Enhancing Generalised Linear Models with Data Mining

Dr. Inna Kolyshkina, Sylvia Wong, and Steven Lim

Enhancing Generalised Linear Models with Data Mining

Inna Kolyshkina, PricewaterhouseCoopers(Actuarial)
Sylvia Wong, PricewaterhouseCoopers(Actuarial)
Steven Lim, PricewaterhouseCoopers(Actuarial)

1. Introduction

Generalised linear models (GLM) appear to be a tool that has become very popular and have shown to be effective in the actuarial work over the past decade, see for example Haberman & Renshaw (1998). A detailed description of the GLM methodology is outside of the scope of this paper, and can be found in the other sources such as McCullach and Nelder(1989).

Data mining methodologies are more recent and their popularity in the actuarial community is increasing. They have been used in insurance for risk prediction/assessment, premium setting, fraud detection, health costs prediction, treatment management optimization, investments management optimization, customer retention research and acquisition strategies. Recently a number of publications have examined the use of data mining methods in an insurance and actuarial environment (eg, Francis (2001), Francis (2003). The main reasons for the increasing attractiveness of the data mining approach is that it is very fast computationally and also overcomes some well-known shortcomings of traditional methods.

However the advance of the new methodologies does not mean that the proven, effective techniques such as GLM should be wholly replaced by them. This paper discusses how the advantages and strengths of GLM can be effectively combined with the computational power of data mining methods presenting an example of the combining multivariate adaptive regression splines (MARS®) and GLM approaches by running MARS® model and then building a GLM with MARS® output functions used as predictors. The results of this combined model are compared to the results achievable by hand-fitted GLM. Comparisons are made in terms of time taken, predictive power, selection predictors and their interactions, interpretability of the model, precision and model fit.

2. Enhancing the Linear Modelling Approach by Combining it with Data Mining

GLM being a linear technique shares the usual shortcomings of the linear modelling approach.

Linear models

- operate under the assumption that data is distributed according to a distribution in the exponential family

- are affected by multicollinearity, outliers and missing values in the data

- are often troublesome to use for selecting important predictors and their interactions

- are troublesome to use with categorical predictors that have large numbers of categories (for example, postcode, occupation code etc) as this can lead to unreliable results due to sparsity-related issues

- take longer to build because of the need to address the issues above by transforming both numeric and categorical predictors and choosing predictors and their interactions by hand which can prove to be a lengthy task.

Data mining techniques in contrast

- are typically fast,
- easily select predictors and their interactions,
- are minimally affected with missing values, outliers or collineanty and
- effectively process high-level categorical predictors.

This suggests that combining a linear approach with data mining tools can expedite the modelling process, allowing the modeller to attain equal or better model accuracy in less time with the same level of interpretability. Such models, usually combining decision trees, multivariate adaptive regression splines and GLM have been used by our team in a number of projects (see Kolyshkina and Brookes, 2002).

3. Multivariate Adaptive Regression Splines (MARS®)

Multivariate adaptive regression splines (MARS®) is becoming increasingly popular in the actuarial community; for example, Francis (2003) describes application of MARS® to insurance fraud analysis.

We provide below a brief introduction to the MARS® methodology, a more detailed description can be found in other sources (see for example, Friedman, 1991, Hastie et al. (2001)).

MARS® is an adaptive procedure for regression, and can be viewed as a generalisation of stepwise linear regression or a generalization of the recursive partitioning method to improve the latter's performance in the regression setting (Friedman, 1991; Hastie et al, 2001).

The MARS® procedure builds flexible regression models by fitting separate splines (or basis functions) to distinct intervals of the predictor variables. Both the variables to use and the end points of the intervals for each variable-referred to as "knots" -are found via an exhaustive search procedure, using very fast update algorithms and efficient program coding. Variables, knots and interactions are optimized simultaneously by evaluating a "loss of fit" (LOF) criterion. MARS® chooses the LOF that most improves the model at each step. In addition to searching variables one by one, MARS® also searches for interactions between variables, allowing any degree of interaction to be considered.

