Journal of Applied Finance & Banking, vol. 4, no. 5, 2014, 45-53
ISSN: 1792-6580 (print version), 1792-6599 (online)
Scienpress Ltd, 2014

Using Multi-class AdaBoost Tree for Prediction Frequency of Auto Insurance

Yue Liu[1], Bing-Jie Wang[1,2] and Shao-Gao Lv[3]

Abstract

In this paper, AdaBoost algorithm, a popular and effective prediction method, is applied to predict the prediction of claim frequency of auto insurance, which plays an important part of property insurance companies. Using a real dataset of car insurance, we reduce the frequency prediction problem to be a multi-class problem, in turn we employ the mixed method called multi-class AdaBoost tree (a combination of decision tree with adaptive boosting) as our predictor. By comparing its results with some most popular predictors such as generalized linear models, neural networks, and SVM, we demonstrate that the AdaBoost predictor is more comparable in terms of both prediction ability and interpretability. The later objective is particularly important in business environments. As a result, we arrive at the conclusion that AdaBoost algorithm could be employed as a robust method to predict auto insurance. It is important to practical contribution for insurance company in terms of conclusion explanation and decision making suggestions.

JEL classification numbers: C52, G11
Keywords: Auto Claim, GLM, Data Mining, Neural networks, Decision Tree, Multi-Class AdaBoost

1 Introduction

Automobile insurance plays an important roles in current non-life insurances. However, ratemaking is a complex and difficult task for various reasons. Firstly, many factors such as vehicle make and vehicle use are relevant. Only considering each of them individually as independence can be hurtful [7]. On the other hand, taking account of all interactions is intractable and is sometimes suffered from the curse of dimensionality [17]. Secondly, an

[1]College of Finance, Southwestern University of Finance and Economics, Cheng Du, 611130, China.
[2]Statistics School, Southwestern University of Finance and Economics, Cheng Du, 611130, China. The second author is the corresponding author.
[3]Statistics School, Southwestern University of Finance and Economics, Cheng Du, 611130, China.

Article Info: *Received :* June 11, 2014. *Revised :* July 9, 2014.
Published online : September 1, 2014

another difficulty comes from the distribution of claims: asymmetric with fat tails with a large majority of zeros and a few unreliable and very large values, i.e., an asymmetric heavy tail extending out toward high positive values. Modeling data with such a distribution is quite difficult because outliers, sampled from the tail of the distribution, have a strong influence on parameter estimation. Thirdly, one more difficulty is due to the non-stationary nature of the relationship between explanatory variables and the expected claim amount. This has an important effect on the methodology to use, in particular, with respect to the task of model selection. Of course, there are many other problem to be faced with when modeling data from auto insurance, and we refer the readers to the references [17, 7] for details.

There exist various auto insurance modelling literatures on such models [1, 12, 10, 22]. For example, Generalized Linear Models (GLMs) [1] are widely used for building insurance models. These models are based on a traditional approach to statistical modeling where the available data are drawn from a given stochastic data model (e.g., Gaussian, Gamma, Poisson, etc.). They are attractive because of producing interpretable parameters which are combined in a multiplicative fashion to obtain an estimate of loss cost, defined here as the portion of the premium which covers losses and related expenses.

In the past two decades, with rapid development in computation and information technology, an immense amount of data has been created. The field of statistics was required a burning desire for new tools, so as to analyze the increasing size and complexity hidden implicitly among the data. Most of these tools originated from an algorithmic modeling culture rather than a data modeling culture (Brieman, [2]). In contrast to data modeling, algorithmic modeling does not assume any specific model for the data, but treats the data mechanism as unknown. As a result, algorithmic models significantly increase the class of functions that can be approximated relative to data models, and useful information or structure in the data can be extracted automatically. Most popular approaches such as neural networks, SVM(support vector machine) and decision tree have emerges, gain a lot of successful application in many fields. On the whole, they are more efficient in handling large and complex data sets and in fitting non-linearities to the data. However, probably because of this lack of interpretability in most algorithmic models, their application in terms of social science problems have been very limited so far. As for auto insurance modelling, as far as we know, Chapados et al. [6] used several data-mining methods to estimate car insurance premiums. Francis [9] illustrates the application of neural networks to insurance pricing problems such as the prediction of frequencies and severities. Kolyshkina, Wong, and Lim [11] demonstrate the use of multivariate adaptive regression splines (MARS) to enhance GLM building.

