Genetic programming as a method to develop powerful predictive models for clinical diagnosis

Ivar Siccama
Chordiant Software Inc.
The Netherlands
ivar.siccama@chordiant.com

Maarten Keijzer
Chordiant Software Inc
The Netherlands
maarten.keijzer@chordiant.com

ABSTRACT

In the field of medicine it is of vital importance to accurately predict the presence of a disease (diagnostic prediction) or the future occurrence of a certain event (prognostic prediction). Genetic programming provides a method to develop such prediction models in an optimal way. In this paper we discuss as an example the diagnostic prediction of pulmonary embolism (PE), and compare the method of genetic programming with the logistic regression technique, which is well-known in the medical field.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Medical Information Systems

General Terms

Algorithms, Performance

Keywords

Genetic Programming, Diagnosis, Prognosis, Logistic Regression, Pulmonary Embolism

1. INTRODUCTION

The use of predictive models in clinical diagnosis has so far been restricted to paper score-cards, often developed decades ago. The advantage of these scorecards is that they can be calculated by hand; a corresponding disadvantage is that they *need* to be calculated by hand and are thus severely limited in their complexity and modelling capabilities. The advances made in non-linear statistics, knowledge discovery and data mining have largely been ignored for clinical use.

With the advent of electronic patient records, desktop computers PDAs, and corresponding networking and database infrastructure, the time has now come to exploit the use of predictive models in a clinical setting. A medical practitioner, be it physician or nurse, could benefit from these models. Predictive models may not only be used for diagnosis where, due to a conflict with the core competence of a physician, acceptance may be slow, but can also play a crucial role in the suggestion of treatments, prevention of errors and the advice of clinical tests. Predicting duration or the possibility of complications can play a crucial role in the planning of operation rooms, hospital beds and intensive care departments. As this latter use of predictive models leads to immediate cost-reduction, this use of predictive models is readily accepted.

Predictive models come in various flavours. The time-honoured workhorse of predictive modelling in the medical field, in particular in epidemiology, is logistic regression. This technique works under the assumption of linear addition of evidence, and comes with a set of analyses in the form of odd-ratios, relating the unit change in the inputs to the probability of a particular event occurring. Due to the linearity assumption logistic regression models can exhibit sub-optimal predictive performance. This can be countered partially by performing a cubic spline interpolation and introduction of collinearity factors. These improvements must however be found manually. In contrast with linear models, non-linear models do not have a-priori bounds on predictive performance. As in the field of medicine any increase in predictive performance is literally of vital importance, non-linear models should be considered. In our case we study a variant of genetic programming that is designed to create robust and highly predictive models.

Below a case-study is presented where a non-linear model is built using genetic programming and compared with a standard logistic regression model created by epidemiologists. Not only is the non-linear model significantly more accurate than the logistic regression model, it also uses one predictor less in predicting pulmonary embolism. This gives a clear example where the use of more powerful predictive modelling strategies can deliver better performance. When the ultimate aim is to put predictive models on the desk of the medical practitioner, such reliable, high performance is of utmost importance.

This paper will mainly focus on this particular study, but will also give a quick overview of two other projects that have been tackled with the same motivation: bringing highly predictive models into use in the medical field.

GECCO'05, June 25–29, 2005, Washington, DC, USA.

2. PULMONARY EMBOLISM

2.1 Method

Data were used from a prospective diagnostic study among 398 patients in secondary care of 18 years or older who were suspected of PE. Variables known were *patient history*, *physical examination*, *chest radiography* and *leg ultrasound*. Of all patients, 170 were diagnosed as having PE (prevalence = 43 %), which was determined using a VQ scan and pulmonary angiography. The data set was split, randomly, in two parts: a derivation set of approximately 67 % (165 patients) and a validation set of approximately 33 % (133 patients). The derivation set was used for model development and the validation set to test the validity of the two models. In addition bootstrapping was used to crosscheck the validity of the models.

The aim of both models was to develop a prediction model to estimate the presence or absence of PE as good as possible with a minimum of diagnostic tests (predictors). The final logistic regression model was obtained by selecting predictors with P-values < 0.10 using the Likelihood Ratio test. In the genetic programming technique models were represented by binary trees, composed from a set of binary operators. A limit on the maximum depth of each tree allowed up to 8 predictors to be used. Crossover and mutation operated on the branches and nodes of the trees. The used pool size was 40 and the search was terminated when no significant progress was observed (after 2000 generations).

2.2 Results

The final logistic regression model used 8 predictors (see Figure 1, where the model is presented as a nomogram) and the final genetic programming model used 7 predictors (see figure 2). The performance was tested using the external validation set and using the internal validation technique of bootstrapping. Applying the logistic regression model to the validation set yielded an ROC area of 0.68 (95 % CI: 0.59-0.77) and application of the final genetic programming model on the validation set resulted in an ROC area of 0.73 (95 % CI: 0.64-0.82). The performance of the genetic programming model is therefore significantly higher than that of the logistic regression (with one diagnostic test less). Bootstrapping tests confirmed this result.

Usage in a clinical setting can be through the use of risk categories, where the non-linear scores of the genetic programming model are distributed into score intervals with an equal number of patients. An example is shown in Table 1. For the highest risk category (containing 20 % of all patients) the probability of PE is 79 %.

3. DISCUSSION

It is to be expected that in the near future data collection in hospitals will become more and more automized with the introduction of Electronic Patient Records. This will allow larger data sets to be collected to develop better models. Also, it will become possible to apply on the same computer infrastructure the results of prediction models, the results of which can be used by medical practitioners in their diagnoses and decisions. Since the computer will do the calculation, there will be no need for simple models that can be calculated manually, but the models may be of a more complex nature and therefore potentially more powerful, which can be of vital importance.

