Hybrid Predictive Modelling for Motor Insurance Claim

Zuriahati Mohd Yunos1, Siti Mariyam Shamsuddin2, Roselina Sallehuddin3, and Razana Alwee4

1,2,3,4School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.

Email : zuriahati@gmail.com

Abstract. The objective of this study is to propose a new hybrid model to predict the Malaysia motor insurance claim by estimating the two important components; claim frequency and claim severity. The proposed model are integrating between grey relational analysis and back propagation neural network. We proposed the hybrid model to handle the issue of the insurance data and the complexity of classical statistical technique. Moreover, the classic statistical techniques are incapable of handling huge information in the insurance data. Thus, hybrid model is proposed because it has a high learning ability and capability to learn. Finally, a comparative analysis is carried out to evaluate the predictive model performance between GRABPNN and BPNN. The results produce by MAPE show a small percentage and thus, show that GRABPNN model provides more accurate predictive results compared to BPNN model.

1. Introduction
Predictive modelling is a process that involves problem identification, analysis of data, model development, as well as prediction and validation [11]. Predictive modelling in the insurance industry helps actuaries and other insurance analysts employing predictive models to enhance business operations that previously used human expertise. Moreover, the influences of predictive modelling are also dependent on the quality of the data used to generate the model. Insurance is a unique type of agreement between the insurer or insurance company and the insured or client in which the insurers permit that upon the occurrence of specific events, whether to make a payment to clients or cover the specific costs. There are two types of insurance such as life insurance and general insurance. Among the primary products in general insurances are motor insurance, fire insurance, personal accident insurance, health care insurance, and traveling insurance. This study focuses on developing a predictive model for motor insurance claim by estimating the two important components, namely, claim frequency and claim severity [4, 6, 8]. Claim frequency is defined as the number of claims per exposure unit, whereas claim severity is defined as the average claim cost per claim [8]. The modelling of claim frequency and claim severity needs an information of exposure, number of claims and amount of claim (cost). The expected claim frequency and claim severity can be calculated through a process of identifying grouping risk, which has the same characteristics (combination of rating factors or variables).

2. Related Work
There are two important issues highlighted in motor insurance claims which are the data and techniques, and these issues are interrelated with one another. The first issue is related to the characteristics of insurance data, which contain huge information or large number of variables, as well as uncertainty, noisy and incomplete information as agreed by [5, 6, 7, 9]. This issued has been discussed in [20] on how to identify the rating factors and rating classes to build the grouping risk. Another study by [23] reported that the the grouping risk can be determined by a number of factors such as the vehicle cubic capacity, the geographic zone where the insured lives, age of the insured. This includes the data of gender criterion, the driver’s age, the vehicle characteristics,
and the driving behaviour [22, 24]. [3] also agreed that insurance data have complex data structures. Another problem is the existing of extreme values in the data which cannot be ignored or dealt as outliers [14]. The work of [18] has thoroughly investigated the insurance data consists a large number of variables and many important variables redundancies between them because they share the same information. The second issue is related to the complexity of statistical analysis that has become more apparent. Due to this, actuaries had to solve the problem of finding a model that can explain realistically the event of risk [3, 4] and a model that is able to handle complex problems in exploiting varying information [12]. There are some statistical methods and mathematical algorithms applied in insurance claim practices such as generalized linear model (GLM) [19], Almer’s simple multiplicative model [15], an additive model in 1960 [16], log-linear model [17]. However, these techniques have limitations such as lack of discriminatory power, restrictions on the independent distribution, or poor ability to interpret [21].

Thus, we recommed a feature ranking using GRA as the technique attempt to maintain each feature which is most likely will be able to provide improvements in predicting accuracy and its not eliminate the features but ranking it. In addition, BPNN is choose as the technique use past experience to train the network to provide more consistent and reliable evaluations on claim frequency and claim severity.

3. Methodology

3.1 Experimental data

The data is provided by Insurance Services Malaysia Berhad (ISM), which is based on 1.2 million policies for the year 2001 until 2003 and are used to evaluate the proposed hybrid model, GRABPNN. The claim data motor insurance consist of two different types: third party property damage (TPPD), and third party bodily injury (TPBI). The input factors or rating factors used for predicting frequency claims and severity are coverage, vehicle cubic capacity, vehicle made, vehicle year and location. Any missing values are eliminated and data is normalized using min-max formula to smoothing data distribution for obtaining better data generalization. Then, the data is partitioned into two parts: training sets and testing sets based on 70%:30% ratio. Table 1 lists all the inputs and output used for claim frequency and claim severity. For claim frequency, each data is used to determine the number of claims made by the insured (clients); and for claim severity, the data are used to compute the amount claimed by the insured or amount paid by the insurer to insured.

