ECG Classification Using kNN and LDA for Continuous Heart Monitoring

Adrian Ocneanu , Colin Jones , and Andy Adler

Computer and Systems Engineering, Carleton University
{adrian_ ocneanu, colin_ jones}@carleton.ca
adler@sce.carleton.ca

Abstract: In this work, an ECG data representation and encoding schema is investigated. Its aim is to support mobile and continuous heart monitoring, for athletes and cardio-vascular disease (CVD) patients.

For data analysis and encoding, a linear discriminant analysis (LDA) was performed on ECG data capturing several heart conditions, obtained from PhysioNet. On-line performance, in terms of classification of unknown heartbeats, using k-nearest neighbours (kNN), was computed and reported.

We show that such an approach allows for simple, well-established, and robust data classification tools to be deployed. Using this representation schema, we hypothesize there is a potential for cheaper and more user-friendly apparatuses in the market.

1 Introduction

A leading cause of death in Canada, and the world, are the cardio-vascular diseases [1]. Nearly one third of yearly deaths are credited to this [2] and strategies to tackle these diseases go through prevention, treatment and are accompanied, in most cases, by remote and continuous monitoring of the evolution of the heart's condition [4]. As a result, solutions do exist, but commercially available portable ECG monitoring devices cost between $1,000+ and $12,000 [5].

Besides this cost drawback, they are also used for very short periods of time, of 1-2 days, many times requiring the patient to record their symptoms by hand [6].

For athletes' monitoring, on the other hand, things don't look any simpler. The intensity of training and requirements for the heart actually gives room for many heart affections [3] that need close monitoring. Therefore, this turns into a serious health concern, as athletes suffer severe heart problems, sometimes leading to their death [4]. These heart problems may very well pass unnoticed, but they should be detected and treated beforehand.

One reason for this status-quo is the lack of long term biological signals data. Logging long term data from patients or athletes is a step medical or sports training science hasn't benefited from yet. Medicine making is mainly reactive, rather than preventive, and athletes performance is scientifically un-structured and poorly understood.

With these points in mind, we set out to explore a data encoding schema based on pattern recognition tools, that would allow for online classification of patients/athletes heartbeats.

2 Method

We are interested in discriminating between different heart affections coming from an online ECG signal. Our hypothesis is that each heart beat, of a patient suffering from one of these diseases (or healthy, for that matter), will be different than the ones for patients suffering from other diseases.

Our main claim is that different heart diseases will display different PQRS complexes and we can discriminate between them, making this a classic multi-class discrimination problem.

We did the following: (i) using an ECG database, extracted individual heartbeats, using the Pan Tompkins algorithm [11], (ii) performed a linear discriminant analysis (LDA), with 8 distinct classes, and (iii) iterated over a 10-fold cross-validation, using a k nearest neighbours classifier.

Based on our classification results, we claim our hypothesis was confirmed. As a result, we argue that these pre-computed eigen-heartbeats can easily be pre-

loaded on mobile monitoring devices to allow for online classification.

2.1 PhysioNet database

For this study, a dataset from PhysioNet was used, the Physikalisch-Technische Bundesanstalt (PTB) [7]. This database consists of 290 patients that present various heart diseases, as indicated in Table 1. The labels for the patient are the *reason* for which they were admitted in the hospital, most times as part of a routine check-follow-up, rather than presenting also *actual* symptoms, as we understand the dataset's description.

Table 1: PTB Patients

Diagnostic class	**Number of subjects**
Myocardial infarction	148
Cardiomyopathy/Heart failure	18
Bundle branch block	15
Dysrhythmia	14
Myocardial hypertrophy	7
Valvular heart disease	6
Myocarditis	4
Miscellaneous - omitted from analysis	*4*
Healthy controls	52

