Using Hidden Markov Models and Spark to Mine ECG Data

Jamie O'Brien
Saint Mary's College of Maryland
St. Mary's City, Maryland
Email: jcobrien@smcm.edu

***Abstract*—New potential risk factors for cardioembolic strokes are being considered in the medical community. The presence of these factors can be determined by reading an electrocradiogram (ECG). Manual ECG analysis can take hours. We propose combining accurate Hidden Markov Model (HMM) techniques with Apache Spark to improve the speed of ECG analysis. The potential exists for developing a fast classifer for these risk factors.**

I. Introduction

The proliferation of medical data in modern hospitals provides a rich environment for data mining. Electrocardiograms (ECGs) provide a wealth of information that can be used to diagnose cardiovascular diseases (CVDs). In Agarwal and Soliman [1], it is suggested that the ECG can be used to detect cardioembolic stroke risk factors. Aside from those factors included in the Framingham Risk Score, emerging factors include:

1) cardiac electrical/structural remodeling,
2) higher automaticity,
3) heart rate & heart rate variability.

Currently, the manual analysis of ECG patterns is time-consuming. It can take several hours to complete Acharya et al [2].

II. Problem Statement

We want to find a better method of detecting the emerging risk factors listed in Section I. We want to combine an effective Hidden Markov Model (HMM) classifier for ECGs with the fast, distributed processing power of Apache Spark.

A. Atrial Fibrillation—A Verified Stroke Risk

In atrial fibrillation (AF), the heart's atrial walls do not produce an organized contraction—instead, they quiver [3]. Even though AF is a component of the Framininghamm Stroke Risk Score [4], it is often undetected; the condition has evaded detection even in patients known to have paroxysmal atrial fibrilation. The detection rates may vary depending on the algorithms used, but seem to improve with longer monitoring times [5]. The difficulty of accurately detecting AF motivates the search for additional stroke risk factors.

B. Hidden Markov Models

Hidden Markov Models (HMMs) have been used with great effect in classifying ECGs. Andereão et al were able to demonstrate an accuracy of 99.97% in detecting the QRS complex of the heartbeat [6]. Their approach was to create a general model of the heartbeat, and then tune the model to each individual by using data from the first 20 seconds of their ECG. The general model of the heartbeat was composed of discrete states representing the P, Q, R, S, and T waves, the PQ and ST intervals, and the isoline. Andreão et al's work was able to detect premature ventricular contractions (PVCs). We hope to use a similar model for detecting ectopic beats and bundle blocks.

C. Apache Hadoop

Hadoop pairs a high-bandwidth distrubted file system with MapReduce programming Svachko et al [7]. This allows for a task to be broken up across many computers, the components calculated independently, and the results collected. In this way, Hadoop may improve the performance of signal processing tasks. This performance improvement is the core of the Cloudwave system described in Jayapandian et al [8]. The authors of that work used Hadoop to process multimodal bioinformatic data. A stand-alone machine was able to process 10 signals in 22-36 minutes. Their Hadoop cluster was able to process the same data in 4-6 minutes.

D. Apache Spark as a replacement for Hadoop MapReduce

While the Cloudwave system described in Jayapandian et al [8] is impressive, the highly iterative nature of data mining tasks may cause significant overhead under Hadoop's MapReduce architecture. Apache Spark avoids this issue by using the concept of resilient distributed datasets (RDDs). These RDDs can be cached in memory. This make the data available for iterative and parallel programming alike without having to be constantly reloaded Zaharia et al [9].

III. Method

The in-progress research explores the applicability of Hidden Markov Models on ECG readings, with the goal of detecting the emerging factors mentioned in [1]. Here we note the strategy for constructing our system.

We obtained ECG signals from the QT Database (QTDB), using the WaveForm Database application suite. We also

obtained two sets of annotations: one, marked `atr`, contains annotations that marks beats as normal, or as having some abnormality (preventricular contraction, for instance); the second set of annotations, marked `pu0`, contains waveform markers, such as p, t, and N (for normal qrs complex). Any records from the QTDB that did not contain annotations from `atr` were excluded, as we would not be able to verify our results against them.

We transformed the `pu0` annotations to provide clearer information. The standard for annotating waves is to open a wave with a paren, note the wave, and then close it with a paren. For instance, the p wave would be marked by the annotations (, p,). We wrote a script to process these annotations, and change them to the form pBegin, p, pEnd, so that all parenthesis were removed. This meant that the annotations themselves could now become a set of states for use in a Hidden Markov Model. The states derived from the annotations were: pBegin, p, pEnd, q, r, q, tBegin, t, tEnd, unknownBegin, and unknownEnd.

