Detection of human alertness using supervised learning

Anders Hørsted

Kongens Lyngby 2011
IMM-PHD-2011-70

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

This is the summary/abstract

Resumé

På dansk

Preface

This thesis was prepared at Informatics Mathematical Modelling, the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The thesis deals with different aspects of mathematical modeling of systems using data and partial knowledge about the structure of the systems. The main focus is on extensions of non-parametric methods, but also stochastic differential equations and neural networks are considered.

The thesis consists of a summary report and a collection of ten research papers written during the period 1996–1999, and elsewhere published.

Lyngby, December 1999

Henrik Aalborg Nielsen

Acknowledgements

I thank my...

Contents

Chapter 1

Data

In this chapter the main data set used to create the classifier is introduced. The first section describes the general structure of the data set and then a section about the initial data cleanup follows. After the cleanup section comes the main section about feature exploration. Finally a section about feature transformations finishes the chapter.

1.1 Introducing the data

The data is available through the kaggle website [4] and consists of measurements for 500 *trials*[1]. For each trial 2 minutes of sequential data is recorded and the interval between two recordings of a *row* is 100 ms, giving approximately 1200 rows for each trial. For each row 33 features are measured. The 33 features are:

- TrialID - The trial id. Is zero based.
- ObsNum - The observation number within the trial. Is zero based, so a row with ObsNum=33 is the 34th row in the trial.

[1] On the kaggle website there is two data sets available. One training data set (500 trials) and one test data set (100 trials). Since the competition is finished and it isn't possible to get a test data set with the IsAlert-feature included, only the training data set is used.

- IsAlert - A binary variable describing whether the driver was alert or not. This is the feature that should be classified.
- P1-P8 – Eight physiological features.
- E1-E11 – Eleven environmental features.
- V1-V11 – Eleven vehicular features.

The spokes person from Ford has repeatedly [1, 2] denied to disclose any additional information about the features, so nothing is known about what the different features represents, apart from the feature type (physiological, environmental, vehicular). Also nothing is known about the data type of the features (nominal, ordinal, ratio etc.).

1.2 Introductory data cleaning

In this section the first data cleaning is done. [3]

Bibliography

[1] Mahmoud Abou-Nasr. Kaggle forum: Discrete or continuous values, 2011. URL http://www.kaggle.com/c/stayalert/forums/t/266/discrete-or-continuous-values/1635#post1635.

[2] Mahmoud Abou-Nasr. Kaggle forum: About the parameter, 2011. URL http://www.kaggle.com/c/stayalert/forums/t/317/about-the-parameter/1861#post1861.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science + Business Media LLC, 2006.

[4] Kaggle.com. The ford challenge data files, 2011. URL http://www.kaggle.com/c/stayalert/Data.