Testing NIR data processing system with genetic algorithms

Timo Mantere

University of Vaasa

Department of Electrical Engineering and Automation

P.O. Box 700

FIN-53851 Vaasa, Finland

timan@uwasa.fi

http://www.uwasa.fi/~timan

The Research Motive

The fast and accurate diagnosis of melanoma is highly important

It is estimated that one out of every 70 persons will get melanoma during their lifetime

The earlier studies shows, that there exists spectral regions of NIR data that contain the information needed for doing melanoma diagnosis

The Motive behind GA Testing

The original motivation for generating simulated data with genetic algorithms for testing classification system was that we had only a limited number of skin spectra data; samples of skin cancer and samples of healthy skin

The more wide testing the performance and reliability of every system is always important, so we decide to test the system with simulated data and also try to find system boundaries and weaknesses

We also have earlier experience of testing the software and the imaging systems with GA generated simulated data's

Research Questions?

Can the NIR spectra data of skin cancer be effectively simulated with genetic algorithms?

Is the simulated NIR data too similar with real data or always detachable from the real data?

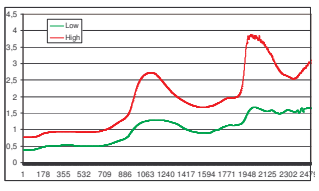
Can the simulated data give us better understanding of NIR data classification boundaries, and reveal the cases of over learning with teaching data?

The Method

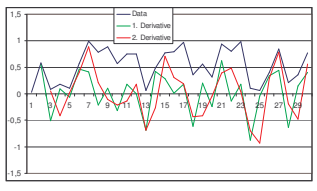
We are generating simulated NIR data by using the original data as boundary

1. The envelope curve of original data
2. The 1. derivative of original data
3. The 2. and higher derivatives of original data

Original data is grouped, so simulated data is generated using the properties of different groups as boundaries

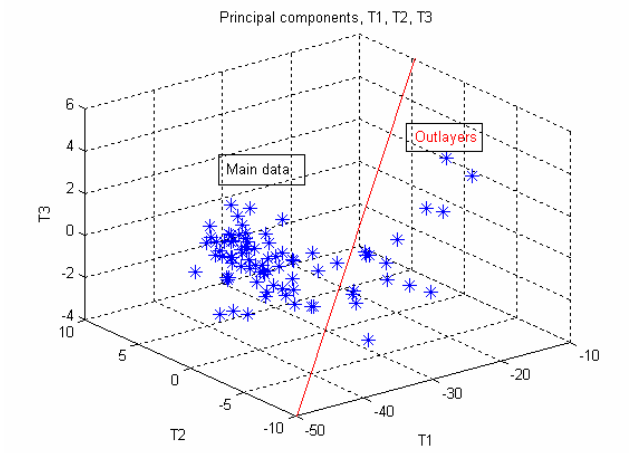


Envelope curve of original data



Example of 1. & 2. derivatives of data serie (random numbers between [0, 1])

Original Data

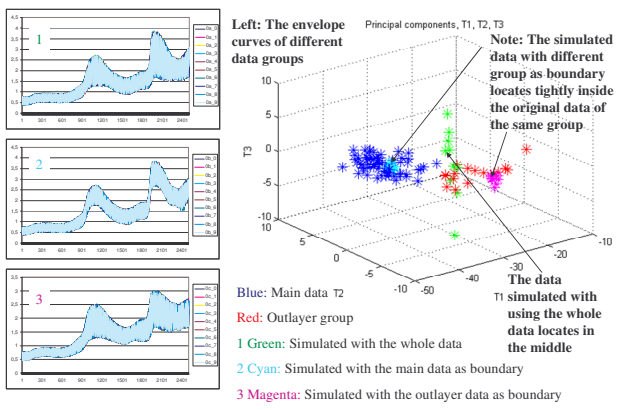


Original data contains outlayers as can easily be seen from the ICA graph

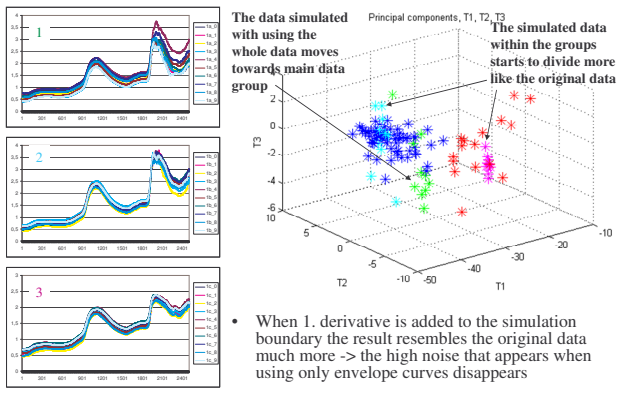
The main data is separated to its own group

