Testing NIR based skin spectra analyzer system and software with the simulated data generated by genetic algorithm

T. Mantere,* P. Välisuo and J.T. Alander
*Department of Electrical Engineering and Automation University of Vaasa, P.O. Box 700, FIN-65101 Vaasa. *E-mail: timan@uwasa.fi*

Introduction

We have tested a NIR based melanoma (skin cancer), detection system with simulated NIR spectral data. The objective was to determine if the NIR spectral data of skin can be effectively simulated with genetic algorithms. 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.

Near-infrared (NIR) spectroscopy is based on absorption of electromagnetic radiation at wavelengths in the range 780 - 2500 nm. The concentrations of components, such as water, protein, fat, and carbohydrate can be detected by NIR spectroscopy. The major advantage of NIR spectroscopy is that it is non-invasive, and it is therefore suitable for performing skin examinations.

The motivation for generating simulated data for testing classification systems was that we have only a limited number of actual samples of skin spectra data, i.e. samples of skin cancer and healthy skin. However, wider testing of the performance and reliability of the system is important, so we decided to do it with simulated data. We also have earlier experience of testing software systems with GA (genetic algorithm)-generated test data.[1]

Genetic algorithms[2] are computer-based optimisation methods that use Darwinian evolution as a model and inspiration. The solution bases of a problem are encoded as chromosomes consisting of several genes. These virtual chromosomes are tested against a problem represented as a fitness function. The better the fitness value gained by a chromosome, the better is its chance to survive and to be selected to be a parent of new individuals.

Independent Component Analysis[3] (ICA) is a computational method for separating a multivariate signal into additive subcomponents, assuming the mutual statistical independence of the non-Gaussian source signals. When the independence assumption is correct, ICA is said to give very good results with mixed signals in blind separation. ICA finds the independent components, sources, by maximising the statistical independence of the estimated component.

The proposed method

We did not have enough skin samples to generate a reliable model, therefore we decided to generate new simulated samples with the help of real samples. The simulated samples were generated with GA optimisation by using constraints: the envelope curves (Figure 1) of original spectra and their 1st and 2nd derivatives.

This meant that the fitness function was such, that if the value of the individual was not inside the envelope curve, a penalty was added that corresponded to the amount of difference from the curve. The fitness value was minimised, and when an individual with the fitness value of zero was found, it meant that the spectra fulfilled all constraints.

In the previous paper[4] we used GA for wavelength selection and classified the original skin samples with PLS. This time we used ICA-based classification, and omitted the wavelength selection. Instead we used the whole spectra of 2500 wave-numbers (equivalent to 1000 - 2400 nm), selected from the area of interest (mainly water peaks). This large optimisation problem is extremely difficult for all optimisation methods, also for GAs.

We created a special GA version that was able to perform this task in which we used standard real-coded GA, so that the fitness function expanded every time that a proper solution for the current fitness function was found. In the beginning the fitness function is only one number long.