2.2 Linear Discriminant Analysis

Dimensionality reduction is a requirement for most pattern recognition approaches. A common tool in this sort of representation, when dealing with multiple classes, and discrimination is desired, is linear discriminant analysis (LDA) [8]. LDA, put in simply, finds projection directions that maximize the inter-class distance and minimizes the intra-class distance, as indicated in Figure 1 (from [13]). This is in opposition to principal component analysis (PCA) [9], that attempts to reconcile and best represent all data points, as indicated in Figure 2 (from [12]). If the number of classes to be considered is C, then we are looking for $(C-1)$ projection vectors, ω_i, that can be arranged in a projection matrix, $W = [\omega_1|\omega_2|...|\omega_{C-1}]$:

$$y_i = \omega_i^T x \Rightarrow y = W^T x \tag{1}$$

where x is the input data and y are the coefficients of this projection. The within-class scatter, S_W, can be written as:

$$S_W = \Sigma_{i=1}^{C} S_i \tag{2}$$

where

$$S_i = \Sigma_{x \in \omega_i}(x-\mu_i)(x-\mu_i)^T$$

$$\mu_i = \frac{1}{N_i}\Sigma_{x \in \omega_i} x$$

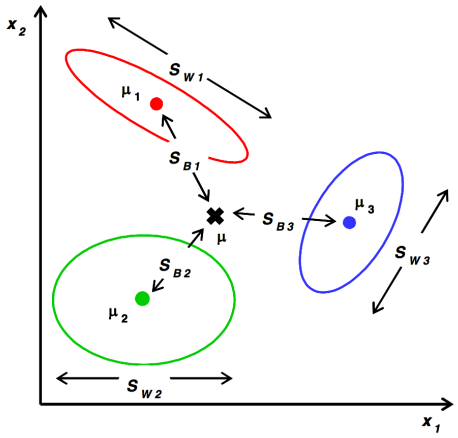

Figure 1: Multiclass Linear Discriminant Analysis (LDA)

The between-class scatter, S_B, can be written as:

$$S_B = \Sigma_{i=1}^{C} N_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{3}$$

where

$$\mu = \frac{1}{N}\Sigma_{\forall x} x = \frac{1}{N}\Sigma_{i=1}^{C} N_i \mu_i$$

Combining the two, and recalling we are looking to maximize the between-class to within-class scatter ratio, we can write a new objective function, which needs to be maximized:

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|} \tag{4}$$

The solution of this equation is a matrix W^* whose columns are the highest value eigenvectors coming from the following generalized eigenvalue problem:

$$W^* = [\omega_1^*|\omega_2^*|...|\omega_{C-1}^*] = argmax\frac{|W^T S_B W|}{|W^T S_W W|} \Rightarrow$$
$$\Rightarrow (S_B - \lambda_i S_W)\omega_i^* = 0$$

It follows that only $(C-1)$ of the λ_i values will be non-zero, and that the directions of maximum class separability are the eigenvectors corresponding to the highest eigenvalues of $S_W^{-1} S_B$.

