}$$

where the log-likelihood function of w is expressed using

$$\begin{aligned} l(w) &\equiv \log p(Y_L|X_L, w \equiv \log \prod_{i=1}^{L} P(y_i|x_i, w) \\ &\equiv \sum_{i=1}^{L} x_i^T w^{(y_i)} - \log \sum_{j=1}^{K} \exp\left(x_i^T w^{(j)}\right). \end{aligned} \tag{5}$$

During the testing phase, the estimations of the regression coefficients (w) are entered into the MLR model to find out the posterior class probability densities of each feature vector in the K classes. The class label of a particular feature vector is determined from the index corresponding to the maximum posterior class probability of the given test pixel vector [4], [8].

IV. Results and Evaluation

In this article, we have chosen a hierarchical multinomial model, implemented in Matlab, which was tested on the database described above, and the results are represented by the confusion matrix. This is a tool for measuring the quality of a classification system. Each column of the matrix represents the number of occurrences of an estimated class, while each row represents the number of occurrences of a real class (or reference). Figures 8 and 9 show its main properties and performances criteria.

In particular, all the parameters depicted in Figure 9 are defined as follows

- **CorrectRate:** Correctly Classified Samples / Classified Samples

	N	V	A	J	other
N	42016	7	9	3	0
V	16	3947	3	0	0
A	4	12	692	0	0
J	0	0	3	29	0
Other	0	0	0	1	0

Fig. 8. The confusion matrix.

CorrectRate	**0.9313**
ErrorRate	**0.0687**
LastCorrectRate	**0.9313**
LastErrorRate	**0.0687**
InconclusiveRate	**0**
ClassifiedRate	**1**
Sensitivity	**0.9286**
Specificity	**0.9417**
PositivePredictiveValue	**0.8125**
NegativePredictiveValue	**0.9798**
PositiveLikelihood	**15.9405**
NegativeLikelihood	**0.0758**
Prevalence	**0.2137**

Fig. 9. The performance of LRM.

- **ErrorRate:** Incorrectly Classified Samples / Classified Samples
- **LastCorrectRate:** the following equation applies only to samples considered the last time the classifier performance object was updated. This is Correctly Classified Samples / Classified Samples
- **LastErrorRate:** the following equation applies only to samples considered the last time the classifier performance object was updated, which is Incorrectly Classified Samples / Classified Samples
- **InconclusiveRate:** Nonclassified Samples / Total Number of Samples
- **ClassifiedRate:** Classified Samples / Total Number of Samples
- **Sensitivity:** Correctly Classified Positive Samples / True Positive Samples
- **Specificity:** Correctly Classified Negative Samples / True Negative Samples
- **PositivePredictiveValue:** Correctly Classified Positive Samples / Positive Classified Samples
- **NegativePredictiveValue:** Correctly Classified Negative Samples / Negative Classified Samples
- **PositiveLikelihood:** Sensitivity / (1 - Specificity)
- **NegativeLikelihood:** (1 - Sensitivity) / Specificity
- **Prevalence:** True Positive Samples / Total Number of Samples.

In this work we have adapted the algorithm of MLR methods, so that the output is not only a repartition of probabilities, but a class. A variety of observations have been carried out and we have obtained the classification rate of 93.13%, which is a clear indication that our proposed method is suitable to this type of data. Furthermore, it suggests that it successfully addresses the challenge of the classification of cardiac arrhythmias, with a sensitivity of 92.86% and specificity of 94.17%. We also note that our approach improves the criterion of transparency and interpretability of the process, by positively impacting on the results and the processing of probabilities. Furthermore the readability of the results has also been enhanced, which is an important aspect of the interpretation process carried out by cardiologists expert. As a consequence, it is clear that our approach provides an advantage over other methods of classifications

V. Conclusion

In this paper we have presented a classifier based on Multinomial Logistic Regression (MLR). Currently MLR offers a major advantage in the classification due to their simplicity. In the medical field, experts need automatic diagnostic support to facilitate and justify their decisions, which tends to lack in several techniques cited in the literature in particular neural networks.
The method presented in this paper offers physicians an explicit knowledge base on probability acquired from a medical database. The contribution in the classification process is demonstrated by an accuracy of 93.13%, and our method offers good flexibility and transparency in the system of detection.

In future research, we are aiming to investigate the integration of the approach presented in this paper with fuzzy partition rules [6], to provide an efficient and scalable tool in arrhythmia detection.

References

[1] Yu S. and Chou T., Integration of independent component analysis and neural networks for ECG beat classification, Expert Systems with Applications, Volume 34, Issue 4, 2008

[2] Li, Jun, Jose M. Bioucas-Dias, and Antonio Plaza. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. Geoscience and Remote Sensing, iEEE Transactions on 48.11 (2010): 4085-4098.

[3] Bohning, Dankmar. Multinomial logistic regression algorithm. Annals of the institute of Statistical Mathematics 44.1 (1992): 197-200.

[4] Camps-Valls, Gustavo, and Lorenzo Bruzzone. Kernel-based methods for hyperspectral image classification. Geoscience and Remote SenSing, iEEE Transactions on 43.6 (2005): 1351-1362.

[5] Gacto MJ, Alcala R, Herrera F. Interpretability of linguistic fuzzy rule-based systems: an overview of interpretability measures. Information Sciences 2011; 181(20):43404360. doi:10.1016/j.ins.2011.02.021.

[6] Behadada O, Trovati M, CHIKH MA, and Bessis N. Big data-based extraction of fuzzy partition rules for heart arrhythmia detection: a semi-automated approach Concurrency and Computation: Practice and Experience, 2015

[7] Karthikeyani V and Parvin Begum I. Comparison a Performance of Data Mining Algorithms (CPDMA) in Prediction Of Diabetes Disease, International Journal on Computer Science and Engineering, Vol. 5 Issue 3, 2013

[8] Duda R. and Hart M., Pattern classification and science analysis, John Wiley and Sons, New York. 1973