ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

Estimation of Claim Severity in Non-Life Insurance: A Non-Parametric Approach

K.M. Sakthivel[1], C.S. Rajitha[2]

[1]Assistant Professor, Department of Statistics, Bharathiar University, Coimbatore, Tamilnadu, India

[2]Research Scholar, Department of Statistics, Bharathiar University, Coimbatore, Tamilnadu, India

Abstract— In non-life insurance setting of the right premium for the customer at the beginning of the insurance contract is absolutely necessary for an insurance practice. For that the accurate and authentic estimate of the number of claim occurrences and claims size is extremely important. Different methods are available in the literature for predicting the claim size of a policy for forthcoming years such as Generalized linear models (GLMs), Poisson regression models, Credibility models, Bayesian Models etc. But due to some changes in exposure, classification of rating factors, migration in risk classes, the above mentioned classical methods will not provide a suitable model for prediction of future claim size. Hence a dynamic empirical model will address this problem. Recent studies shown that Artificial Neural Networks (ANN) is powerful tools for prediction by observing variation present in the data and predict future observations based on the characteristics of trained data sets. In this paper, we have shown that ANN will produce relatively better result compare to GLM.

Keywords- Claim Severity, Generalized Linear Model, Artificial Neural Network

I. INTRODUCTION

One of the most important problems in insurance business is to set the premium for the customers. The essential feature of an insurance practice is to set the premium at the beginning of the insurance contract. To determine the correct premium for next year in an insurance company, precise and reliable estimate of the number of occurrence of claims and the total claim amounts is extremely important. In a competitive market, it is advantageous for the insurer to charge a fair premium according to the expected loss of the policyholder. In personal car insurance, for instance, if an insurance company charges too much for old drivers and charges too little for young drivers, then the old drivers will switch to its competitors, and the remaining policies for the young drivers would be underpriced. These results in the adverse selection issue (Dionne et al., 2001): the insurer loses profitable policies and is left with bad risks, resulting in economic loss both ways. Therefore appropriately set the premiums for the insurer's customers are one of the crucial task to predict the size of actual (currently unforeseeable) claims. One difficulty in modeling the claims is that the distribution is usually highly right-skewed, mixed with a point mass at zero. Such type of data cannot be transformed to normality by power transformation, and special treatment on zero claims is often required. Traditional methods used generalized linear models (GLM; Nelder and Wedderburn, 1972) for modeling the claim size (e.g. Renshaw, 1994; Haberman and Renshaw, 1996). However, the authors of the above papers performed their analyses on a subset of the policies, which have at least one claim. Alternative approaches have employed Tobit models by treating zero outcomes as censored below some cutoff points (Van de Ven and van Praag, 1981; Showers and Shotick, 1994), but these approaches rely on a normality assumption of the latent response. Alternatively, Jorgensen and de Souza (1994) and Smyth and Jorgensen (2002) used Generalized linear model (GLMs) with a Tweedie distributed outcome to simultaneously model frequency and severity of insurance claims. They assume Poisson arrival of claims and gamma distributed amount for individual claims so that the size of the total claim amount follows a Tweedie compound Poisson distribution. Due to its ability to simultaneously model the zeros and the continuous positive outcomes, the Tweedie GLM has been a widely used method in actuarial studies (Mildenhall, 1999; Murphy et al., 2000; Peters et al., 2008). Despite of the popularity of the Tweedie GLM, a major limitation is that the structure of the logarithmic mean is restricted to a linear form, which can be too rigid for real applications. In auto insurance, for example, it is known that the risk does not monotonically decrease as age increases (Anstey et al., 2005). Although nonlinearity may be modeled by adding splines (Zhang, 2011), low-degree splines are often inadequate to capture the non-linearity in the data, while high-degree splines often result in the over-fitting issue that produces unstable estimates.