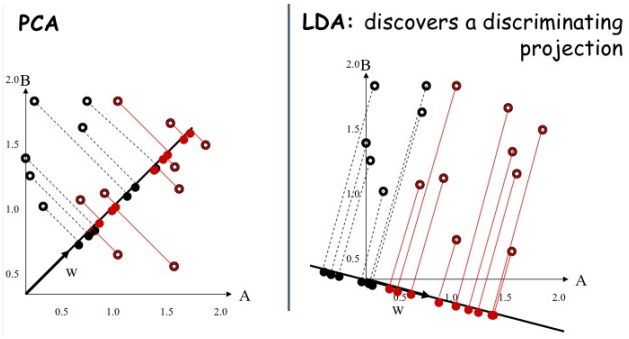

Figure 2: PCA vs. LDA

2.3 Online Classifier based on k-Nearest Neighbours

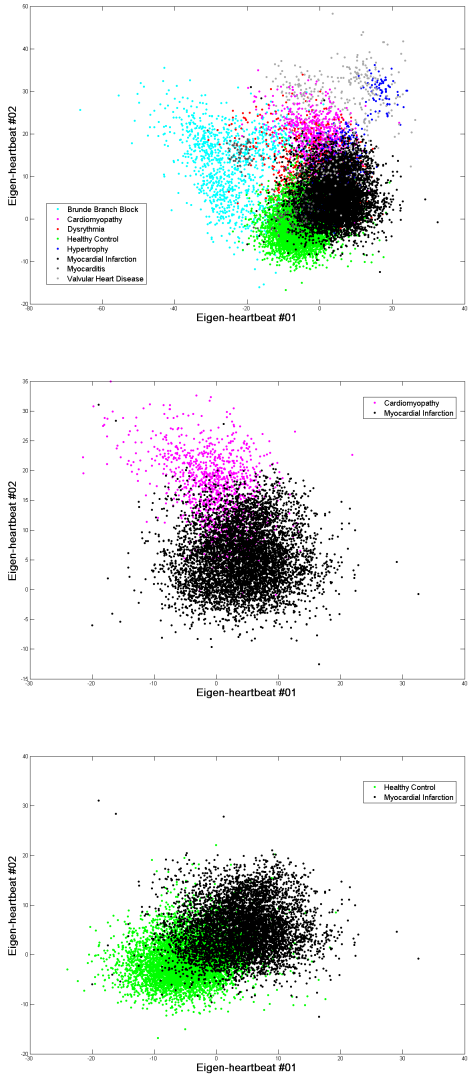

Figure 3: LDA results with emphasis on classes overlapping (represented using the top 2 eigen-heartbeats)

Using the representation mentioned in Section 2.2, an online classifier of heartbeats is proposed. This means that an incoming ECG signal can be projected, once it's aligned (using an aligning algorithm, like the Pan/Tompkins one), or continuously (shifted time-wise, with each incoming sample), onto this vector space, and this representation can be coupled with a simple k-nearest neighbour voting schema.
An issue with such an approach is not knowing the best number of k to use, therefore a parameter search was performed. This was simply done by varying k and the error rates for classification were computed, as shown in Section 3.

3 Results

The LDA performed on the ECG data provides the eigen-vectors, or bases for signal projections. These can be used for the representation of incoming online signals and their classification, beat by beat, as belonging to a certain type of heart condition (healthy condition included).

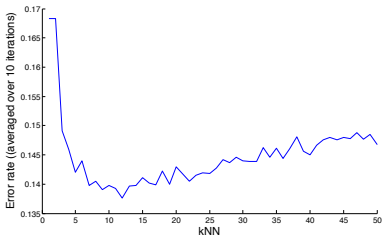

Figure 4: Error rates over a range of *kNN* values

To validate the robustness of this approach, a 10-fold validation was performed on the data. For each 9/10 training subset, error rates were computed while attempting to classify the remaining 1/10 testing subset. These error rates were averaged over the 10 iterations and assigned as the errors specific to the particular value of k used in the process.
k was changed from 1 to 50 and the results can be seen in Figure 4.
As several classes suffer from the overlap mentioned in Section 2.1, a confusion matrix was computed and the result presented in Table 2.

Table 2: Confusion Matrix

Classification \ True Class	B. B. Block	Cardiomyopathy	Dysrhythmia	Healthy Controls	M. Hypertrophy	M. Infarction	Myocarditis	V. H. Disease
B. B. Block	53247	800	694	6479	1	371	64	44
Cardiomyopathy	729	48691	1436	851	265	6955	59	214
Dysrhythmia	2201	1424	22413	1517	59	6974	5	107
H. Controls	683	1325	227	348549	748	27738	412	168
M. Hypertrophy	0	143	41	485	24704	3008	3	16
M.Infarction	460	5479	1777	47084	1071	250826	219	384
Myocarditis	77	31	8	2114	4	764	9052	0
V. H. Disease	98	1775	546	111	596	1532	93	12249

Note that the confusion matrix is not and doesn't have to be symmetrical. In fact, there are 2 cases where there is no obvious correspondence, as indicated in yellow, in Table 2. Firstly, class 6 (myocardial infarction) is mistaken for class 3 (dysrhythmia) but not the other way around. Secondly, the same goes for some elements in class 4 (healthy control) that get misclassified as being members of class 1 (brunde branch block), but not the other way around.