| Notation | Claim Frequency | Claim Severity |
|----------|-----------------|----------------|
| F1       | Coverage        | Coverage       |
| F2       | Vehicle made    | Vehicle made   |
| F3       | Vehicle cc      | Vehicle cc     |
| F4       | Vehicle year    | Vehicle year   |
| F5       | Location        | Location       |
| F6       | Exposure        |                |
| F7       |                 | Number of claim |

3.2 Model Development

Figure 1 shows the research framework consist of two phases; feature ranking and model development. In phase 1, feature ranking using GRA is employed to determine the most and least significant features, and rank the six important features as listed in Table 1. The feature ranking is completed by rearranging the GRG value by put the greatest GRG value in the first position; the second largest is located in the second location, and subsequently. In phase 2, the output from the GRA feature ranking is employed and combined with the BPNN and known as, GRABPNN and compare with single model BPNN. The process of integrating the techniques is known as hybridization. The motivation of integration is to understand the underlying relationship between input nodes and hidden nodes, and to enhanced the prediction accuracy. There are nine network structure in GRABPNN are developed and to be tested on claim frequency and claim severity. There are six input nodes and one output node are used as listed in Table 1. The number of hidden nodes are determine by using the number of input nodes (n) and used formula $n, 2n, 2n+1$ [12].
4. Experimental Results and Analysis
This section discusses the result obtained from the experiment. The discussion is divided into two parts. Part 1 discusses on how GRA rank the input and select the optimum features that will be fed into BPNN. The second part will discuss on the result obtained from BPNN using the optimum features. Then, the performance of the proposed model will be compared with the standard BPNN.

4.1 Feature Ranking using GRA
Based on Table 2 and 3, it was found that for each claim types, the most significant features for claim frequency are exposure (F6) and coverage (F1), and for claim severity are number of claim (F7) and coverage (F1) with a stronger correlation close to 1 which are greater than 0.5. The feature ranking result are based on GRG value in descending order and can be seen in the column 'rank' from the most to the less significant.

| Table 2. The GRG value of claim frequency |
|------------------------------------------|

| Claim frequency | TPPD claim | TPBI claim | Rank |
|-----------------|------------|------------|------|
| F6              | 0.97229    | 0.96695    | 1    |
| F1              | 0.64829    | 0.65069    | 2    |
| F4              | 0.58924    | 0.59232    | 3    |
| F5              | 0.58831    | 0.58212    | 4    |
| F3              | 0.57892    | 0.5803     | 5    |
| F2              | 0.57855    | 0.57686    | 6    |

| Table 3. The GRG value of claim severity |
|------------------------------------------|

| Claim severity | TPPD claim | TPBI claim | Rank |
|----------------|------------|------------|------|
| F7             | 0.90598    | 0.90227    | 1    |
| F1             | 0.61384    | 0.61701    | 2    |
| F4             | 0.60368    | 0.60135    | 3    |
| F5             | 0.58931    | 0.59039    | 4    |
| F3             | 0.58403    | 0.58336    | 5    |
| F2             | 0.58205    | 0.58205    | 6    |
4.2 Hybrid Model using GRABPNN

In this study, the model development is done through trial-and-error with the aim to obtain the best predicting results. The number of input nodes is similar to a number of features. The maximum and minimum numbers of input nodes used are 6 and 4. The elimination number of input nodes is done one by one from 6 to 4. The output from the GRA was then fed into the network. Then, a comparative analysis was carried out to evaluate the predictive model performance between GRABPNN and BPNN. The best model is chosen based on four error measurements, mean squared of error (MSE), root mean square of error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The best model is chosen based on four error measurements, mean squared of error (MSE), root mean square of error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE).

Table 4. The best model for claim frequency

| Predict Model | Claim types | N/W   | MSE   | RMSE  | MAE   | MAPE  |
|---------------|-------------|-------|-------|-------|-------|-------|
| GRABPNN       | TPPD        | 6-13-1 | 310.5 | 17.6  | 10.1  | 20%   |
|               | TPBI        | 4-8-1  | 96.1  | 9.8   | 5.3   | 25.7% |
| BPNN          | TPPD        | 6-13-1 | 369.50| 19.22 | 10.73 | 21.9% |
|               | TPBI        | 4-8-1  | 79.13 | 8.90  | 5.33  | 25.8% |

