KDD For Insurance Risk Assessment: A Case Study*

Graham J. Williams and Zhexue Huang
CSIRO Division of Information Technology
GPO Box 664 Canberra 2601 Australia
Email: Graham.Williams@cbr.dit.csiro.au

March 1996

Abstract

Insurance is a business of risks. Identifying and understanding areas of risk is an important task performed by an insurer. An assessment of risk is used to set the appropriate premium for insurance policies. This paper describes a KDD exercise which uses decision tree techniques to identify significant areas of risk within an insurance portfolio. The real world dataset used contains information about policies and insurance claims on those policies. Decision trees can be constructed to identify and describe areas of high risk which are then evaluated, as a separate exercise, in terms of claim frequency and claim costs. The paper stresses the idea of interactive post-processing, or evaluation, of the patterns that are illuminated by traditional data mining tools.

Keywords: Knowledge discovery in databases, KDD Process, data mining, insurance risk analysis.

1 Introduction

Assessing the performance of an insurance portfolio requires both overall and within-portfolio analyses. The overall analysis demonstrates whether the portfolio in the past was profitable or not, whereas a detailed within-portfolio analysis reveals which particular areas within the portfolio made significant profits and loses (Coutts 1984). The latter information is particularly important for a company to know in order to maintain both its competitiveness in the market and its profitability. Without knowing where the areas of significant risk are in its portfolio an insurer will be unable to sustain operation, even though the overall portfolio may perform well temporarily. A portfolio should balance its exposure to risk.

The overall analysis of an insurance portfolio can be performed using straight forward statistical techniques, based on the total premiums earned and the total of the claims paid. More sophisticated approaches are required for within-portfolio analysis (Brockman and Wright 1992, Coutts 1984).

A common approach to within-portfolio analysis is to partition the risk of the whole portfolio into small areas of risk. These are identified by a set of risk rating factors usually represented by a contingency table. Only a few variables are usually considered at a time (the claim frequency, the claim cost, and the exposure, for example) for each cell of the contingency table.

When a portfolio is partitioned a model can be fitted to each cell relating the level of the risk rating factors to the claim frequency and claim cost. Historical data is used to estimate parameters of the model. The model can then be used to predict the expected claim frequencies and claim costs for different risk rating factor levels.

With such an approach though the risk rating factors must be categorical. A continuous factor has to be categorised into a small number of levels. A trade-off is needed in the detail of the analysis and fit of the

*The authors acknowledge the support provided by the Cooperative Research Centre for Advanced Computational Systems (ACSys) established under the Australian Government's Cooperative Research Centres Program.

model. Also, the interaction between factors is often ignored because of the difficulty with handling them. With many variables and with categorical variables with many values, the number of possible interactions is very large.

Alternative approaches to the task of insurance risk analysis have primarily been explored within the statistical/actuarial domains. Siebes (1994), however, considered the problem of insurance risk analysis in the context of data mining by employing probability theory. Insurance claims on policies in any one year were viewed in terms of Bernoulli experiments. This work lead to the idea of developing equal probability, homogeneous descriptions of classes of the insurance portfolio.

In this paper we consider some initial exploratory work in using an iterative and interactive approach to insurance risk analysis within the KDD context. The approach is motivated by the desire to involve expert intuition as an integral aspect of the process, supported by formal analyses. The approach is also driven by the view of KDD as an iterative process involving a number of stages, one of which is the traditional Data Mining stage (Fayyad, Piatetsky-Shapiro and Smyth 1996, Williams and Huang 1996b).

While the primary focus of this paper is on the data mining and evaluation (post data mining) stages, the pre data mining stages are also highlighted. Section 2 describes the concept of an insurance portfolio database which serves as the starting point of the KDD exercise. Section 3 summarises the process through which the data supplied by our collaborators was transformed into a Working Dataset for the data mining stage. Section 4 describes the data mining tool used for this exploratory analysis and Section 5 demonstrates the types of results and post data mining evaluations performed. NRMA Insurance Limited, one of Australia's largest general insurers, were the collaborators for the project and supplied the database and domain expertise.

2 Insurance Portfolio Databases

A portfolio database in an insurance company contains a set of insurance policies purchased by customers. A policy insures up to a specified value a particular entity (motor vehicle, household contents, buildings) from loss. When damage or loss occurs to the insured entity a claim is made against the policy for compensation. A policy is valid for a certain time period and is usually renewed upon payment of the premium. The period during which a policy is active is referred to as the period of exposure (as in the insurer's exposure to risk) and is usually measured in days. At any particular time the exposures associated with active policies in the portfolio database will differ depending upon their dates of validity.

A key factor to the success of an insurance portfolio is balancing the trade-off between the setting of competitive premiums and covering the risk associated with the exposure. The insurance market is very competitive and setting premiums too high leads to a loss of market share. Yet setting premiums too low leads to loss of profit. Premiums are generally based on a small number of key factors (such as age of driver, type of vehicle, and address of owner) identified by various analyses and intuitions. Because of the size of many insurance portfolios, analysis is often restricted to straightforward techniques.

The performance of a portfolio is usually assessed in the context of the previous year's data. The analyses performed on the data are used by the underwriter to envisage the near future performance of the portfolio and to adjust the policy rating structure to reflect changes in the market and in the behaviour of the insured. Essentially the analyses are used each year to tune the rules which set premiums for the coming year.