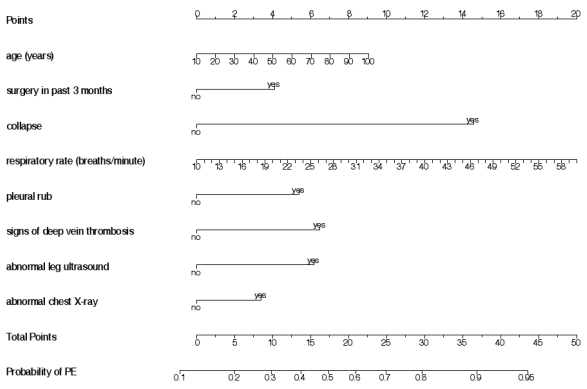
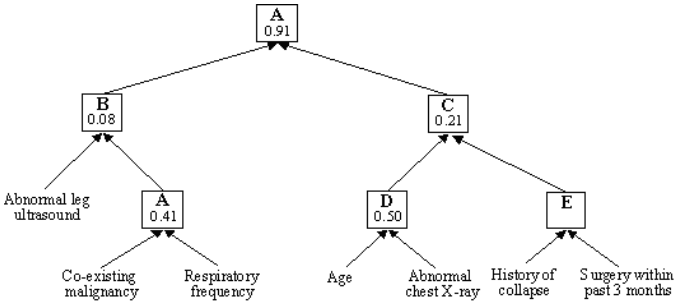


Figure 1: Nomogram presentation of the logistic regression prediction model



$$
\begin{aligned}
A &= 1 - p\sqrt{1-x} - (1-p)\sqrt{1-y} \\
B &= pf(x) + (1-p)f(y) \text{where} f(x) = 2x - 0.593k(2x-1)^3 \\
C &= px^2 + (1-p)y^2 \\
D &= px^2 + (1-p)y^2 \\
E &= \frac{1}{2} + \frac{1}{2}\sin(x^2 + \frac{1}{2}\pi y^2 - 1)
\end{aligned}
$$

Figure 2: The final model created by genetic programming, represented as a binary tree. The nodes A-E represent binary operators, where x and y are the inputs from the left and right of the boxes, the values used for the parameter p are shown in the boxes

Score interval	Probability of PE (%)
0.00-5.79	15.1
5.80-6.92	21.1
6.93-6.95	43.4
6.96-8.08	59.3
8.09-8.82	79.2

Table 1: Score intervals, each containing 20 % of all patients, and the corresponding probability of PE for each interval. Given the category the patient falls into, the medical practitioner can make a decision on the treatment.

This example shows that a powerful prediction model can be obtained using Genetic Programming. An extensive analysis on this particular problem can be found in [1]. The use of standard computer technology will make it possible to replace the paper nomogram in Figure 1 with a simple computer application that, given the patient characteristics from the EPR infrastructure, will highlight the predicted score-interval from Table 1. Such a simple application provides valuable information to use in the treatment plan.

The search for such a complex, non-linear model, is ideally suited for a genetic algorithm (whereas, for the more structured formula of a logistic regression, the method of maximum likelihood optimization is sufficient). It should be noted that studies on medical data are often done on small samples, and a thorough method of validation is essential. Both bootstrapping and an external validation set were used here. All methods of developing prediction models carefully have to estimate the amount of overoptimism, especially since the proposed relations are more difficult to interpret.

Another advantage of genetic programming lies in the automatic variable selection. In another project, the authors have developed a prediction model using mass spectra data from cerebral spinal fluid, where the number of variables (peptide peaks) is very large (these results are still to be published, but an abstract containing the preliminary setup can be found in [2]). Used in combination with methods of variable reduction the results are very promising, and could lead to a diagnostic application similar to the one described above.

An example where a prognostic application is already in use in a hospital is described in [3] and [4]. Here predictive models are used to assess the preoperative risks using the results of an anamnesis and physical examination. After creating predictive models on the hospital's own records, the models have been integrated in the existing hospital infrastructure. These applications are now used by the anesthesiologists for planning the operation, assessing the risk before an operation, and even monitoring the risk during the operation in the operation room. This particular application is a complete realization of the goal of putting predictive models there where they can be of most use: in the hands of the medical practitioner.

4. ACKNOWLEDGMENTS

The authors would like to acknowledge the efforts of the Julius Centrum, University Medical Center, Utrecht on the pulmonary embolism project; Diakonessenhuis Hospital Utrecht for the work done in preoperative risk screening; and the Erasmus Medical Center, Rotterdam for the proteomics project. This work was funded in part by Senter in the context of the Medicast project.

5. REFERENCES

[1] C. Biesheuvel, I. Siccama, D. Grobbee, and K. Moons. Genetic programming outperformed multivariable logistic regression in diagnosing pulmonary embolism. *J. Clin. Epidemiol.*, 57(6):551–560, jun 2004.

[2] L. Dekker, I. Siccama, G. Jenster, P. S. Smitt, and T. Luider. A new method to analyze matrix-assisted laster desorption/ionization time-of-flight peptide profiling mass spectra. *Rapid Commun Mass Spectrom.*, 19(7):865–870, 2005.

[3] P. Houweling, H. Korsten, and I. Siccama. Het voorspellen van postoperatieve misselijkheid en braken. *Ned. Tijdschrift voor Anesthesiologie*, 16(7), 2003.

[4] R. T. Meesters, I. Siccama, and P. L. Houweling. Decision rules and prediction models in preoperative risk management. In E. S. of Anaesthesiologists, editor, *Euroaneasthesia 2004*, 2004.