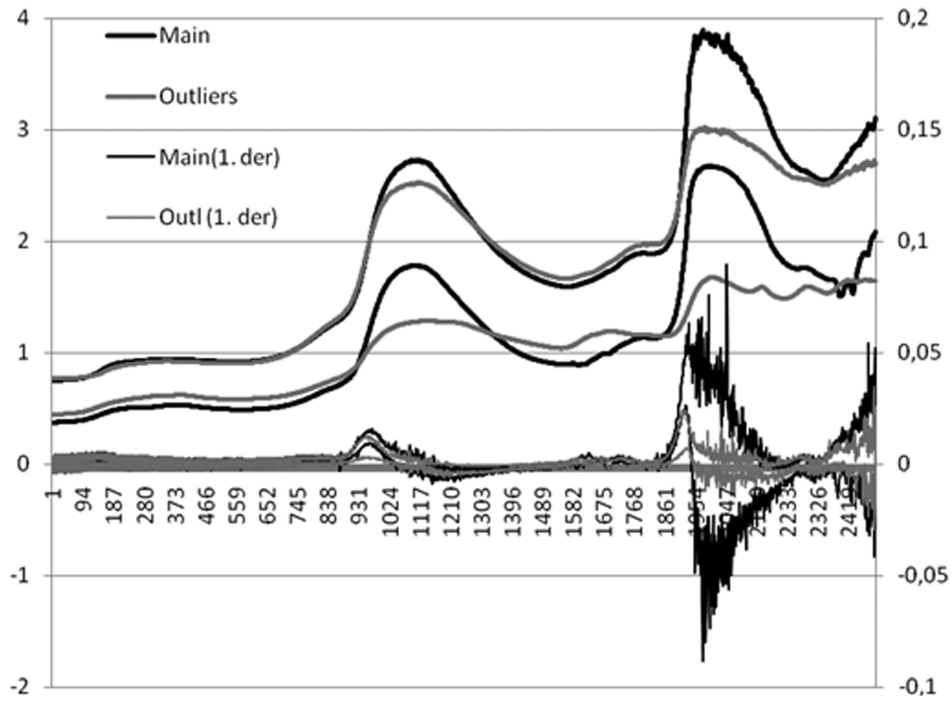

Figure 1. The envelope curves of original data and the envelope curve of 1st derivative of the original spectra, main data (black), the outlier groups (gray).

When we found an individual with a perfect fitness value, the fitness function expanded to two numbers and so on, until it reached 2500 numbers.

The individuals are always free to change after the new fitness function is introduced. However, thanks to elitism it is likely that the best individual for the previous fitness function survives, and also performs well in the new situation.

In addition, we programmed gene duplication crossovers and mutations:

(a) the gene is copied from the parent's neighboring gene locations
(b) the gene value is an arithmetic crossover between parent's neighboring genes
(c) the gene is copied from the neighboring gene of the same individual
(d) the gene value is the mean value of the same individual's neighboring genes

The reasoning behind these operations is that the neighboring values of NIR spectra are quite close to each other. Therefore copying neighboring values helps the GA to find and keep proper values that fulfill the constraints.

The GA parameters: population 30, elitism 16.7 %, uniform crossover rate 95 %, Gaussian mutation probability 5 %: stopping criteria: proper NIR spectra found, *i.e.* no boundary violations.