The "optimal" MARS® model is selected in a two-phase process. In the first phase, a model is grown by adding basis functions (new main effects, knots, or interactions) until an overly large model is found. In the second phase, basis functions are deleted in order of least contribution to the model until an optimal balance of bias and variance is found. By allowing for any arbitrary shape for the response function as well as for interactions, and by using the two-phase model selection method, MARS® is capable of reliably tracking very complex data structures that often hide in high-dimensional data (Salford Systems, 2002). MARS® is fast, requires less data preparation than some other techniques, can easily handle missing values or noisy data, and the output, for both the model and the basis functions, is easy to interpret. MARS® is implemented in a software package produced by Salford systems. The package is easily available, inexpensive and can work with data in most formats (SAS, SPSS, dbf etc). The output MARS® produces can be combined with any GLM software with minimal effort as it is easy to code in any program language such as SAS which is the main data analysis software package used by many actuaries.

4. How the Use of MARS® Can Expedite GLM Building

Most of the shortcomings of linear models outlined above can be overcome by using MARS® as a way of pre-processing predictors before putting them in a GLM. This will also significantly reduce the time needed for model building. This can be done by feeding MARS® output (in the form of basis functions created by MARS®) as inputs into a GLM.

MARS® is minimally affected by multicollinearity, outliers and missing values in the data, easily handles categorical predictors with large numbers of categories and requires less data preparation than linear methods, it quickly selects important predictors and their interactions and transforms numeric and categorical predictors in such a way that the resulting variables are easy to interpret. The modeller though needs to make sure that the transformed predictors make business sense, and that the MARS® model is stable.

We have seen that although MARS® output functions are not created specifically to be used as the input for a linear model, in practice about 90% of them turn out to be significant predictors in a GLM. Another feature of the MARS® output functions that makes them useful is that they are linearly independent as stated in Friedman (1991) which means that the multicollinearity issues do not arise in the GLM that uses them as explanatory variables.

In the case study below this technique was applied to summarised data, but it would be even more efficient on the individual level data with many predictors, both numeric and categorical.

5. Case Study. Queensland Industry CTP data provided by Motor Accident Insurance Commission (MAIC)

5.1 Background

The methodology described above was applied in order to model the ultimate incurred number of claims based on reported claim data. The data used was industry-wide auto liability data from Queensland (commonly called Compulsory Third Party or "CTP" in Australia).

5.2 Data Description

Individual claim data was aggregated into the number of claims reported for each accident month and development month for input to the GLM. The variables used for the analysis were accident month, accident quarter, number of casualties, development month, development quarter, number of vehicles in the calendar year, and number of vehicles exposed in the month.

5.3 Modelling Methodologies Description

5.3.1 Hand-fitted GLM

An initial GLM was created without using MARS®. This was a Poisson model with the log link, using the number of vehicles exposed in the month as the offset. The transformations and interactions of the input variables were created manually for the purposes of both best model fit and interpretability. The model fit was assessed by usual methods such as ratio of deviance to the degrees of freedom, predictor significance, link test and residual analysis. All the assessments showed adequacy of the model fit.

5.3.2 MARS®- enhanced GLM

A second model was created by preprocessing the variables in MARS® as described above and then including them in a GLM as the inputs. First we built a MARS® model with the ratio of incurred number of claims to number of vehicles exposed in the month as the dependent variable. The model output included explanatory variables pre-processed by MARS® : We then used these variables as the inputs in the Poisson model with log link and the number of vehicles exposed in the month as the offset. The GLM output showed that most of these variables were significant. We then used backward elimination to refine the model by excluding the inputs that were not significant. The resulting model fit was assessed in the same way as for the hand-fitted model above.

5.4 Comparison of Models

5.4.1 Timing

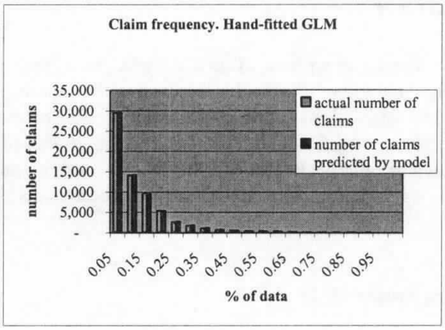
The hand-fitted model of this type would usually take about 5-7 days to build and refine.

The MARS® - enhanced model involved running the model in MARS® with different settings such as finding the optimal level of predictor interaction, then copying and pasting MARS® output into SAS and running and refining the GLM. This took about half a day. The MARS® analysis took less than an hour.