We separate the outlayers to their own group for the simulation purposes

Simulated with Envelope Curve as Boundary

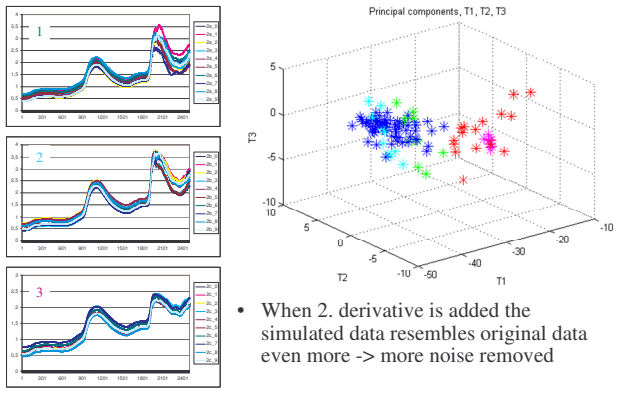


Simulated with EC & 1. Derivative as Boundaries



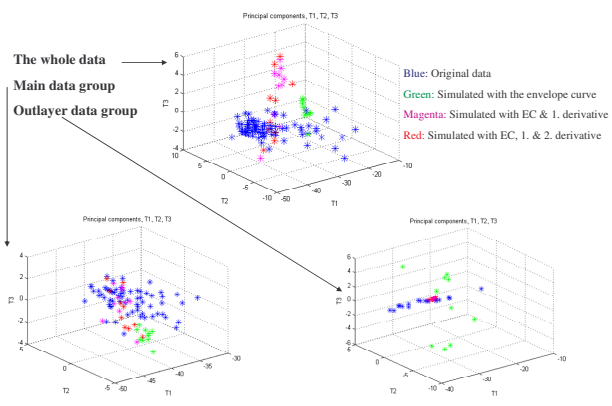
- When 1. derivative is added to the simulation boundary the result resembles the original data much more -> the high noise that appears when using only envelope curves disappears

Simulated with EC, 1. & 2. Derivative as Boundaries



- When 2. derivative is added the simulated data resembles original data even more -> more noise removed

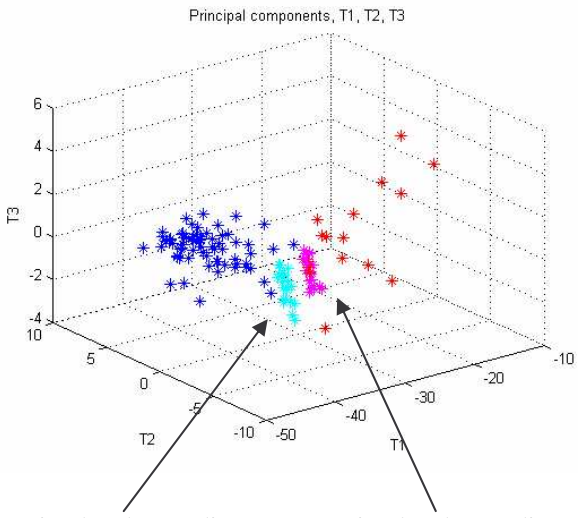
The Effect of Simulation Boundaries



Inside the groups, the different level of boundaries seem to have different effect on how tight group the simulated data locates

Searching the group boundaries

With genetic algorithms we can optimize the simulated data close to the group boundaries e.g. in order to find them



Data simulated according the main data group and optimized to be close to the outlayer data group

Data simulated according the outlayer data group and optimized to be close to the main data group

The Results and Conclusions

The results show that NIR spectra can be effectively simulated with genetic algorithms by using contour envelope curves and derivatives of the real data, with both melanoma and normal skin

When we are using more derivative boundaries, the simulated data becomes undetectable from the real data

Genetic algorithms can be used to optimize the simulated data e.g. for finding the envelope curves of those samples that will be classified as melanoma or as normal skin, i.e. finding the classification boundary

Acknowledgements

The EU Interregional project NIRCE is gratefully acknowledged for funding this research