Recently, The application of Boosting [10] to insurance pricing has gained good performance in perdition accuracy. Yet it is worth noting that the setting considered in [10] is the limited binary classification or regression problem. In this paper, we will foucs on the multi-class classification problem, since the response takes values in discrete integers. As pointed out by Freund and Schapire [15], straightforward extensions of the binary weak-learning condition to multiclass do not work. Requiring less error than random guessing on every distribution, as in the binary case, turns out to be too weak for boosting to be possible when there are more than two labels. On the other hand, requiring more than 0.5 accuracy even when the number of labels is much larger than two is too stringent, and simple weak classifiers like decision stumps fail to meet this criterion, even though they often can be combined to produce highly accurate classifiers.

In this study, we experimented with using a relatively new learning method for the field of

credit rating prediction, multi-class AdaBoost [4], to predict claim frequency. We were also interested in interpreting the models and helping users to better understand bond raters behavior in the bond-rating process. The remainder of the paper is structured as follows. A background section about statistical learning follows the introduction. Then, a literature review about boosting is provided, followed by descriptions of the analytical methods. We also include descriptions of the data sets, the experiment results and analysis followed by the discussion.

2 Analytical Methods

A branch of statistical learning (or machine learning) is mainly concerned with the development of proper refinements of the regularization and model selection methods in order to improve the predictive ability of algorithms. This ability is often referred to as generalization, since the algorithms are allowed to generalize from the observed training data to new data. One crucial element of the evaluation of the generalization ability of a particular model is the measurement of the predictive performance results on out-of-sample data, i.e., using a collection of data, disjoint from the in-sample data that has already been used for model parameter estimation. We provide some brief descriptions of three methods in this section, and focus more on adaptive boosting algorithms, adopted in this paper.

2.1 Neural Networks

Neural networks have been extremely popular for their unique learning capability and have been shown to perform well in different applications in our previous research such as medical application and game playing [5]. One of popular neural networks is called to be back propagation neural network, which consists of a three layer structure: input-layer nodes, output-layer nodes and hidden-layer nodes. Back propagation networks are fully connected, layered, feed-forward models. Activations flow from the input layer through the hidden layer, then to the output layer. A back propagation network often begin with a random set of weights. The network adjusts its weights each time it sees an input-output pair. Each pair is dealt with at two stages, a forward pass and a backward pass respectively. The forward pass involves presenting a sample input to the network and letting activations flow until they reach the output layer. During the backward pass, the networks actual output is compared with the target output and error estimates are computed for the output units. The weights connected to the output units are adjusted usually by a gradient descent method. The error estimates of the output units are then used to derive error estimates for the units in the hidden layer. Finally, errors are propagated back to the connections stemming from the input units. The back propagation network updates its weights incrementally until the network converges. The main drawbacks for neural networks are that only local solution is found, and tend to lead to overfitting, also is short of interpretability. For algorithm further details, we refer the readers to reference [3].

2.2 Support Vector Machine

Support vector machine (SVM) is a novel learning machine introduced first by Vapnik [21]. It is based on the Structural Risk Minimization principle from computational learning

theory. It contains a large class of neural nets, radial basis function (RBF) nets, and polynomial classifiers as special cases. Yet it is simple enough to be analyzed mathematically, because it can be shown to correspond to a linear method in a high dimensional feature space nonlinearly related to input space. In this sense, support vector machines can be a good candidate for combining the strengths of more theory-driven and easy to be analyzed conventional statistical methods and more data-driven, distribution free and robust machine learning methods.

Assume that There is an input space, denoted by $X \subseteq R^{\mathrm{d}}$, and a corresponding output space, denoted by $Y \subseteq R$,and a training set, denoted b D, $D = ((x_1, y_1),(x_2, y_2),....,(x_n, y_n))$, n is the sample size. The task of classificaiton is to construct a heuristic function $\mathrm{f}(x) \approx y$ on the population distribution. The nature of the output space Y decides the learning type. $Y = \{\text{-}1,1\}$ leads to a binary classification problem, $Y = \{1,2,3,...K\}$ leads to a multiple class classification problem, and $Y \subseteq R^{\mathrm{d}}$ leads to a regression problem. SVM belongs to the type of maximal margin classifier, in which the classification problem can be represented as an optimization problem,

$$\min_{\omega,b,\xi} \langle \omega, \omega \rangle + C \sum_{i=1}^{n} \xi_i \tag{1}$$

subject to $\mathrm{y_i}(\langle \omega, \phi(x_i) + b \rangle) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,2,...,n$ (2)

