Towards intelligent healthcare information processing systems

Thomas Mazzocco
tma@cs.stir.ac.uk

University of Stirling, Scotland (UK)

Signal Processing Workshop 2011
Brno University of Technology

Outline

Introduction

First case study

Second case study

Conclusions

Introduction

- Huge amount of information is nowadays collected in healthcare domain (patients' clinical history, diagnostic test results, etc.)
- A key challenge is how to mine the plethora of information and build intelligent hybrid and transparent models which can effectively help clinicians to make decisions
- Many pilot predictive models have been successfully developed over the last decades
- Data mining and (biologically inspired) intelligent techniques can be employed to develop and improve such models

First case study

The development of an intelligent logistic regression models to aid the diagnosis of early dementia

Mazzocco, Thomas; Hussain, Amir; “Novel logistic regression models to aid the diagnosis of dementia”, (Elsevier) *Expert Systems with Applications*, Available online 8 September 2011, ISSN 0957-4174, 10.1016/j.eswa.2011.09.023

Background

- Different sources of evidence are used in the dementia diagnosis process
- Timely diagnosis of dementia is a condition for improving dementia care
- General practitioners play a central role in the diagnosis process but 50-80% of cases are missed
- There is a limited range of readily available diagnostic instruments

Our aims

- To develop a new predictive model usable as a decision support system for dementia diagnosis
- To improve performance of available tools based on Bayesian Belief Networks
- To combine statistical approach with intelligent techniques
- To gain some insights about dementia diagnosis process

Benchmark vs. our model

- A dataset of 164 patients suspected of dementia is considered
- For each patient 14 variables about their clinical history are recorded along with physician's diagnosis

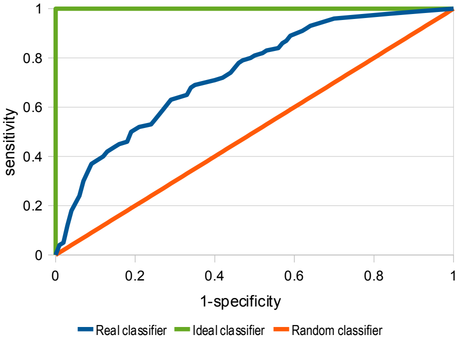
	benchmark model	our model
Technique	Bayesian belief network	logistic regression + AIS
Variables	expert driven	expert or data driven

ROC curve

- The outcome of the model is a probability between 0 and 1
- The probabilities given by the model need to be converted to a ‘yes’ or ‘no’ value, using a cut-off point
- The decision threshold used to separate positive and negative outcomes has a certain grade of arbitrariness depending on the desired false positive over false negative ratio
- The ROC curve is built representing for each possible value of the decision threshold, a pair of true-positive and false-positive performance rates, giving the predictive power of the model regardless the chosen threshold

ROC curve

Example of ROC curve, area under curve (AUC) is 0.5 for a random classifier and 1.0 for an ideal classifier



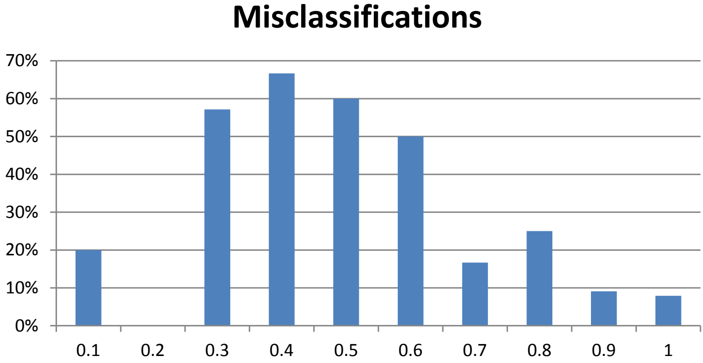
Model performance

The performance of our logistic regression model (both using expert driven and data driven variables selection) is reported in this table along with the benchmark (Bayesian belief network model)

	previous model (expert driven)	our model (expert driven)	our model (data driven)
AUC	0.764	0.783	0.879
R^2	n/a	0.371 - 0.602	0.365 - 0.601
Variables	10	16 + constant	8 + constant

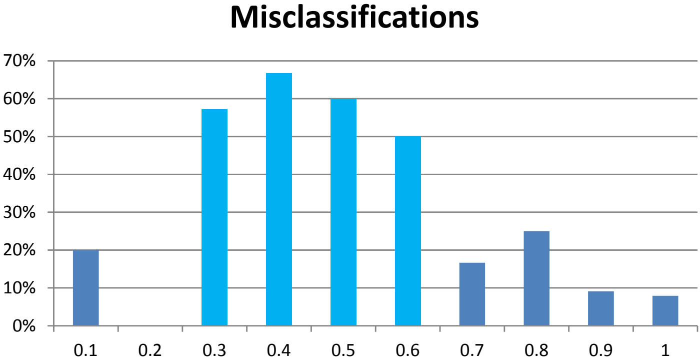
Model performance

In order to analyze the distribution of incorrect predictions, they have been grouped over 10 intervals of 0.1 each