Neural networks (NNs) are being used as an alternative to all these traditional techniques and gaining popularity in recent years. Studies in Artificial Intelligence shows that Artificial Neural Networks (ANN) is powerful tools for prediction and classification, areas where regression models and other related statistical techniques have traditionally been used, due to their nonlinear nonparametric adaptive learning properties. NNs are data dependent and therefore, their performance improves with sample size. But statistical methods, such as Regression perform better only for extremely small sample size.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

Neural networks are being used in areas of prediction and classification, areas where regression models and other related statistical techniques have traditionally been used. A number of researchers have illustrated the connection of neural networks to traditional statistical methods. Ripley (1994) discusses the statistical aspects of neural networks and classifies neural networks as one of a class of flexible nonlinear regression methods. Sarle (1994) translates neural network jargon into statistical terminology and shows the relationship between neural networks and statistical models such as generalized linear models, projection pursuit and cluster analysis. They have explained that neural networks and statistics are not competing methodologies for data analysis and there is considerable overlap between the two fields. Warner and Misra (1996) have presented a comparison between regression analysis and neural networks in terms of notation and implementation. They have also discussed when it is advantageous to use neural network model in place of a parametric regression model, as well as some of the difficulties in implementation. Schumacher, Robner, and Vach (1996) and Vach, Robner, and Schumacher (1996) have presented a comparison between feedforward neural networks and the logistic regression. The conceptual similarities and discrepancies between the two methods are also analyzed. T.E. Dalkilic et al. (2009) have applied NN approach with fuzzy rules instead of using least square approach to describe the relationship between response variables and independent variables of a multiple linear regression model concerning the estimation of future claim payments. And have built an algorithm using adaptive network for finding out the parameters of the regression model. Itedal Sabri Hashim Bhia (2013) shows that the forecasts and estimations of the insurance premiums revenue using ANN adapted to be substantial and utilitarian to distribute it for forecasting the insurance premium revenue.

In this paper, we aim to model the insurance claim size data using Artificial Neural Networks and compare the efficiency of ANN with GLMs. The organization of the paper is as follows: Section 2 gives details about GLMs. In section 3 we provide a detailed description about NNs computation and various types of neural networks. Section 4 gives details of data and software used for this paper. Section 5 provides results. Section 6 contains conclusions.

II. GENERALIZED LINEAR MODELS

GLM is a natural generalization of the familiar classical linear models. The class of GLMs includes, as special cases, linear regression, analysis-of-variance models, log-linear models for the analysis of contingency tables, logit models for binary data in the form of proportions and many others. It is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. The use of classical linear models in actuarial work is not new. Thus, such models have been an established part of the description of claim frequency rates and average claim costs in motor insurance. The use of GLMs in actuarial work can be traced back to the early 1980s. Thus, McCullagh and Nelder (1983, 1989) give many examples of the fitting of GLMs to different types of data, including average claim costs data from a motor insurance portfolio (originally modelled by Baxter et al. (1980) using a weighted least squares approach) and claim frequency data for marine insurance.

In a generalized linear model, each outcome Y of the dependent variables is assumed to be generated from a particular distribution in the exponential family, a large range of probability distributions that includes the normal, binomial, Poisson and gamma distributions, among others. The mean, μ, of the distribution depends on the independent variables, X, through:

$$E(Y) = \mu = g^{-1}(X\beta)$$

where $E(Y)$ is the expected value of Y; $X\beta$ is the linear predictor, a linear combination of unknown parameters β; g is the link function.

In this framework, the variance is typically a function, V, of the mean:

$$Var(Y) = V(\mu) = V(g^{-1}(X\beta))$$

It is convenient if V follows from the exponential family distribution, but it may simply be that the variance is a function of the predicted value. The unknown parameters, β, are typically estimated with maximum likelihood, maximum quasi-likelihood, or Bayesian techniques. In this paper this method is also used for modeling Swedish motor insurance data.

III.NEURAL NETWORKS

NNs have been shown to be very trusted systems in many applications such as prediction, classification, forecasting, and modeling due to their ability to learn from the data and also their nonparametric nature. The NN solution may assist in predicting the reliable claim frequency for the forthcoming year.