However, we found that it was not practical to simply map the states annotated in `pu0` to the beat classifications annotated in `atr`. When attempting to map the state sequence to PVC, for instance, no significant correlation could be found in a sample of PVC beats. We hypothesized that the duration of the states was also significant. It may be necessary to mark states as being faster or slower than normal. The duration between, for instance, pBegin and pEnd could tell us if the p wave were of normal duration.

With this in mind, we are determining a way to map the ECG signal itself to states. In [10], we find an algorithm for decomposing ECG signals into line segments. This algorithm moves a dynamically-sized window along the ECG signal. The window checks the distance between the endpoints and every point in-between, using normalized distances where needed. We can adjust the allowed error to accomodate noisy signals.

We modify this algorithm to output a list of 4-tuples of the form (starting point, length, mean of segment, standard deviation of segment). This converts the continuous ECG signal into a set of data points. We must then convert this set of data points into states that correspond with the waveforms of the heart beat: the p wave, qrs complex, t wave, and the intervals between them.

IV. The Classification Process

We begin by slicing an ECG signal between its R-R intervals. We then take a slice and segment it using the algorithm described in [10]. These segments are then labeled by the state they most match, using a decision tree. The progression of states is treated as an observation, and fed into the HMM to determine which beat type most accurately matches the observation.

V. Further Work

This work will not be complete until the HMM itself is built and can be tested. In anticipation of this, we have separated the QTDB into a training set comprising approximately 80% of the annotated data, and a testing set with the remaining approximately 20%. The training set is composed of five sub-groups, each approximately 20% of the size of the training set. We intend to use these sub-groups for cross-validation.

After the model is built and its performance is evaluated, we can begin the construction of the Apache Spark implementation of the model. The purpose of this will be to compare the performance of the Spark implementation against the locally-run implementation. The parameters for this experiment will be determined when the HMM itself is complete.

VI. Conclusion

This research may provide a effective method for detecting the emerging risk factors for a cardioembolic stroke mentioned in section I. This would assist researchers who are investigating these risk factors.

Acknowledgment

We would like to thank the National Science Foundation (NSF) for their generous grant, and the University of Colorado, Colorado Springs for hosting the Research Experience for Undergrads (REU) program.

References

[1] S. Argwal and E. Soliman, "Ecg abnormalities and stroke incidence," 2013. [Online]. Available: http://www.medscape.com/viewarticle/808752

[2] R. Acharya, A. Kumar, P. Bhat, C. Lim, N. Kannathal, and S. Krishnan, "Classification of cardiac abnormalities using heart rate signals," *Medical and Biological Engineering and Computing*, vol. 42, no. 3, pp. 288–293, 2004.

[3] F. H. Martini, J. L. Nath, and E. F. Bartholomew, *Fundamentals of Anatomy and Physiology (9th Edition)*. Benjamin Cummings, 1 2011.

[4] F. H. Study, "Stroke," https://www.framinghamheartstudy.org/risk-functions/stroke/stroke.php, (Visited on 07/14/2014).

[5] M. A. Rosenberg, M. Samuel, A. Thosani, and P. J. Zimetbaum, "Use of a noninvasive continuous monitoring device in the management of atrial fibrillation: a pilot study," *Pacing and Clinical Electrophysiology*, vol. 36, no. 3, pp. 328–333, 2013.

[6] R. V. Andreão, B. Dorizzi, and J. Boudy, "Ecg signal analysis through hidden markov models," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 8, pp. 1541–1549, 2006.

[7] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*. IEEE, 2010, pp. 1–10.

[8] C. P. Jayapandian, C.-H. Chen, A. Bozorgi, S. D. Lhatoo, G.-Q. Zhang, and S. S. Sahoo, "Cloudwave: Distributed processing of big data from electrophysiological recordings for epilepsy clinical research using hadoop," in *AMIA Annual Symposium Proceedings*, vol. 2013. American Medical Informatics Association, 2013, p. 691.

[9] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: cluster computing with working sets," in *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, 2010, pp. 10–10.

[10] A. Koski, "Modelling ecg signals with hidden markov models," *Artificial intelligence in medicine*, vol. 8, no. 5, pp. 453–471, 1996.