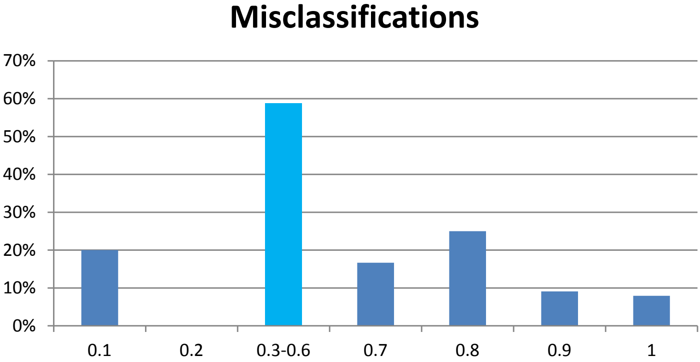
Model performance

As can be seen, misclassifications are concentrated in the range 0.2 to 0.6



Model performance

It is observed that almost 60% of predictions in the range between 0.2 and 0.6 are incorrect



Artificial immune system

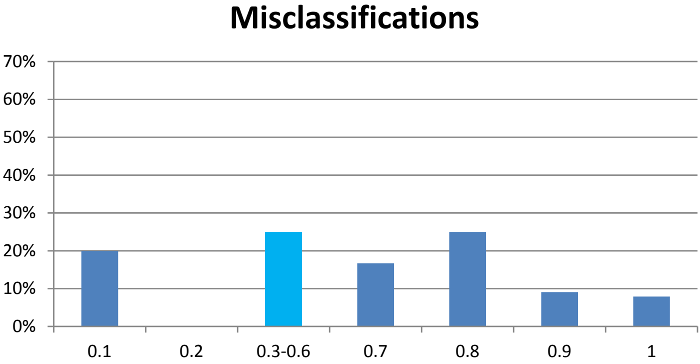
Artificial immune systems (AIS) are adaptive systems, inspired by theoretical immunology and observed immune functions, principles and models, which are applied to problem solving.
(De Castro-Timmis, 2002)

A new AIS algorithm has been developed:

- based on negative selection algorithm
- with a validation mechanism (backward elimination approach)
- making an optimal use of dataset during training
- to be used for datapoints where misclassification is more likely

Model performance

After processing the data of the patients having a prediction in the range 0.2 - 0.6 with our AIS model, misclassifications ratio decreased to 25%, in line with other intervals



Discussion

- 3 variables included in the previous model do not improve the logistic regression model, 1 variable not included in the previous model significantly improved the model performance
- A complete data-driven approach (both in variables selection and in model building) resulted in a more accurate predictive model when compared with the hand-cradfted one
- While logistic regression model can provide with a good predictive power, intelligent techniques (i.e. AIS) can help further reducing misclassifications

Second case study

The development of a side-effects mapping model in patients with lung, colorectal and breast cancer receiving chemotherapy

Mazzocco, Thomas; Hussain, Amir; “A side-effects mapping model in patients with lung, colorectal and breast cancer receiving chemotherapy”, *13th IEEE International Conference on e-Health Networking Applications and Services (Healthcom), 2011*, pp.34-39, 13-15 June 2011

Background

- About 300,000 individuals were diagnosed with cancer in the United Kingdom (5% of the population)
- Different treatments are available and the survival rates have been improving over the last 30 years
- Chemotherapy is often administered, but exposes patients to risk of significant side-effects that could have a negative impact on patients quality of life and daily living
- Poorly informed patients are less likely to comply with treatment and are more likely to experience anxiety

Related work

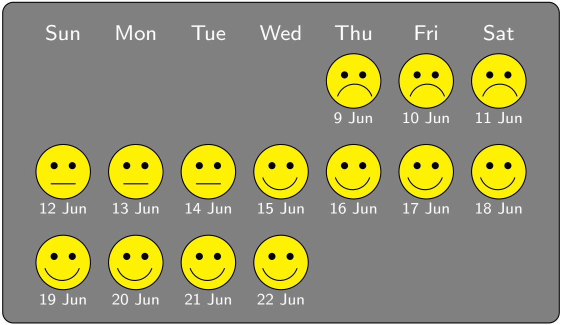
- Technology may enhance communication between healthcare professionals and patients:
  - improvements in quality of life
  - symptom control
  - reductions in the rate of hospitalizations
  - emergency department visits
  - cost savings
- Patients have positive views of using technology, reporting improvements in communication with healthcare providers
- We build upon the **ASyMS©** (Advanced Symptom Management System) which has been developed as an example of the use of technology in cancer care

ASyMS©: register, monitor, predict symptoms

- Mobile telephone-based remote symptom monitoring system
- Patients complete a symptom questionnaire on a mobile phone and this information is sent directly to their hospital-based healthcare professional
- Self-care advice is then given and the healthcare professional may contact the patient if necessary
- The patients symptom history is used to predict the likely side effects a patient could expect over the course of treatment