A. Neural Computation

ANN is a machine learning technique that is designed to model the way in which the brain performs a particular task or function of interest; the network is usually implemented by using electronic components or is simulated in software on a digital computer (Simon Haykin, 2001). An ANN, usually configured for a specific application, is composed of a large number of highly interconnected parallel processing elements, called neurons, working in agreement to solve specific problems, through a learning process (A. Ibiwoye et.al., 2012). An

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

artificial neuron is an eclectic simulation of biological neuron, and it consists of its own dendrites, synapses, cell body and axon terminals. It receives stimulation from nearby cells, or from its environment, and generates a modified action potential or nerve signal (Ajibola et al., 2011). According to Stergious and Siganos (2007), ANN approach has a unique capability for deriving meaning from complicated or imprecise data and is useful in detecting patterns or trends that are too complex to be noticed by humans or other computer techniques. Agatonovic-Kustrins et al. (2000) explained that the behavior of a neural network is determined by the transfer functions of its neurons, by the learning rule, and by the architecture itself. Figure1 is the architectural framework of the commonly used artificial neural network consisting of layers of input units connected to layer of hidden units which are connected to a layer of output units.

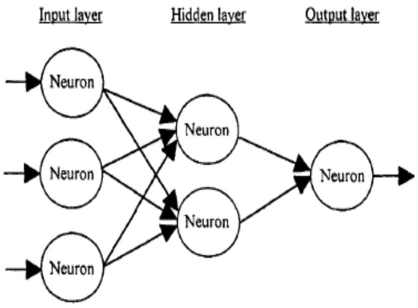

Fig. 1 ANN Overview

The behavior of an ANN depends on both the weights and the input-output (transfer function) that is specified for the units (Stergious and Siganos, 2007).These functions fall into one of three categories namely: linear, threshold and sigmoid functions. Several possibilities of using transfer functions of different types in neural models are discussed in (Duch and Jankwoski, 2001). During training, the inter-unit connections are optimized until the error in predictions is minimized and the network reaches the specified level of accuracy. Nguyen, (2005) investigated the predictive power of various neural network models in predicting corporate failure.

NNs are mostly good at recognizing complex patterns. The central idea of a NN is to extract linear combinations of the inputs as derived features, and then model the target variable as a nonlinear function of these features (T. Hastie, R. Tibshirani, and J. Friedman 2004). The NN algorithm family is quite large. A NN is a statistical model that can be represented by a network diagram.

B. Neural network types

In the last few decades different types of neural networks have been investigated. Among these the feed forward NN was the simplest type of artificial NN. In the feed forward network, the network moves in the forward direction. In the case of feedback networks the networks include connections back to previous layers or even back to the neuron itself. The former class of the network has been used in this paper.

IV. DATA ANALYSIS

The data give details of third party motor insurance claims in Sweden for the year 1977. These data were compiling by the Swedish Committee on the Analysis of Risk Premium in Motor Insurance, summarized in Hallin and Ingenbleek (1983) and Andrews and Herzberg (1985). The data are cross-sectional; describing third party automobile insurance claims. The data set consists of a claim file with 2182 claims. The outcomes of interest are the sum of payments (the severity), in Swedish kroners. Outcomes are based on 6 categories distance driven by a vehicle, broken down by 7 geographic zones, 7 categories of recent driver claims experience, 9 types of automobile and the number of claims (claim frequency). Even though there are 2,205 potential distance, zone, experience and type combinations ($5 \times 7 \times 7 \times 9 = 2{,}205$), only $n = 2{,}182$ were realized in the 1977 data set. Here we estimate the claim size (total payments) using GLM and ANN.

A. Claim size prediction using GLM

For estimating the claim size using GLM first the data set is divided in to two sets randomly, train set and test set. 75% of the data belongs to the train set and remaining 25% of the data belongs to the test set. Here the variable total payments (claim size) are considered as response variable and all other variables are considered as independent variables. And also estimate the total payments by considering all claims and for positive claims using R software.

B. Claim size prediction using ANN

For estimating the claim size using ANN first we train 75% of the given data using the neuralnet package with feed forward NN with back propagation algorithm. During the training phase the model learns how to use some of the fields in a record to predict the value of another field. After that testing 25% of the remaining data using the weights generated by the training process. In this method also payments(claim size) are estimated based on the other variables such as distance driven by a vehicle, broken down by 7 geographic zones, 7 categories of recent driver claims experience, 9 types of automobile, the number of insured in policy years and the number of claims (claim frequency). Also the outcome is estimated for all claims (Fig. 2) and for positive claims (Fig. 3).

