SIEMENS

May 2012

**Proposal for a Master Thesis in Statistics**

# Novelty Detection Methods in High-Dimensional Spaces With An Application On Spectral Data

Novelty detection involves identifying new or unknown data that a machine learning system is not aware of during training. For example, it can be used for modeling the normal behavior of a system, enabling the detection of any divergence from normality.

Another common application of novelty detection is to safeguard data-driven models, e.g. neural networks, against undefined extrapolation behavior. Learning a neural network model always means minimizing some error function defined over a finite training set. It is expected that the model will be most accurate in regions of input space for which the density is high since these regions have the most influence on the error function. So if some test point falls into a region with high density then the model is likely to perform well. If it falls into a region with low density (e.g. an outlier) then the model will likely perform poorly and the predicted output should be rejected if the confidence is below some threshold. Therefore, density estimation plays a prominent role in novelty detection.

In many applications, like spectroscopy, microarray analysis or image processing, the input data is high-dimensional which considerably hinders density estimation – a problem often cited as the curse-of-dimensionality.

A common remedy here is to use dimensionality reduction techniques, e.g. principal component analysis, before applying density estimation. Other prominent approaches for dealing with the high dimensionality involve kernel methods.

In this thesis, different novelty detection methods are to be investigated with respect to spectral data. Spectral data arises from spectroscopic measurements where a sample to be analyzed is illuminated in some frequency range and the reflected intensity is measured over several hundreds or thousands of wavelengths. Thus, the resulting measurements are described as inputs in a high-dimensional space and suitable dimensionality reductions techniques have to be developed.

Implementation will be done in Matlab, so prior experience with Matlab or Octave programming is a prerequisite.

A sound background in Machine Learning and/or Statistics is required, specifically multivariate analysis methods. Experience with spectral data is not necessary.

For further information please contact:

Dr. Clemens Otte, clemens.otte@siemens.com, T. 089 636 44246
Prof. Dr. Thomas Runkler, thomas.runkler@siemens.com

LMU Mentors:

Prof. Gerhard Tutz
Dr. Fabian Scheipl (Room 239) fabian.scheipl@stat.uni-muenchen.de, T 089 2180 2248

References:

Books
- Hastie, Tibshirani, Friedman; *The Elements of Statistical Learning*, Springer (2009)
- Theodoridis, Koutroumbas; *Pattern Recognition,* Academic Press (2008)
- Izenman; *Modern Multivariate Statistical Techniques - Regression, Classification, and Manifold Learning*, Springer (2008)

Papers:
- Markou, Singh; *Novelty detection: A review, part 1: Statistical approaches*, Signal Processing 83, 2481–2497, (2003)
- Markou, Singh; *Novelty detection: A review, part 2: Neural network based approaches*, Signal Processing 83, 2499-2521, (2003)
- Rosipal, Krämer; *Overview and Recent Advances in Partial Least Squares,* SLSFS 2005, LNCS 3940, pp. 34–51, (2006)