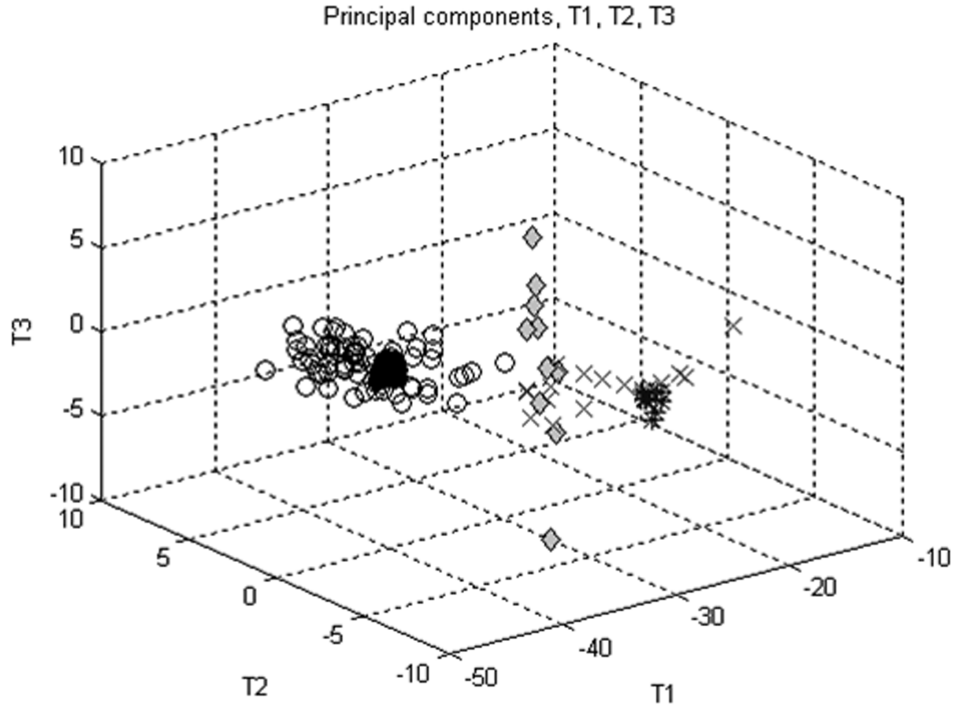

Figure 2. The original and NIR spectra simulated with envelope curve and 1st derivative, and analyzed by ICA. Original main spectra's (o) and outlier's (×). Spectra simulated by all original data's (◊), main group (•), and outlier group (*).

Results

First, we used only the envelope curves of the original data as boundaries. The results in Figure 2 shows how the NIR spectra simulated by using the whole original data were located in middle of the ICA classification, but those that are simulated by using the main or an outlier group were tightly concentrated in the middle of its reference group.

The reason is that spectra simulated using only the envelope curve as boundary have a lot of fluctuations (noise). If this noise is removed by *e.g.* low-pass filtering, the simulated spectra become practically identical.

We added the 1st derivative envelope to the boundaries. The NIR spectra simulated (Figure 3) with the main data group as reference seemed to have much more variation than in Figure 2.

NIR spectra simulated by the outlier group were still almost as tightly together. NIR spectra simulated with the whole original data were no longer in between the groups, but were instead located with the main data group. The explanation is that the envelope curve of the 1st derivative outlier group is much narrower than with main data group, so that the whole data envelope is more similar to the main data envelope.

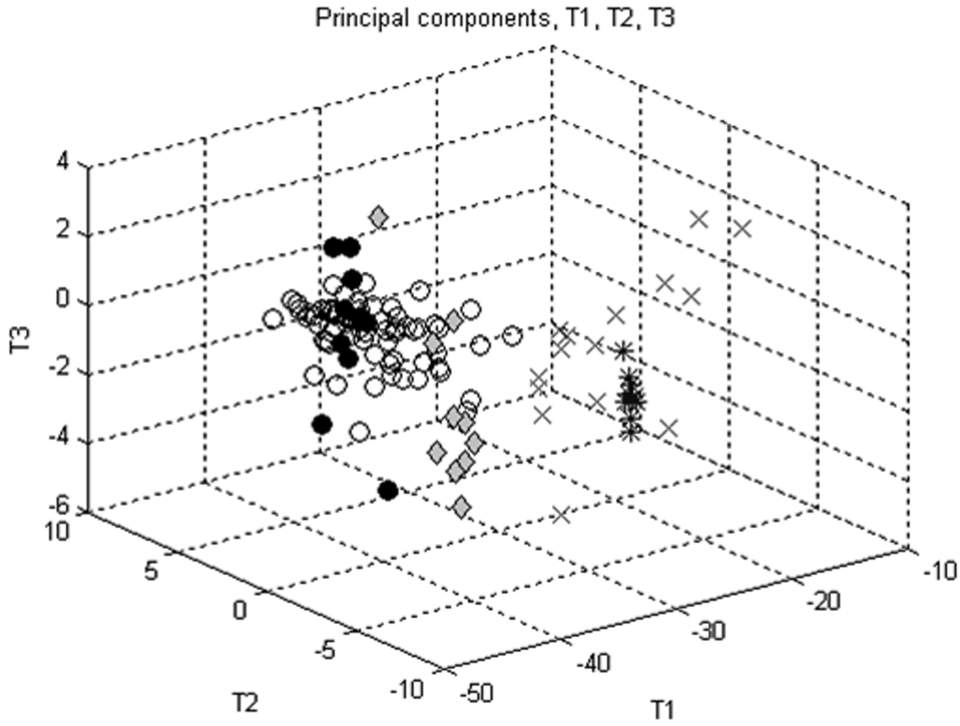

Figure 3. The original and NIR spectra simulated with envelope curve and 1st derivative and analyzed by ICA (symbols as in Figure 2.).