ASyMS©: register, monitor, predict symptoms

A diary is presented on patients mobile phones where, for each day, a smiley, sad or neutral face is used to depict the overall side-effects situation predicted for that particular day



Sun	Mon	Tue	Wed	Thu	Fri	Sat
				☹ 9 Jun	☹ 10 Jun	☹ 11 Jun
😐 12 Jun	😐 13 Jun	😐 14 Jun	☺ 15 Jun	☺ 16 Jun	☺ 17 Jun	☺ 18 Jun
☺ 19 Jun	☺ 20 Jun	☺ 21 Jun	☺ 22 Jun			

Our aim

To improve the predictive power of ASyMS©:

- generalizing the previous model designing a more powerful and comprehensive side-effect risk model

  a single mathematical model, which predicts on a day-by-day basis the symptoms that patients with cancer receiving chemotherapy are going to experience

- extending the breast cancer model to cover colorectal and lung cancer conditions

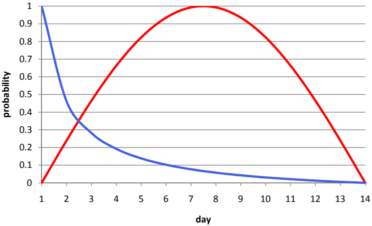
ASyMS© vs. our model

	ASyMS©	our model
Dataset	24 patients	56 patients
Conditions	breast cancer	breast, colorectal, lung cancer
Model	different for each symptom	same for all symptoms

- In both models, data about symptoms were collected over 4 cycles of chemotherapy, each lasting 14 days
- Selected symptoms: diarrhoea, fatigue, hand and foot sore, mucositis, nausea

Equation model

- Two main time-dependant tendencies over time were outlined: the ‘peak effect’ and the ‘inverted U-shape effect’



$$S(d) = sin\left(\frac{d-1}{d_{max}-1}\pi\right) = sin\left(\frac{d-1}{13}\pi\right)$$

$$H(d) = \frac{d_{max}}{d_{max}-1}\left(\frac{1}{d} - \frac{1}{d_{max}}\right) = \frac{14}{13}\left(\frac{1}{d} - \frac{1}{14}\right)$$

Equation model

- The same formal model for all symptoms combining two effects (inverted U-shape and peak) and possible differences between cycles
- Three groups: lung, colorectal and breast cancer
- Coefficients determined using regression

$$P(d) = a \cdot S(d) + b \cdot H(d) + \sum_{n=1}^{4} c_n \cdot D_n$$

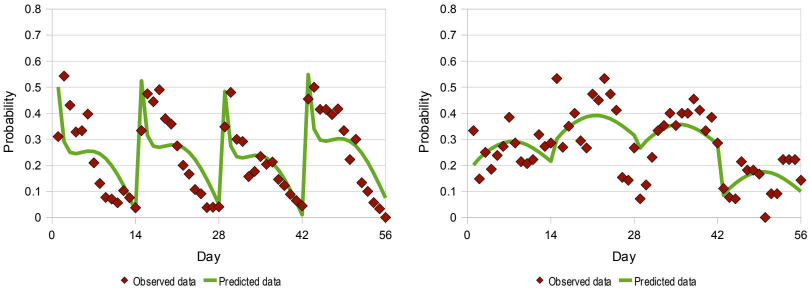
$P(d)$ probability of experiencing the symptom on day d

a, b, c_n coefficients determined using regression

D_n dummy variable to identify the n-th cycle

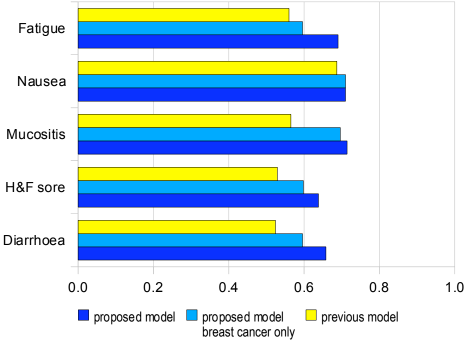
Examples of mapping

Observed versus predicted data for nausea in breast cancer (left) and for mucositis in lung cancer (right)



Model performance

Receiver-operating characteristic (ROC) curve has been used to evaluate the model's performance: areas under curve (AUC) are here compared



Discussion

- The combination of the empirically observed effects produced an average increase of 19% on model performance
- The proposed model has been both generally improved and may be reusable in different contexts
- With a more accurate model
  - patients know which symptom they should expect and when
  - health professionals take appropriate action to avoid or minimize expected discomforts

Conclusions

- Two models have been successfully developed building on existing tools combining statistical approaches with intelligent techniques
- An intelligent hybrid data-driven approach can outperform traditional (hand-crafted) benchmark predictive models
- The developed models are flexible and transparent, making clear the reason for a certain prediction
- The outlined frameworks seem to be general enough to be re-used for different applications