Where ϕ: $X \to F$ is a nonlinear map given in advance from X to high dimensional feature F. By the dual technique, we can rewrite the above problem as the following equivalent form

$$\mathrm{maxW}(\alpha) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}} - \frac{1}{2} \sum_{i,j=1}^{\mathrm{n}} y_i y_j K(x_i, x_j) \tag{3}$$

$$\text{s.t.} \sum_{i=1}^{\mathrm{n}} yi\alpha i = 0, 0 \leq \alpha i \leq C, i = 1,2,...,n \tag{4}$$

Where $K(x,u) = \langle \phi(x), \phi(u) \rangle$ is called a kernel function. This dual formulation is belongs to classical quadratic convex optimization, and many existing software can solve it efficiently.

2.3 Adaptive Boosting

Boosting (Schapire and Freund, [15]) refers to a general technique of combining rules of thumb, or weak classifiers, to form highly accurate combined classifiers. This idea, known as the strength of weak learnability (Schapire [16]), is the inspiration of boosting. Among so many boosting algorithms, the most classical and popular one is AdaBoost. The AdaBoost algorithm proposed by Freund and Schapire [9], with its various versions proven to be highly competitive in terms of prediction ability in a wide range of applications, has draw much attention in the machine learning community as well as in relative areas of statistics. AdaBoost is the abbreviation for adaptive boosting, which shows an essential feature of the algorithm. In the iterative process of AdaBoost, misclassified observations will get larger weight in building the next classifier, so the algorithm can adapt to previous mistakes and correct its training error in successive iterations. In this way, the algorithm

will generate a sequence of weak classifiers with different votes to form a final classifier system, where classifiers with smaller training error will have more votes.

Consider a binary classification problem, given the training dataset D as above, the algorithm of AdaBoost is outlined in Table 1.

Friedman, Hastie and Tibshirani [8] explained the dramatic improvements in prediction performance of AdaBoost with well-known statistical principles, namely additive modeling and maximum likelihood. They showed that for a two-class problem, boosting can be viewed as an approximation to additive modeling on the logistic scale using maximum Bernoulli likelihood as a criterion. When the basic classifier of AdaBoost is decision tree, the algorithm becomes AdaBoost tree. Using additive model, we can express AdaBoost tree as , where stands for decision tree, M is the number of trees, and is the parameter of decision tree. AdaBoost tree is most commonly used among boosting algorithms because its simple, and it can generate predictions of high accuracy as well as good interpretability.

Using the strategy of one-versus-all technique (Culp, Michailidis and Johnson [4]), the AdaBoost algorithm can handle a multi-class problem. For a K-class problem, we model each class against the remaining class to generate K different subsystems. Then run the subsystems simultaneously, and compare the values of $(\mathrm{k}=1,2,...\mathrm{K})$ returned by each subsystem. The value corresponding to the maximum will be the class label.

Table 1: the Algorithm of AdaBoost

Algorithm: AdaBoost

Input: Training dataset D; weak learners;

Output: Final classifier $G(x)$.

1)Initialize the weight distribution of observations $D_1 = \{w_{11},..., w_{1i},..., w_{1N}\}$,where $w_{1\mathrm{i}} = \frac{1}{\mathrm{N}}, i = 1,2,..., N$;

2)For $m=1$ to M do

2.1)Study training dataset D using weight distribution D_m and get the basic classifier $Gm(x) =: x \to \{-1,1\}$; 2.2) Compute the error rate of $G_m(x)$ by $\varepsilon_{(m)} = P[G_m(x_i) \neq y_i] = \sum_{i=1}^{N} I(G_m(xi) \neq y_i)$, where $\sum_{\mathrm{i}=1}^{N} \omega_{mi} = 1$ for each $\mathrm{m} \in \{1,2,..., M\}$;

2.3)Compute the parameter of by $a_m = \frac{1}{2}\ln(\frac{1-\varepsilon_m}{\varepsilon_m})$;

2.4)Update weight distribution $D_{\mathrm{m}} \le D_{\mathrm{m+1}} = \{\omega_{m+1,1},..., \omega_{m+1,N}\}$,

Where $\omega_{m+1,i} = \frac{\omega_{mi}\exp(-a_m y_i G_m(x_i))}{Z_m}, i = 1,2,..., N$

and $Zm = \sum_{i=1}^{N} \omega_{mi}\exp(-a_m y_i G_m(x_i))$ is a normalized factor.

2.5) **End for**

3)Return final classifier $G(x) = sign(\sum_{m=1}^{M} a_m G_m(x))$.

