Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 6

Due: March 29, 2023(100 points)

Haoran Zhang, Chenjie Wu, Qishu Dong
GitHub repo: hrcheung/arrhythmia-detection
Email: zhang.haoran1@northeastern.edu

Your assignment for this week will be to send us a 2-page project proposal. The Google Docs version of this proposal is available here. You can choose the format you want to use, but make sure to include the following information to your document:

1. What problem are you going to be tackling on your project?
2. Why is that an interesting/useful application of data mining?
3. What models/techniques (clustering/classification/etc.) are you envisioning to apply?
4. Where are you going to get the data?

To complete your proposal submission, create your project's repository in your own Github namespace, and upload your provide your URL in Gradescope. For example, my Git handle is kni-neu, and my project repository is:

https://github.com/kni-neu/project

You can use my template or any other template that you might find appropriate. (Or, it is acceptable to use no template at all.) An example set of a set of organized proposal sections is shown below.

1 Problem

Make sure you make clear what problem are you going to solve in as concise and unambiguous manner. This section is typically 2-3 sentences.

We are going to develop algorithms and train accurate models for automated arrhythmia detection, helping doctors diagnose and intervene arrythmia in the earlier stage.

2 Background

This is where you tell us why this solving this problem is important. What will people be able to do once you've solved this problem? How could it conceivably help people? What makes this application of data mining useful?

Arrhythmias are a type of heart rhythm disorder that can have serious consequences, including heart failure, stroke, and sudden cardiac arrest. Accurate and timely diagnosis is critical to ensure appropriate treatment and management of the condition. Automated arrhythmia detection algorithms can help healthcare professionals to diagnose arrhythmias quickly and accurately, even in cases where the condition is asymptomatic. This can lead to earlier arrythmia diagnosis, intervention, and patient treatment.

3. Approach

Here, you can start detailing some specifics. Be sure to cover:

- Where are you going to get data?
- What data mining techniques will you use?

We are using MIT-BIH arrhythmia database from PhysioNet. It contains 48 half-hour ECG recordings, with over 110,000 annotations. There were 47 subjects studied by the BIH Arrhythmia Laboratory between 1975 and 1979. All the patients (records) were from Boston's Beth Israel Hospital.

Feature Engineering
We conduct 3-step feature engineering on the raw data. First, we do the data inspection. This database contains two sets of data: normal (with 4046 entries) and abnormal (with 10506 entries). Each dataset consists of 187 features, with the last column indicating whether the record is normal or abnormal. Secondly, we handle duplicate records and missing values. Although the dataset is relatively clean with minimal errors, we will fill any missing values with zero as a precautionary measure. Third, data aggregation and shuffling. We combine normal and abnormal data and shuffle them to ensure randomness in the data for training purposes. By meticulously engineering features from the MIT-BIH Arrhythmia Database, we aim to enhance the performance of automated arrhythmia detection algorithms.

Modelling
For modelling part, this problem can be considered as a supervised anomaly detection question. We have the label N denoting normal heart rate and A denoting abnormal data.

Potential models include Support Vector Machine, Decision Tree, Random Forest and Neural Networks. We will start with Random Forest because this model takes advantage of multiple decision trees and it is usually more interpretable, compared to deep learning models. SVM is another valid option because it is easier to visualize the result with hyperplane.

We have considered transformers as a stretch step to improve accuracy. More details are coming soon.