

Optimizing SIU Return On Investment Through Data Mining

By Robert A. Threlfall & William J. Lundy

Revised January 19, 2004

Part I. Introduction

Advances in the past decade in computing power and algorithms for machine learning and data mining have made powerful optimization techniques practical and even routine. We give one example here of optimizing the Special Investigation Unit (SIU) return on investment (ROI) ratio once a sufficient number of claims have been analyzed by SAFESTONLINE.

Prediction of an individual claim being fraudulent can be done using various classification and prediction models. Some common approaches include Decision trees, Bayesian Discriminant with conditional independence assumption (also termed class conditional independence), Bayesian belief networks, logistic regression (a type of generalized linear model – GLM) and neural networks. These models have been used successfully in a wide range of applications, such as medical diagnosis^{i,ii} (often doing better than a human expert), credit card fraud detectionⁱⁱⁱ, insurance fraud prediction^{iv}, bankruptcy prediction^v and direct marketing^{vi}.

The general approach is to gather sample learning cases where each case is a database record (a data tuple), consisting of predictive attributes (variables) and a known outcome. For example, for auto insurance fraud, a simple data set could consist of the predictive attributes: age, sex, employment status, annual income, policy inception date, car year and model, and the outcome likelihood class. For the training (learning) dataset the outcome likelihood classification is known and could be categorized by the attributes low, medium or high. From this training set a predictive model is constructed. This predictive model is then validated on another data set with known outcomes. If the model is accurate, it is then used on claims for which the likelihood of fraud is unknown, i.e., previously unseen data, where the outcome classification attribute of either low, medium or high fraud likelihood is unknown.

Part II. Comparison of commonly used models

Each of these different classification and prediction methods have their strengths and weakness. In many applications the above methods are of comparable accuracy. They can be compared according to predictive accuracy, speed, robustness (ability to make correct predictions given noisy or missing data), scalability (the ability to construct the model efficiently given large amounts of data), flexibility (ability to alter and improve the model easily), parsimony (number of variables and parameters used, both explicit and hidden – the ideal model is neither under- nor over-fitted), and level of understanding and insight provided by the model (interpretability).

Decision trees

While decision trees are commonly used and can easily be converted to classification rules (if then rules), they are difficult to scale unless advanced algorithms are used, and they can be quite large and “bushy” unless attribute-oriented induction (AOI) is used. AOI replaces lower-level data with higher-level concepts. For example, the car model attributes: Chevrolet Camero, Chevrolet Cavalier, Ford Mustang, Nissan Altima and so on could be combined into the more general category of “moderately priced sports cars” to help prune the decision tree.

Bayesian Discriminant

Bayesian Discriminant’s strength is that it is rational, fast (hence scaleable for large data sets), simple, can easily accommodate a large number of variables and can be easily automated by computer software. Bayesian Discriminant is theoretically the most accurate classifier provided the predictive variables are independent and the sample data set is representative of the populationⁱⁱⁱ (see endnote i for a real world example). Empirical studies have found its accuracy in some domains to be comparable to decision trees and neural networks^{i,ii,iii}. The accuracy decreases the more the predictive variables are correlated. This can be partially mitigated by adding new variables consisting of the interactions of the correlated variables. This technique should be used sparingly. It is generally better to instead use a generalized linear model such as: i) the logistic regression model as described below, or ii) the log-linear model.

The Bayesian Discriminant model is particularly well suited for insurance fraud detection systems using red flags (indicators). This model scales well, both in terms of handling a large number of claims and a large number of indicators. Finally, because of the conditional independence assumption, it is one of the few models available where the addition or removal of indicators, regardless of whether they are red flags or white flags (exculpatory in nature), does not require the model to be rebuilt.

In its simplistic form, the Bayes Discriminant model consists of binary yes or no attributes, and for each attribute the yes and the no value are assigned a weight, which may be positive or negative. For each case the appropriate weights are summed and membership in the outcome classification is assigned according to whether the sum is below or above a pre-defined threshold. In other words, add the values of the selected flags. Does the final total reach a predetermined threshold? If yes, assign it for investigation, otherwise don’t assign it.

Increasing the threshold reduces the number of false positives (claims that the model considers in need of investigation, but in fact are legitimate), but at the expense of more false negatives (missed fraudulent claims). Conversely, lowering the threshold reduces the number of false negatives, but increases the number of false positives. (Ideally a model should have both a low false positive rate and a low false negative rate.)