Table 5. The best model for claim severity

| Predict Model | Claim types | N/W   | Claim Frequency |
|---------------|-------------|-------|-----------------|
|               |             |       | MSE  | RMSE | MAE | MAPE |
| GRABPNN       | TPPD        | 6-12-1 | 123158238.8 | 11097.67 | 6112.32 | 51.6% |
|               | TPBI        | 6-12-1 | 133139846.4  | 11538.62  | 5836.05  | 70.56% |

Table 4 and 5 displays the best model with the best network structures for each claim type. It was found that for each claim type for claim frequency and claim severity, the best model is given by GRABPNN model. A small value of MAPE is acquired by GRABPNN compared to BPNN for claim frequency and claim severity (highlighted in grey). Therefore, the studies have proven that the proposed hybrid model is good and acceptable in predicting claim frequency and claim severity. Moreover, the application of GRA to determine and rank the most, and the least significant features have enhanced the accuracy of the proposed hybrid model.

5 Conclusions

This study proposes a new hybrid model for modelling the motor insurance claims. The proposed hybrid model was used and tested on two claims types, namely third party property damage (TPPD) and third party bodily injury (TPBI) claims. Each claim type has two elements, namely, claim frequency and claim severity, and the development of the predictive model has been applied on these two elements. Rating factors are used as an input factor in the proposed hybrid model. The implementation of GRA does not require any assumptions about the data and it is suitable in all situations, although the historical data are limited. The proposed hybrid model, GRABPNN has successfully been done and indeed improves the predictive accuracy. Moreover, the study has gained evidence that, given various numbers of features and hidden nodes, rank the informative features; GRABPNN obtained better performance in modelling claim frequency and claim severity for each claim type as compared to other models.

Acknowledgements

The authors gratefully acknowledge the Fundamental Research Grant Scheme, through grant number FRGS:4F825, Research Management Centre (RMC) of Universiti Teknologi Malaysia and Applied Industrial Analytics (Ali@s) for the support and motivation in making this research a success.
References

[1] Bahia H S I 2013 *Inter. J. Intell. Sci.* 3136.
[2] Batty M, Tripathi A, Kroll A, Peter Wu C-S, Moore D, Stehno C, Lau L, Guszczza J and Katcher M 2010 *Predictive Modelling for Life Insurance* (Deloitte Consulting LLP).
[3] Christmann A 2004 *Allg. Stat. Archiv.* 88 375.
[4] David M 2015 *Proc. Eco. Finance* 20 147.
[5] Dugas C, Chapados N, Ducharme R, Saint-Mleux X. and Vincent P 2011 *Inter. J. Bus. Intell. Data Mining.* 6 237.
[6] Frees EW, Peng Shi P and Valdez EA 2008 *ASTIN Bulletin* 39 165.
[7] Frees EW 2014 *Frequency and severity models.* In E Edward, G Meyers and R A Derrig Cambridge. Cambridge University Press.
[8] Ismail N and Jemain A A 2008 *Sains Malaysia* 37 412.
[9] Kim J H T and Kim J 2014 *J. Appl. Math. & Info.* 32 343.
[10] Kitchen F L 2009 *Inter. J. Electronic Finance* 3 311.
[11] Rosenberg M A, Frees E W, Sun J, Johnson Jr P H and Robinson J 2007 *North American Actuarial J.* 11 54.
[12] Zhang G, Patuwo B E and Hu M Y 1998 *Inter. J. Forecasting* 14 35.
[13] Shapiro A F 2009 *Ins. Math. & Eco.* 44 307.
[14] Aftab S, Abbas W, Bilal M M, Hussain T, Shoaib M and Mehmoood S H 2013 ICDIM p 1.
[15] Almer B 1963 *15th Inter. Congress of Actuaries* 2 314.
[16] Bailey R A 1963 *Proc. of the Casualty Actuarial Society* 50 4.
[17] Chang L and Farley W 1979 *J. of Risk and Insurance.* 46 75.
[18] Chapados N, Dugas C, Vincent P and Ducharme R 2008 *ICDM Workshops* p 97.
[19] David M 2015 *Proc. Eco. & Fin.* 20 147.
[20] Laas D, Schmeiser H and Wagner J 2016 *The Geneva Papers on Risk & Ins.* 41 398.
[21] Lindsey J K 2004 *Introduction to Applied Statistics* (Oxford University Press).
[22] Schmeiser H, St‘ormer T and Wagner J 2014 *The Geneva Papers on Risk & Ins.* 39 322.
[23] Shi P and Valdez E A 2014 *Ins. Math. & Eco.* 55 18.
[24] Wamwea C M, Muema B K and Mung’atu J 2015 *American J. Theo. & Applied Stat.* 4 527.