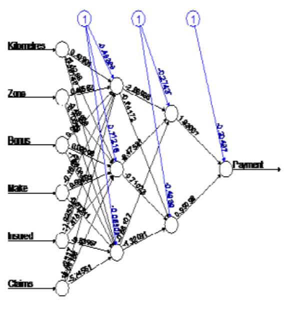

Fig. 2 All Climes

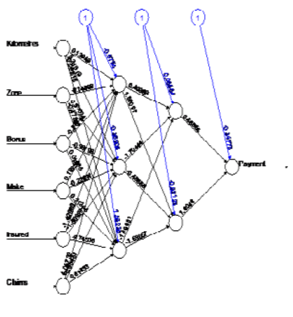

Fig. 3 Positive Climes

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

Here we compare the efficiency of the two methods by using Mean Squared Error (MSE). MSE can be defined as

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(a_j - y_j)^2$$

a_j and y_j represent the target value and network output for the j^{th} training pattern respectively, and N is the number of training patterns.

V. RESULTS

Table 1 shows the MSE of target value and model output of total claim size (payments) using GLMs and ANN. The results shows that when we consider all claims(2182) MSE between target value and network output using ANN is small compared to GLM by considering a single hidden layer and 2 hidden layers. If we consider only the positive claims(1797) MSE using ANN is small compared to MSE using GLM for two hidden layers and large for single hidden layer.

TABLE I. STANDARDIZED MSE USING GLM AND ANN

ANN Setup	Number of claims	MSE	
		GLM	ANN
Hidden=c(3,2)	All Claims	0.000016	0.000016
	+ve claims only	0.000019	0.000016

V. CONCLUSION

As we know the claim size prediction is substantial in insurance practice to set the premium at the beginning of the insurance contract. Though various methods have been developed for different circumstances for prediction purposes, that methods depends on some limiting assumptions such as linearity, normality, independence etc and also takes comparatively more time than rather simplified practical model (ANN) for prediction. Since ANNs are trained using the already available original data the error may be reduced and also Neural Networks is data dependent and their performance improves with sample size. We compared the ANN with GLM, a statistical method for predicting the claim size. Results show that ANN is also suitable for predicting the claim size because the difference between MSE between actual claim size and estimated claim size using ANN is relatively small compare to GLM for a motor insurance data.

REFERENCES

[1] K. J. Anstey, J. Wood, S. Lord and J. G. Walker, "Cognitive, sensory and physical factors enabling driving safety in older adults," Clinical psychology review, vol. 25, pp. 45-65, 2005.

[2] S. Agatonovic – Kustrins, and R. Beresford, (). "Basic Concepts of Artificial Neural Network (ANN) Modeling and its Application in Pharmaceutical Research," Journal of Pharmaceutical Biomedical Analysis, vol. 22 (5), pp. 717-27, 2000

[3] O. O. E. Ajibola, , V. O. S. Olunloyo and O. Ibidapo- Obe, "Artificial Neural Network Simulation of Arm Gait in Huntington's Disease Patient," International Journal of Biomechatronics and Biomedical Robotics, vol. 1 (3), pp. 133-140, 2011.

[4] L. A. Baxter, S. M. Coutts and G. A. F. Ross, "Applications of linear models in motor insurance," In Proc. 21st Int. Congr Actuaries, vol. 2, pp. 11-29. Bern: Stampfli, 1980.

[5] Itedal Sabri Hashim Bahia, "Using Artificial Neural Network Modeling in Forecasting Revenue: Case Study in National Insurance Company/Iraq," International Journal of Intelligence Science, vol. 3, pp.136-143, 2013.

[6] T.E Dalkilic, F. Tank and K.S. Kula, "Neural networks approach for determining total claim amounts in insurance," Insurance: Mathematics and Economics vol. 45, pp. 236-241, 2009.

[7] G. Dionne, C. Gourioeroux, and C. Vanasse, "Testing for evidence of adverse selection in the automobile insurance market: A comment," Journal of Political Economy, vol. 109, pp. 444-453, 2001.