The Bayes weights are calculated according to Bayes Theorem^{vii}. This has a beneficial consequence. By applying

one additional mathematical transformation the model will yield an actual predictive probability. Thus instead of just categorizing a claim as possibly being fraudulent or not, the Bayes Discriminant model can assign the claim an actual probability of being fraudulent. In part IV we will show how this can be used to optimize return on SIU investment.

Bayesian belief networks

Bayesian belief networks can be highly accurate, the predictive variables are not assumed to be independent and they provide a very clear rationale for their prediction, however they are more difficult to construct automatically (expert assessment usually improves the model) and they can be quite complex. They incorporate cause and effect relationships. For example, they are very useful for medical diagnosis^{viii} as the cause and effect relationships between risk factors and diseases are encoded by the way the network is structured.

Neural networks

Neural networks were developed initially as computational analogs of neurons and are able to emulate some of the human brain's pattern recognizing abilities. There are many types of neural networks. The most common one consists of an input layer of nodes connected to typically one hidden layer of nodes which is connected to the output layer of nodes. Each node connection has a weight that is learned during the training period. The input and output layers are fixed by the data, but the hidden layer topology is determined by trial and error. The hidden and output nodes also have a nonlinear transformation function. Given enough hidden nodes a neural network can closely approximate any function.

They are typically accurate even with noisy data and they can respond correctly to patterns that are only broadly similar to the training patterns. They can infer subtle, unknown relationships from data. They are nonlinear and able to solve some complex problems more accurately than linear methods.

Many of the traditional drawbacks of neural networks for data mining are being eliminated through the development of new algorithms and new kinds of neural networks. The cascade correlation algorithm developed by Scott Fahlman at Carnegie Mellon University^{ix} eliminates the trial and error traditionally needed to determine the best number of hidden nodes and their topology. Neural networks traditionally have been viewed as black boxes providing no human understandable basis for their predictions. This limitation is improving. For example, sensitivity analysis^{xiii} can be used to assess the impact a given input variable has on a network output. This analysis can lead to if then type rules amenable to direct understanding.

Drawbacks still exist, such as long and computationally intensive training, susceptibility to learning something different from what its trainer had in mind and memorizing as opposed to learning.

A potential drawback to using neural networks for fraud detection is the lack of training data available for less common outlier events which can altar the outcome classification. With hundreds of indicators of varying

frequency available a great deal of training data is needed to prevent improper learning due to "confusing" combinations of indicators. For example, an insufficient amount of training and validation data could cause a neural network to not learn an important but infrequent indicator (II). This could occur if due to "unfortunate sampling" II was largely present in claims that also included indicators the neural network had previously learned were important. The more indicators used in the system the more likely this improper learning could happen.

For the purposes of fraud detection more advanced neural networks are preferred. An adaptive resonance theory (ART) neural network^x is capable of continuous learning in data that changes over time. This would be useful for organized fraud detection as identities and methods change over time.

Logistic regression

Logistic regression models are amenable to rational interpretation, can be largely or completely automatically constructed and the amount of uncertainty in the prediction can be estimated. Their predictive power can be better than a neural networkⁱⁱ. In addition to their use as a classification tool, they are widely used to compute relative risk in epidemiologic studies. With the additional knowledge of the baseline risk they can be used to compute the actual absolute risk probability as well. This can be used for fraud prediction as well.

The logistic regression model in its simplest form is structurally equivalent to the Bayesian discriminant model, but is optimized differently. In its more advanced form it can incorporate high dimension attribute influences and nonlinear interactions among multiple attributes. In essence the logistic function is an index measuring multiple contributing risk factors. Also, inherent in the logistic model is the idea of risk thresholds. In many applications the risk is considered minimal until a critical threshold is reached, then the risk is considered to rise rapidly until it tapers off at a second threshold where the risk is by and large maximal and only increases slightly with additional risk factors.

Piecewise linear regression

In piecewise linear regression the data being modeled is segmented into intervals such that for each segment the relationship between the predictive variables and outcome variable is linear or close to linear. Thus each segment is modeled by a set of rules that indicate what data intervals the given linear model is to be applied to. In the simple case, for a given segment, all predictive variables except for one are fixed. Thus, each segment is a two dimensional linear snapshot and can be easily plotted and interpreted.