5.4.2 Goodness of fit. Bar charts. Gains chart

The fit of both GLMs was assessed by usual methods such as ratio of deviance to the degrees of freedom, predictor significance, link test and residual analysis. The MARS®- enhanced GLM has shown a similar if not slightly better fit to the hand-fitted GLM.

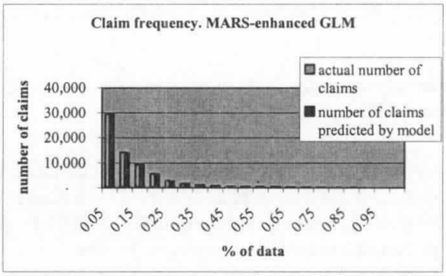
Figure 1. Average actual and predicted values for overall number of claims, hand-fitted GLM



A further diagnostic of model performance is analysis of actual versus expected values of the number of claims.

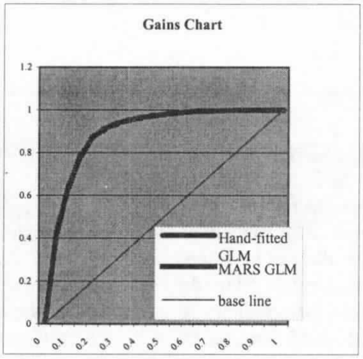
Such analysis can be pictorially represented by a bar chart of averaged actual and predicted values for the number of claims. To create such a chart, the records were ranked from highest to lowest in terms of predicted number of claims for each model, and then segmented into 20 equally sized groups. The average predicted and actual values of the number of claims for each group were then calculated and graphed. The chart for the hand-fitted GLM is shown in Figure 1 and the chart for MARS®-enhanced GLM is shown in Figure 2.

Figure 2 Average actual and predicted values for overall number of claims, MARS-enhanced GLM



Comparison of the charts suggests that the models fit equally well, with the MARS®-enhanced GLM having a marginally better fit. The hand-fitted GLM slightly over-predicts for the fifth group and under-predicts for the third group while the MARS® - enhanced GLM predicts well for the higher expected numbers of claims but slightly overpredicts for the groups with lower numbers of claims. However, the scale of these errors is of little business importance.

Figure 3 Gains chart for number of incurred claims for both models.



Another graphical method used for the comparison of the models was gains charts. Gains charts are described in detail in literature, see for example Berry & Linoff (2000).The gains chart presented in Figure 3 shows that both models are able to predict the segments with high number of claims with a good degree of accuracy. As a rough guide, taking the 15% of records predicted as having the highest number of incurred claims by the model, we end up with 80% of the total number of incurred claims. Taking the 30% of records predicted as having the highest number of incurred claims by the model, we end up with 93% of the total number of incurred claims. The graph above shows that the models perform equally well. Detailed analysis of actual statistical results suggests that the MARS® - enhanced GLM performs marginally better than the hand-fitted model.

It can be seen from the charts above that the MARS® - enhanced GLM fits only slightly better than the hand-fitted model. This effect would be more apparent in raw data modelling than in modelling summarised data as the trends observed are likely to be smoother and easier to identify as random variation cancels out.

5.4.3 MARS® -created vs hand-transformed variables and predictor interactions: similarities and differences

Comparison of the MARS®-created predictor variables with those which were manually created showed a great degree of similarity. For example, if we compare the predictors based on the variable "development month", MARS® placed knots mostly at the same points that were found important by the hand-fitted model. The differences included the fact that the hand-fitted model included the variable "minimum (development month, 10)" which is equal to 10 if development month is greater than 10 and is equal to development month otherwise, while MARS® selected 9 rather than 10 as the "knot point". MARS® also selected interactions of predictors that were not picked up by the hand-fitted model such as the interaction of development month and experience month.

5.4.4 Interpretability of the models

Interpretability of the models was similar. The hand-fitted model was easier to interpret because it included less predictors and less predictor interactions than the MARS®- enhanced model.

5.5 Findings and results

The fit and precision of the models was similar with the MARS®-enhanced GLM showing a slightly better fit than the hand-fitted GLM. The MARS®-enhanced GLM included predictor interactions not picked up by the hand-fitted model. This effect would be more pronounced in raw data modelling than in modelling summarised data, as the trends observed are likely to be smoother and easier to identify as random variation cancels out. The hand-fitted model was easier to interpret because it included less predictors and less predictor interactions than the other model. Building of the MARS®-enhanced GLM was considerably faster and more efficient.
These findings suggest that MARS® is a useful tool to enhance and expedite GLM modelling.