Still, the most common source for error, it would seem, is the mutual misclassification of class 6 (myocardial infarction), on one hand, for class 2 (cardiomyopathy), and class 4 (healthy control), on the other hand, as identified in green in Table 2.

4 Conclusions

We performed a pattern recognition analysis, using the PTB PhysioNet dataset. We conclude that the classification results, coupled with the potential low computational costs of using pre-computed eigen-heartbeats, make this approach Online classification of signals was proven possible and the error rate, with the simple kNN framework used, rounds the 0.1375, for a kNN of 12.
We also discovered that, for certain patients (as far as the labelling given under PhysioNet is concerned), their heartbeats seem healthy. This might be a result of the labeling method, since the patients were in many cases undergoing routine check-ups, after having recovered (partially or fully) from a previously existent condition. Having identified other applications of a full-body monitoring system for athletes and patients, it is the intention for future work to proceed with a more in depth analysis of performance and health conditions tracking over extended periods of time, and the discovery of meaningful trends and patterns that only this sort of historic data would be able to reveal.
We are actively working towards such a monitoring device, based on the omnipresent smart phone (an Android, in our particular case), as the main collection, handling and transmission of information hub. We are currently able to collect, store and analyze 48+ hours long ECG data sets, in one collection session (stopping for battery replacement/recharging). Our goal is to integrate other biological signals into this long term sensing infrastructure.

References

[1] Heart and Stroke Foundation, What is heart disease?, [Online] 2012, http://www.heartandstroke.com/site/c.ikIQLcMWJtE/b.3682421/k.48B2/Heart_disease__What_is_heart_disease.htm, Accessed November 2012

[2] Heart and Stroke Foundation, Statistics on Heart Disease, [Online] 2012, http://www.heartandstroke.on.ca/site/c.pvI3IeNWJwE/b.3581729/k.359A/Statistics.htm#heartdisease, Accessed November 2012

[3] Hanne-Paparo N, Kellermann JJ, "Long-term Holter ECG monitoring of athletes" Med Sci Sports Exerc, 13(5):294-8, 1981.

[4] The Globe and Mail, Why are high-performance athletes having heart attacks?, [Online] 2012, http://www.theglobeandmail.com/life/health-and-fitness/health/conditions/why-are-high-performance-athletes-having-heart-attacks/article4100522/, Accessed November 2012

[5] John Hopkins Medicine, Holter Monitor, [Online] 2012, http://www.hopkinsmedicine.org/health library/test_procedures/cardiovascular/holter_monitor_92,P07976/, Accessed November 2012

[6] NIH Heart, Lung and Blood Institute, What Are Holter and Event Monitors? [Online] 2012, http://www.nhlbi.nih.gov/health/health-topics/topics/holt/printall-index.html, Accessed November 2012

[7] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101 (23): e 215-e 220 [Circulation Electronic Pages] 2000 (June 13).

[8] Fisher, R. A. (1936)."The Use of Multiple Measurements in Taxonomic Problems". Annals of Eugenics

[9] Karhunen, Kari (1947). "ber lineare Methoden in der Wahrscheinlichkeitsrechnung". Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys. 37: 179.

[10] Sung-Nien Yu and Ying-Hsiang Chen. 2009. Noise-tolerant electrocardiogram beat classification based on higher order statistics of subband components. Artif. Intell. Med. 46, 2 (June 2009), 165-178.

[11] Pan J and Tompkins WJ. A Real-Time QRS Detection Algorithm. IEEE Transactions on Biomedical Engineering 32(3):230-236, 1985.

[12] West Virginia University, Data Mining: Fall'06, [Online] 2012, http://csee.wvu.edu/ timm/cs591o/old/FSS.html, Accessed March 2013

[13] Texas A&M University, CSCE 666 Pattern Analysis, [Online] 2013 http://research.cs.tamu.edu/prism/lectures/pr/pr_l10.pdf, Accessed March 2013