Part III. SAFESTONLINE's current model

SAFESTONLINE's current model is comparable to the above Bayesian Discriminant model. Instead of a data driven computation of model parameters (Bayes weights) an in-depth expert assessment was used instead. SAFESTONLINE was patterned after an earlier red flag system developed by a large national insurance company and tested in a major metropolitan area. The results showed that expert

assessment is accurate and can be strongly relied upon within the domain of insurance fraud detection. The system targeted 51% of the claims entered for further investigation. Of the 51% targeted, a representative sample was taken, and 34% of this sample was subsequently (and conservatively) denied. The overall fraud rate for auto theft claims equaled 17% (i.e., 51% times 34%) which strongly suggests that very few fraudulent claims were missed.

In general, research in artificial intelligence^{xi} and biostatistics^{xii} has repeatedly demonstrated that expert assessment is a reliable way to construct a model in the absence of hard data.

Desirability of tuning model through data mining

SAFESTONLINE currently computes a numeric score for each claim processed, and claims above a pre-defined threshold are forwarded to the SIU for further analysis and investigation. It would be desirable to fine tune the model using data mining techniques and to determine to an even finer degree the probability of a claim being fraudulent based on its numeric score.

Knowing the probability of a claim being fraudulent enhances management's ability to select the optimal investigation thresholds for each insurance line given the constraint of limited SIU resources. Further, the thresholds need to be adjusted according to claim reserve as the likelihood of fraud is to some degree dependent on this amount for many lines of insurance. With a probability model, management has a data driven and rational guideline for setting investigative thresholds and related criteria.

The steps to accomplishing these predictions are 1) collecting the outcome of each claim, i.e., the payout amount, the investigation cost and investigation outcome, 2) consistent use of SAFESTONLINE to insure a sufficiently large and representative data set and 3) building the proper models.

The first two steps are mainly dependent on the commitment of management to allocate sufficient resources to insure proper data collection. The last step requires that the right variables are collected and sufficient number of claims are available. To insure that the right predictive variables are measured requires an iterative approach of expert assessment by the SIU and preliminary model building and validation. It is better to start constructing and validating the model early in the data collection process than late.

Part IV. Optimizing return on SIU investment

Once we have a representative sample of claims we can modify SAFESTONLINE's current model to compute an estimate of the probability of an individual claim being fraudulent. The number of claims needed will likely be several thousand per the combination of insurance line and claim reserve amount, where the reserve amount is represented as a range, such as \$5000 to \$10,000, 10,000 to 15,000, 15,000 to 20,000 and so on. Of course, the model would only be predictive and an investigation into the merits of the claim would still be required for any final decision to be made on its validity.

From the estimated probability of a claim being fraudulent we can compute an estimated expected SIU return on investment (ROI) ratio per claim. For example, if for claim A, the expected probability of it being fraudulent is 80 percent, the expected investigative cost is \$800 and the reserve is \$15,000, the expected SIU ratio can be computed as

$$\text{SIU ROI Ratio} = (15000 \cdot .8 + 0 \cdot .2 - 800) / 800 = 11200 / 800 = 112 / 8 = 14 .$$

Thus for every dollar spent by SIU, fourteen dollars would be saved from the denied claim for a net gain of thirteen dollars. In the absence of SAFESTONLINE the SIU ROI ratio is typically seen to be seven or less, as claims are not sufficiently screened before being forwarded to the SIU.

More accurate estimates can be made by taking into account that for some claims the investigation may result in partial payout instead of full or no payout. In addition, if necessary, models can be constructed that can predict the expected investigation cost as well.

Part V. Enhancements

Consideration should also be given to combining multiple information sources such as from RiskShield and NetMap using statistical or other formal objective numerical algorithms to increase accuracy of prediction^{xiii}. In addition, it may be possible to substantially improve the prediction accuracy by jointly combining the Bayes discriminant model with additional *logical rules* (if then type rules). These logical rules should be derived from multi-dimensional pattern recognition algorithms.

There is evidence that the logical rules and the Bayes rules (note, the Bayesian weight of an attribute in the Bayesian Discriminant model is referred to as a Bayes rule) can be somewhat independent and can compliment each otherⁱ. A stronger predictive model may be possible by requiring a claim to pass both the logical rules and the Bayes Discriminant model threshold.

Part VI. Legal issues

SAFESTONLINE can help protect against bad faith, discrimination and invasion of privacy suits. With the passage of the Gramm-Leach-Bliley Act in 1999 it is becoming more necessary to have a uniform and clear reason to instigate an investigation of fraud and gather information that otherwise would be deemed private and strictly off limits. SAFESTONLINE can provide a rational, uniform, non-discriminatory and compelling reason to investigate a claim. Depending on the legal climate SAFESTONLINE's recommendations can be made more or less compelling by raising or lowering the investigation threshold respectively.

