Machine Learning Project 2013-2014 (Oct)

Analyzing runners profiles

1 Introduction

Detecting changes in a runner's movement patterns is an active area of research which plays a major role in automated health care applications. It is useful, for example, to detect muscular fatigue which is prevalent among runners and is normally detected only after the injury of the muscle. Hence detecting them before the onset helps avoid unnecessary strain and injuries in the muscle. Since fatigue depends on so many factors like flooring condition, prolonged activities, repetitive action etc, there is no way to measure it directly. The recent advancement in wearable sensor technology, however, laid the way for gathering data from more different aspects of human activity. Ideally, these sensors can be linked to computational intelligence and used to build fatigue detection systems to promote a healthier lifestyle. For this to be possible it is necessary to apply machine learning techniques to detect fatigue by combine a variety of measurements and derived features [2, 3].

The system used here is developed by the Human Movement Biomechanics research group of KU Leuven[1] and acquires and processes time series data from an ankle mounted tri-axial accelerometer sensor. The data is collected in their training facility in the context of research towards detecting fatigue.

2 Problem and Approach

We want to infer two things when a runner's data is presented: (1) On what surface a person was running and (2) whether it is a trained runner or not. Eventually, it will be necessary to predict these two targets simultaniously because they influence each other (e.g. a bigger impact on the foot can be caused both by a hard surface and the higher speed of a trained runner). But we focus first on the subtasks to simplify the analysis. In this project we want to learn three classifiers:

1. Type of running surface: *asphalt*, *track* or *woodchip*.

   Detect on what kind of surface a person is running. This serves as a basic building block in a fatigue detection system.

2. Type of runner: *trained* or *nottrained*.

   Train a classifier that can detect whether it is a trained runner or not.

3. Type of runner and surface: The six combinations of type of runner and surface.

The first two classifiers are traditional classifiers where there is one single output class. The last classifier can be seen as a traditional classifier with 6 possible classes or you could tackle this as a multi-label classification problem (also called multi-target classification). This problem is more difficult because the number of combinations for these classes is exponential in the number of classes. Also the different tags are possibly correlated and therefore you cannot always consider the multi-label classification problem as simply a bag of single-label classification problems.

As input data, you are given a set of raw accelerometer data measured on the ankle and hip during 132 sessions with 15 runners. Each of these sessions is labeled with the type of runner and the type of surface.

[1] http://gbiomed.kuleuven.be/english/research/50000737

In order to use raw accelerometer data for classification techniques like Naive Bayes, SVMs or Decicion trees it is often necessary to first extract features from the data (see Albert et al. [1] for inspiration). For this particular problem the most informative features are derived from the peaks present in te data caused by the impact with the ground. From the peaks you can derive the magnitude of the impact, time between steps, etc. Included with this assignment you can find Python code to extract the peaks from the raw data (see `HowToExtractPeaks.html` for more information). Be aware that deriving speed is non-trivial although some solutions have been presented [5, 6, 4].

3 Tasks

3.1 Form Groups **By October 25th**

You can work on this project in groups of two and divide the work. Form the groups as quickly as possible and email the names of the members of your group to davide.nitti@cs.kuleuven.be. You work on this assignment alone or in team, do not share your experimental results or implementations with your fellow students.

3.2 Literature Study

Familiarize yourself with the concept of machine learning from accelerometer data. Get a feeling for what has been done before in this context, and what are the open questions. Divide the work between the two people in the team.

3.3 First Report **By November 8th**

In a first phase of the project you explain in a brief report (about 2 pages) how you will address these tasks. Describe what literature you have read and what you have learned from it. Give examples of questions that you want to find answers for, and give an idea of how you intend to find the answers (What data representation, features and machine learning techniques will you use? How will you evaluate your classifier?). Please send your report in PDF format to the aforementioned email address. The quality of this report influences your final score for the project, so do your best to come up with a good plan.
After the reports are handed in, you will get feedback on your report, and, where necessary, we may give more concrete guidelines on how to proceed.

3.4 Machine Learning

Use at least two different machine learning techniques to classify the data. Make sure that you use a correct experimental setup to assess the accuracy of the classifier (e.g. cross-validation, test/train split).
You are not required to use the entire time range of the measurements for learning since there might be periods of time where no running happens. Also at the start (end) of the measurement a runner might be accelerating (decelerating). It is advisable to remove those parts during preprocessing the data.

3.5 Analyze results

What is the accuracy you can achieve? Do you understand why your parameters work? What is the computational cost of pre-processing, learning and classifying?

3.6 Final Report and Prototype **By December 13th**

Report: Write a report with your findings.
Hand in your final report, describing details of your approach and the results obtained:

- Clearly state the **questions** you address with your research
- Describe how you try to **answer** them (what experiments were performed, how do you measure success, etc.),
- Write out the **conclusions** you draw from your experiments together with a scientifically supported motivation for these conclusions. Such a motivation should include descriptions of used techniques and performed experiments and a report and discussion of the results of these experiments. Be careful to report concrete numbers and custom formulas (avoid conclusions using *weasel words*).
- Report the total **time you spent** on the project, and how it was divided over the different tasks mentioned.

The total length of your report should be at most 10 pages. In addition, you should hand in any software you have written to complete the task. If you have created an actual online implementation, also provide the url of your web application.

Prototype: Include the code you used to perform the experiments. By default, your prototype should use the classifier you have found to be the best. We will test unseen examples using your prototype to assess the accuracy. You are free to use any mainstream programming language and any publicly available toolbox. Make sure that:

- The code is intuitive to run and includes a short README file with the necessary steps.
- Your code is self-contained and includes all dependencies.
- Your code is print-friendly (e.g. max 80 columns).
- It is easy to input a path to a new file and let your software print the classification (this will be used to test your software on an unseen runner).

3.7 Peer assessment **By December 13th, individually**

Send by email a peer-assessment of your partner's efforts. This should be done on a scale from 0-4 where 0 means "I did all the work", 2 means "I and my partner did about the same effort", and 4 means "My partner did all the work". Add a short motivation to clarify your score. This information is used only by the professor and his assistants and is not communicated further.

3.8 Discussion **By December 19-20**

There will be an oral discussion of your project report in week 13 where you will also get the chance to demo your prototype.

Time

The time allocated for completing this project is **60 hours per person**. This *does not* include studying the Machine Learning course material, but it *does* include the literature study specific for this project's topic.

Questions

Please direct any questions that you may have about the project to the Toledo forum.

Good luck!

References

[1] Mark V. Albert et al. "Fall Classification by Machine Learning Using Mobile Phones". In: *PLoS ONE* 7.5 (May 2012), e36556.

[2] Mohamed R Al-Mulla, Francisco Sepulveda, and Martin Colley. "A review of non-invasive techniques to detect and predict localised muscle fatigue". In: *Sensors* 11.4 (2011), pp. 3545–3594.

[3] Stephen J Preece et al. "A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data". In: *Biomedical Engineering, IEEE Transactions on* 56.3 (2009), pp. 871–879.

[4] David Sachs. *Sensor Fusion on Android Devices: A Revolution in Motion Processing*. 2010. URL: http://www.youtube.com/watch?v=C7JQ7Rpwn2k.

[5] Kurt Seifert and Oscar Camacho. *Implementing Positioning Algorithms Using Accelerometers*. Application Note AN3397, Rev 0. Freescale Semiconductor, Feb. 2007.

[6] Oliver Woodman. "Pedestrian Localisation for Indoor Environments". PhD thesis. University of Cambridge, Computer Laboratory, Sept. 2010.