Compared with those data mining algorithms providing comparable predictive accuracy, AdaBoost tree can also produce interpretable results (Friedman, Hastie and Tibshirani [8]). The rules generated by a single decision tree are very easy to understand, but the AdaBoostn tree is a combination of many trees, thus making it difficult to catch the discovered information. To solve this problem, the measure of variable importance is used . With this measurement, we can know the relative influence of the explanatory variables on the response, and that will make it easier to understand the knowledge we get from AdaBoost tree.

3 Problem Description

The data used in this study was extracted from a research paper by Ismail and Jemain (2007). It is the data for private car Third Party Property Damage claim frequencies from an insurance company in Malaysia. Explanatory variables are all categoric variables, including coverage type, vehicle make, vehicle use and driver gender, vehicle year and location. Classes of the variables are shown in Table 2. Altogether, there are 240 risk classes. The respond variable is the claim counts of each class.

Table 2: Explanatory variables and their class

Explanatory variables	Classes
Coverage type	Comprehensive or Non-comprehensive
Vehicle make	Local or Foreign
Vehicle use and driver gender	Private-male or Private-female or Business
vehicle year	0-1year, 2-3year, 4-5year or 6+year
Location	central, North, South or East

In this paper, we reduce the frequency prediction problem to be a multi-class problem by classifying claim counts into 3 classes. The first class is labeled as zero, including 132 observations with the claim counts of 0, the second class is labeled as twenty, including 50 observations with the claim counts of 1 to 20, and the third class is labeled outtwenty, including 58 observations with the claim counts larger than 20. It is worth noting that that the class distribution is imbalanced, and this will have some influence on the prediction.
In this study, GLM (multinomial logistic regression), two-layer BP network and Gaussian radial basis kernel SVM are compared with our proposed algorithm. Also, we have compared AdaBoost tree with decision tree to validate the boosting effect. Using R and Rattle, we divide 240 samples into two parts at random, one as training set (including 110 samples), another as testing set (including 130 samples). Using the training set we build a model with a certain algorithm, and with the testing set we make predictions of target value. In this way we get an error matrix, and then we calculate the error rate of prediction, so accuracy rate is (1-error rate). We implement each algorithm 5 times in succession, calculate and record the accuracy rate of each time. Finally, we evaluate each model with the mean and variance of the accuracy rate. For BP network, we choose the maximum number of iterations to be 200 as the stopping criterion, and the initial random weights are set on [-0.1, 0.1]. For SVM, we choose the cost of constraints violation C to be 10 and kernel width for the radial basis kernel to be 0.2. The minsplit and cp parameter play an important role in the performance of decision tree, and we select 15 and 0.02 respectively,

and we set minbucket, which is suggested to be 1/3 of minsplit (Williams [20]), to be 5. For the stopping criterion of multi-class AdaBoost tree, we allow 800 iterations at most.

4 Experiment Results and Analysis

The paper first validates the boosting effect of multi-class AdaBoost tree by comparing it with decision tree. Table 3 presents the result. As is seen, our proposed algorithm has improved the average predictive accuracy by 3.7 while keeping a small predictive variance. The result supports the theory we introduced before.

Table 3: The prediction performances

Algorithm	Average Prediction accuracy	Variance of prediction accuracy
Decision tree	0.8092	0.0008
Multi-class AdaBoost tree	0.8462	0.007

As we emphasize, one of the advantages of AdaBoost tree is that it can generate a prediction model of interpretability. We calculate the variable importance measurement, and Table 4 presents the result. Although the difference between the score of each variable is relatively small, we can still see that coverage type is the most important variable in predicting claim frequency, which is in accord with practical intuition.

Table 4: Importance of variables

Variables	Score
Coverage type	0.0009
Vehicle make	0.0008
Vehicle use and driver gender	0.0008
vehicle year	0.0007
Location	0.0007

We have also compared multi-class AdaBoost tree with GLM, neural networks and SVM. Table 5 presents a comprehensive comparison of all the algorithms. We can see from the result that multi-class AdaBoost tree outperforms GLM, neural networks and SVM in predictive accuracy by 2.91, 4.61 and 1.53, respectively. More importantly, multi-class AdaBoost tree is also
competitive in the sense of model interpretation.