A useful metric, in this context, is "lift". Lift measures how well the model is identifying the target group. The higher the lift the better the model is doing. It is defined as

$$\text{Lift} = (\text{target response}) / (\text{population response}) .$$

For SAFESTONLINE the target group would be all those claims recommended for investigation, and the target response would simply be the percentage of claims SIU determined to be fraudulent.

The population response is simply the prevalence of insurance fraud in general. For example, if the general prevalence of fraud is 10 percent and if 50 percent of all claims targeted by SAFESTONLINE were found fraudulent, then the lift would $50/10 = 5$.

The higher the lift the more compelling it is to act upon the model's prediction. Conversely, a model with a low lift such as 1.5, would not be very powerful and would be providing only a weak justification to single out claims for investigation.

Part VII. Summary

With SAFESTONLINE the first round of expert assessment and model building is already complete, yielding dramatic savings. As data is collected the current rules in SAFESTONLINE can be improved and more advanced models, such as discussed here, for further cost savings can be accomplished. While this optimization process can be time consuming and somewhat costly, the alternative of no further optimization is almost certainly more costly in the long run. This cost can be partially mitigated by cross over utilization of existing data mining expertise and technology typically present in the marketing departments of large insurance companies.

General References

Kenneth P. Burham and David R. Anderson (2000). *Model Selection and Inference: a practical information-theoretic approach*. Springer-Verlag, New York.

Dan Hammerstrom (1993). Working with neural networks. *IEEE Spectrum*, July, 46-53.

Jiawei Han and Micheline Kamber (2001). *Data Mining: concepts and techniques*. Academic Press, San Diego, California.

David Heckerman (1997). Bayesian Networks for Data Mining. *Data Mining and Knowledge Discovery* 1, 79-119.

David G. Kleinbaum (1994). *Logistic regression: a self learning text*. Springer-Verlag, New York.

Endnotes

ⁱ F. F. Fenech, Z. Junousov, A. Mazovetsky and V. Olchanski (1995). A Computerized Health Screening and Follow-up System in Diabetes Mellitus. *Diabetic Medicine* 3: 271-276.

ⁱⁱ B. A. Teather, G. Della Riccia, D. and Teather (1995). Learning from noisy medical data: a

comparative study based on a real diagnostic problem. In *Mathematical and Statistical Methods in Artificial Intelligence*, (ed. Riccia, G. Della, Kruse, R. and Viertl, R.), pp. 247-256. Springer-Verlag, New York.

ⁱⁱⁱ Jiawei Han and Micheline Kamber (2001). *Data Mining: concepts and techniques*. Academic Press, San Diego, California.

^{iv} Trevor Dwyer (2002). Neural Networks and Decision Tree Prediction in SQL Server. Abstract in *SQL Server Magazine Live Conference and Expo*, Orlando, Florida.

^v Jae Shim, Joel Siegel, Anique Qureshi and Robert Chi (1999). *Information Systems Management Handbook*. Prentice Hall, Paramus, New Jersey.

^{vi} Bob Stone, Ron Jacobs, H. Robert Wientzen (2001). *Successful Direct Marketing Methods*, Seventh Edition. McGraw-Hill, Chicago, Illinois.

^{vii} Bayes Theorem states that the probability of A given B = (the probability of B given A) times (the probability of A) divided by (the probability of B) or in mathematical notation, $P(A | B) = P(B | A)P(A) / P(B)$. The theorem follows directly from the identity $P(A \wedge B) = P(A | B) P(B) = P(B | A) P(A)$.

^{viii} Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen and David J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York.

^{ix} Dan Hammerstrom (1993). Working with neural networks. *IEEE Spectrum*, July, 46-53.

^x Ibid.

^{xi} Avron Barr and Edward A. Feigenbaum (ed.) (1982). *The Handbook of Artificial Intelligence*, Volume II. Addison Wesley, Reading, Massachusetts.

^{xii} P. Armitage and T. Colton (ed.) (1998). *Encyclopedia of Biostatistics*, Volumes 1-6. John Wiley and Sons, New York.

^{xiii} National Research Council, Committee to Review the Scientific Evidence on Polygraph (2002). Appendix K: Combining Information Sources in Medical Diagnosis and Security Screening in *The Polygraph and Lie Detection* (prepublication manuscript). The National Academies Press, Washington, D.C.