[8] N. Duch , N. Jankowski, "Transfer Functions Hidden Possibilities for Better Neutral Networks," In 9 th European Symposium on Artificial neutral Networks, Belgium, pp. 81-94, 2001.

[9] S. Haberman, and A. E. Renshaw, "Generalized linear models and actuarial science," Statistician, vol. 45, pp. 407-436, 1996.

[10] T. J. Hastie and R. J. Tibshirani, Generalized additive models, vol. 43, CRC Press. 1990.

[11] A. Ibiwoye, O. O. E. Ajibola and A. B. Sogunro, "Artificial Neural Network Model for Predicting Insurance Insolvency". International Journal of Management and Business. Res., vol. 2 (1), pp. 59- 68, 2012.

[12] B. Jorgensen, and M. C. de Souza, "Fitting Tweedie's compound Poisson model to insurance claims data," Scandinavian Actuarial Journal, pp. 69-93, 1994.

[13] P. McCullagh, and J. A. Nelder, Generalized Linear Models, London: Chapman and Hall, 1983.

[14] P. McCullagh, and J. A. Nelder, Generalized Linear Models, 2nd edn. London: Chapman and Hall, 1989.

[15] S. J. Mildenhall, "A systematic relationship between minimum bias and generalized linear models," In Proceedings of the Casualty Actuarial Society, vol. 86, pp. 393-487, 1999.

[16] K. P. Murphy, M. J. Brockman and P. K. Lee, "Using generalized linear models to build dynamic pricing systems," In Casualty Actuarial Society Forum, Winter, pp. 107-139, 2000.

[17] J. Nelder, and R. Wedderburn, "Generalized Linear Models," Journal of the Royal Statistical Society. Series A (General), vol. 135, pp. 370-384, 1972

[18] H.G. Nguyen, "Using Neutral Work in Predicting Corporate Failure".Journal of Social Sciences, vol. 1(4), pp. 199-202, 2005.

[19] G. W. Peters, P. V. Shevchenko and M. V. Wuthrich, "Model uncertainty in claims reserving within Tweedie's compound Poisson models," ASTIN Bulletin, vol. 39(1), pp. 1-33, 2009.

[20] A. E. Renshaw, "Modelling the claims process in the presence of covariates," ASTIN Bulletin, vol. 24, pp. 265-285, 1994.

[21] B. D. Ripley, "Neural networks and related methods for classification," Journal of the Royal Statistical Society, Series B (Methodological), vol. 56(3), pp. 409–456, 1994.

[22] W. S. Sarle, "Neural networks and statistical models," In Proceedings of the 19th annual SAS users group international conference, 1994.

[23] V. E. Showers, and J. A. Shotick, "The effects of household characteristics on demand for insurance: A tobit analysis," Journal of Risk and Insurance, pp. 492-502, 1994.

[24] M. Schumacher, R. Robner and W. Vach, "Neural networks and logistic regression: Part I," Computational Statistics & Data Analysis, vol. 21(6), pp. 661–682, 1996.

[25] Simon Haykin, Neural Networks, A comprehensive foundation, Second Edition, Prentice Hall, Pearson Education, Singapore, 2001.

[26] G. Smyth, and B. Jorgensen, "Fitting Tweedie's compound Poisson model to insurance claims data: Dispersion modelling," ASTIN Bulletin, vol. 32, pp. 143-157, 2002.

[27] W. Vach, R. Robner and M. Schumacher, " Neural networks and logistic regression: Part I," Computational Statistics & Data Analysis, vol. 21(6), pp. 683–701, 1996.

[28] W. Van de Ven, and B. M. van Praag, "Risk aversion and deductibles in private health insurance: Application of an adjusted tobit model to family health care expenditures," Health, economics, and health economics, pp. 125-48, 1981.

[29] B. Warner and M. Misra, "Understanding neural networks as statistical tools," The American Statistician, vol. 50(4), pp. 284–293, 1996.

[30] W. Zhang, "cplm: Monte Carlo EM algorithms and Bayesian methods for fitting Tweedie compound Poisson linear models," R package, http://cran.r-project.org/web/packages/cplm/index.html, 2011.