When the 2nd derivative was added to the boundaries (Figure 4.) the simulated NIR data get even more clearly classified into their reference groups.

The spectra simulated with all boundaries are seemingly similar with the original spectra. This was interpreted that GA is capable of generating simulated spectra that are practically inseparable from the real spectra.

Conclusions

The results showed that using GA, the NIR spectra can effectively be simulated by using contour envelope curves of real data and its derivatives. So GA was able to generate simulated NIR spectra that were difficult to separate from the actual spectra. When using more derivative boundaries, the simulated spectra become even more similar to the actual spectra.

Naturally, the quality of simulated spectra are dependent on the original samples, their number and quality. If the original samples adequately represent the whole reference group, then we can simulate almost perfect spectra. If the sample group is small, it makes simulation less accurate.

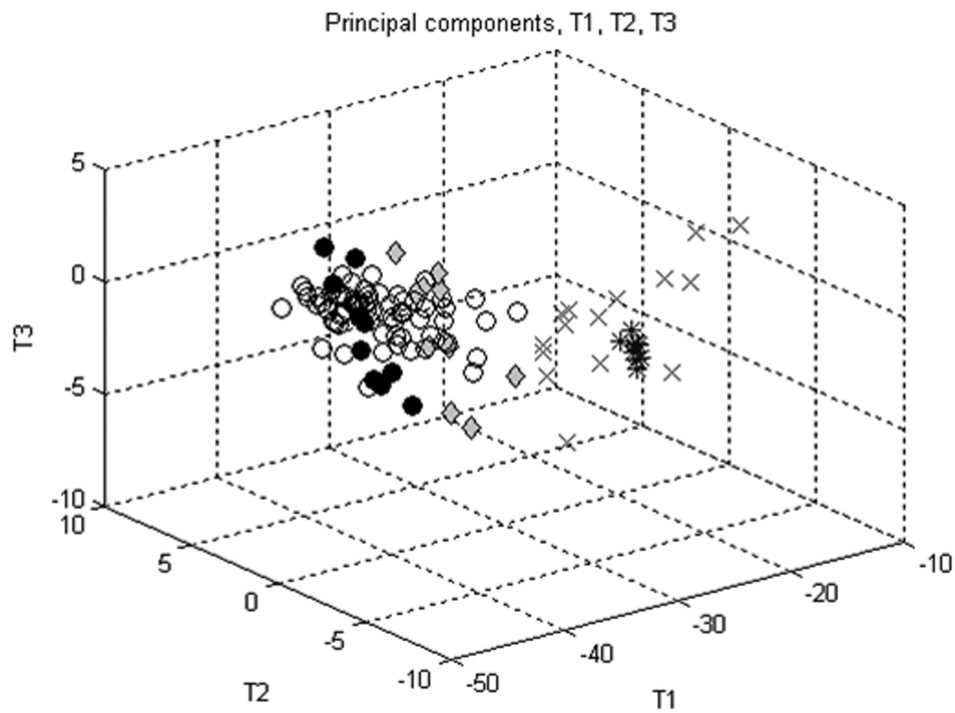

Figure 4. The original and NIR spectra simulated with envelope curve and 1st and 2nd derivative and analyzed by ICA (symbols as in Figure 2.).

Acknowledgements

The European Union Interreg Botnia-Atlantica project Field-NIRCe is acknowledged for funding.

References

1. T. Mantere, *Automatic Software Testing by Genetic Algorithms*, Acta Wasaensia (2003).
2. J. Holland, *Adaptation in Natural and Artificial Systems.* MIT Press, Cambridge, MA, USA (1992).
3. A. Hyvärinen, J. Karhunen, E. Oja, *Independent Component Analysis*. John Wiley & Sons Ltd, Chichester, UK (2001).
4. T.E.M. Nordling, J. Koljonen, J. Nyström, I. Boden, B. Lindholm-Sethson, P. Geladi, J.T. Alander, in *Proc. of the 3rd European Medical&Biological Engineering Conference* (2005).