Table 5: Comprehensive comparison of all the algorithms

Algorithm	Prediction accuracy	Variance	Model interpretation
GLM	0.8185	0.0008	Yes
Neural networks	0.8015	0.0012	No
SVM	0.8323	0.0005	No
Decision tree	0.8092	0.0008	Yes
Multi-class AdaBoost tree	0.8462	0.0007	Yes

5 Conclusion

This study reduced the claim frequency prediction problem of auto insurance to be a multi-class problem, and used multi-class AdaBoost tree for the prediction. We compared the predictive accuracy of our proposed algorithm with that of decision tree to validate the boosting effect. The experimental result showed that average predictive accuracy of AdaBoost tree increased significantly. In addition, this study compared multi-class AdaBoost tree with GLM, neural networks and SVM. The experimental result showed that AdaBoost tree was more competitive in terms of both prediction ability and model interpretation. From the results, we concluded
that AdaBoost tree provides a promising alternative for prediction of claim frequency of auto insurance. There are several directions in which the study could be improved in future research. Firstly, in this study, we reduced the frequency prediction problem to be a multi-class problem, and the classification of claim frequency is subjective. In further study, we will consider the method of AdaBoost regression tree for frequency prediction, and this may be more helpful for decision making of insurance companies. Besides, although there is other information concerning claim frequency of auto insurance, this study did not use much of it in model construction, so we may consider some other machine learning algorithm based on Bayes risk in the future research, and this may help better characterize the relationship between claim frequency and risk classes.

ACKNOWLEDGEMENTS: This research is partly supported by The National Natural Science Foundation of China (No.11301421), and the Fundamental Research Funds for the Central Universities (No. JBK140210).

References

[1] Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D. Schirmacher, E., Thandi, N. (2007). A practitioners guide to generalized linear models. Casualty Actuarial Society (CAS), **9**, 1–116.

[2] Brieman, L. (2001). Statistical modeling: The two cultures. Statistical Science, **16**, 199–231.

[3] Bishop C.M. (1995) Neural Networks for Pattern Recognition, Oxford Univ. Press, New York.

[4] Culp M, Michailidis G, Johnson K. (2009) On multi-view learning with additive models. The Annals of Statistics, **3**, 1–489.
[5] Chen H. Buntin P., She L., Sutjahjo S., Sommer C., Neely D. (1994) Expert prediction, symbolic learning, and neural networks: an experiment on greyhound racing, IEEE Expert, **9** 21–27.
[6] Chapados, N., Bengio, Y., Vincent, P., Ghosn, J., Dugas, C., Takeuchi, I., et al. (2001). Estimating car insurance premia: A case study in high-dimensional data inference. University of Montreal, DIRO Technical Report, 1199.
[7] Derron M. A theoretical study of the no-claim bonus problem. ASTIN Bulletin, 1963, **3**
[8] Friedman J, Hastie T, Tibshirani R.(2000) Additive logistic regression: a statistical view of boosting . The Annals of Statistics, **28** 337–407.
[9] Freund Y, Schapire R E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.(1997) Journal of Computer and System Sciences. **55**, 119–139.
[10] Guelman L. Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Systems with Applications, 2012, 39
[11] Kolyshkina, I., Wong, S., Lim, S. (2004). Enhancing generalised linear models with data mining. Casualty Actuarial Society 2004, Discussion Paper Program.
[12] Ismail N, Jemain A A. Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models. (2007) Casualty Actuarial Society Forum Casualty Actuarial Society.
[13] Meng S W. (2012) Neural networks and prediction of claim frequency of auto insurance. Statistical Research, **29** 23–53.
[14] Ruobonen M. (1988) A model for the claim number process. ASTIN Bulletin, **18**,12–34.
[15] Robert E. Schapire and Yoav Freund. (2012) Boosting: Foundations and Algorithms. MIT Press.
[16] Schapire R E. (1990) The strength of weak learnability. Machine Learning (Boston, MA: Kluwer Academic Publishers), **5**,28–33.
[17] Tremblay L. (1992) Using the Poisson inverse Gaussian in Bonus-Malus system. ASTIN Bulletin, **22**, 87–99.
[18] Willmot G. (1986) Mixed compound poisson distribution. ASTIN Bulletin, **16**, 124–143.
[19] Walhin J F, Paris J. (1999) Using mixed Poisson processes in connection with Bonus-Malus system. ASTIN Bulletin, **29**, 81–99.
[20] William G. (2011) Data Mining with Rattle and R. Springer Science+Business Media, LLC.
[21] Vapnik V. (1995) The Nature of Statistical Learning Theory, Springer-Verlag, New York.
[22] Virginia Wheway. (2004) Variance reduction trends on boosted classifiers, Journal of Applied Mathematics, **8**, 141–154.