5.6 Future directions

For large data sets, our team has found that combining decision trees (CART®) with MARS® and GLM proves quite effective as described in Kolyshkina & Brookes, 2002.

Also as an additional check of fit of a GLM model, a parallel model built in MARS® can be used. The model will be built as a part of the stage described previously and will not require additional time. The model equation can be copied and pasted from MARS® to SAS or another package directly. Comparison of the predicted values of this model to the GLM model graphically and numerically can suggest some insights and inform the choice of the best of the models.

5.7 Conclusion

The results described above demonstrate that the use of a data mining technique, MARS, to enhance GLM building makes the model-building process considerably faster and more efficient. This approach allows to achieve higher computational speed by expediting the process of the selection of predictors and their interactions and variable transformation. The precision of this model is higher than for the hand-fitted model as shown by traditional GLM assessment methods as well as by using additional goodness-of-fit analyses such as gains chart. The effects described would be even more pronounced in raw data modelling than in for modelling summarised data as described in the case study, especially for large data sets with many potential predictors. The interpretability of the MARS®-enhanced GLM is similar level to that of the hand-fitted model.

ACKNOWLEDGEMENTS

We would like to thank Mr Daniel Tess and Ms Raewin Davies (PricewaterhouseCoopers Actuarial, Sydney) for support, advice and thoughtful comments on the manuscript. We would also like to thank Motor Accident Insurance Commission for providing the data for the analysis.

REFERENCES

[1] Berry, M.J.A. and Linoff, G. (2000). Mastering Data Mining. The Art and Science of Customer Relationship Management. John Wiley &Sons, Inc.

[2] Francis, L. (2001). Neural networks demystified. Casualty Actuarial Society Forum, Winter 2001, 252–319.

[3] Francis, L. (2003). Martial chronicles. Is MARS® better than neural networks?. Casualty Actuarial Society Forum, Winter 2003, 27-54.

[4] Haberman, S. and Renshaw, A. E. (1998). Actuarial applications of generalized linear models. In Hand, D. J. and Jacka, S. D. (eds). Statistics in Finance. Arnold, London.

[5]Han , J., and Camber M. (2001) Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.

[6] Hastie, T., Tibshirani R. and Friedman, J. (2001). The elements of statistical learning: Data Mining, Inference and prediction. Springer-Verlag, New York.
[7]Kolyshkina, I. and Brookes, R. Case study.Modelling Risk in Health Insurance: A Data Mining Approach. In Simoff, S.J., Williams G.J. and Hegland, M (eds) AI 2002 Workshop Proceedings Data Mining. Accepted for publication.

[8]Kolyshkina, I, Petocz, P. and Rylander,I. Modelling Insurance Risk: A Comparison of Data.
Mining and Logistic Regression Approaches. In Simoff, S.J., Williams G.J. and Hegland, M (eds) AI 2003 Workshop Proceedings Data Mining. Accepted for publication.

[9]Lewis, P.A.W. and Stevens, J.G., "Nonlinear Modeling of Time Series using Multivariate Adaptive Regression Splines," Journal of the American Statistical Association, 86, No. 416, 1991, pp. 864-867.

[10]Lewis, P.A.W., Stevens, J., and Ray, B.K., "Modelling Time Series using Multivariate Adaptive Regression Splines (MARS®)," in Time Series Prediction: Forecasting the Future and Understanding the Past, eds. Weigend, A. and Gershenfeld, N., Santa Fe Institute: Addison-Wesley, 1993, pp. 297-318.

[11]McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models (2nd edition). Chapman and Hall, London.

[12]Salford Systems (2002). MARS® ® (Multivariate Adaptive Regression Splines) [On-line] http://www.salford-systems.com, (accessed 08/10/2002).

[13]Smyth, G. (2002). Generalised linear modelling. [On-line] http://www.statsci.org/glm/index.html, (accessed 25/09/2002).

[14]Steinberg, D. and Cardell, N. S. (1998a). Improving data mining with new hybrid methods. Presented at DCI Database and Client Server World, Boston, MA.