There are two extremes in setting premiums: all policies attract the same premium; or each individual policy has an individual premium determined for it based on all of the details of the policy. Neither extreme is particularly useful nor practical. The former would likely penalise those who are less likely to lodge claims in favour of those who a more likely to do so, and would probably drive them to the competitors! The latter would generally be impractical because of the complexity of the rules that would be required. However, the goal of good insurance premium setting is likely to lean towards the latter than to the former. An insurer would prefer to fine tune their premium setting more so than is often currently performed, identifying more factors that affect the risk and taking more information into account in setting premiums.

With the focus on dealing with very large databases, Data Mining (and KDD in general) provides an environment in which to perform such extended analyses of insurance portfolio databases. The ACSys Data Mining Environment (Williams and Huang 1996a)—consisting of a collection of integrated tools together with a variety of commercial and research data mining systems based on high performance parallel computers—provides the framework in which these analyses are performed.

3 Preprocessing

As observed elsewhere (Brachman and Anand 1996, Williams and Huang 1996b), while data mining is a key stage of the whole KDD process, the other stages of the process often require considerable (and often considerably more) effort. Starting from the data extracted from the source database maintained by NRMA Insurance Limited (consisting of many millions of policies), a number of transformations had to be performed before a suitable Working Dataset was built. This involved many iterations, some of which were performed after the initial data mining runs were performed. (The initial runs indicated different directions to take, different data to obtain, and different transformations to be performed.)

The Source Dataset was extracted from the NRMA's motor vehicle insurance portfolio database. An initial trial dataset of some 125,000 records, covering a time period of only 6 months (1 July 1994 to 31 December 1994), was used prior to working with the full dataset. Each record contained 93 fields of data. Figure 1 summarises the preprocessing performed.

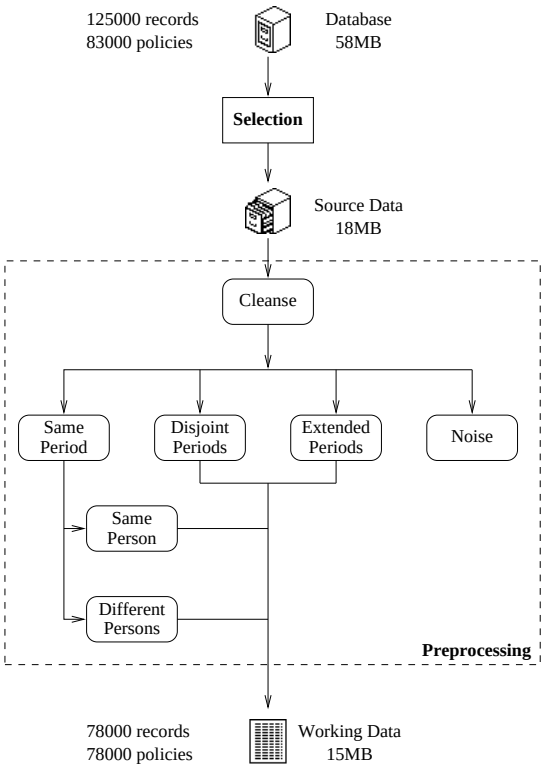

Figure 1: Preprocessing.

The first task was to remove from the database those fields/attributes/variables which were irrelevant to the task at hand. This process was complicated by the fact that some "obviously" irrelevant attributes (e.g., office at which application for insurance was lodged) could contain surprising information. On the other-hand, leaving irrelevant attributes in the data set can lead to aberrant results. A panel of KDD experts (from Machine Learning, Statistics, and Databases) and domain experts (from the NRMA) considered each attribute in turn, removing only those that were seen as clearly irrelevant by all.

The second task, labelled "Cleansing" in Figure 1, involved performing various transformations on the data (e.g., birth dates transformed into ages). Various computed fields were also added to the data at this stage. One, for example, calculated the period of exposure associated with the policy in the context of the period of interest (last six months of 1994).

The third task involved collapsing the transaction-oriented data that was supplied into a policy-oriented dataset which is required for the types of analyses intended to be performed. The original database was oriented towards managing the insurance portfolio—it was not designed with Data Mining in mind. A transaction in the original database records some change made to, or incident associated with, a policy. Hence, in the 125,000 records, there were only about 83,000 policies. Some policies had multiple owners

(hence multiple records) while others had a variety of changes, such as renewals, cancellations, etc. Other policies appeared more than once when there was more than a single claim made on the policy in the period of interest.

A careful study of the dataset had to be made to understand all of its nuances. This finally resulted in the definition of a set of rules to merge the multi-record policies into single policy records. These rules were implemented as a variety of operations and were tuned iteratively as the data and problems became better understood.

An intermediate dataset was created with some 78,000 individual policies, each described by 43 attributes, some of which were generated in the merging process to record, for example, the number of owners of each individual policy and the total amount of all claims made against a policy. Policies that were cancelled or met other criteria (e.g., contained critical but missing data) were also removed from the dataset.

The intermediate dataset was intended to be used for multiple data mining objectives (e.g., fraud detection). Consequently, some of its attributes were still irrelevant to the risk analysis and thus further processing was performed to build the final risk analysis Working Dataset. This dataset contained some 75,000 unique policies, each being represented by 23 attributes, including exposure and total claims cost.

An initial observation that has some impact upon the type of analysis performed was that of the 75,000 policies, only about 4,000 of them actually had claims against them. While not particularly surprising, such a low relative occurrence (about 5%) of an important aspect of the data does require attention (e.g., appropriate use of prior probabilities), but is beyond the immediate scope of this paper.

A second observation of relevance to the data mining is that different policies have different exposures, even though the majority have exposure for the whole period. This indicates that the exposure should be used to weight the examples in some way. This important aspect of the data is also left for future exposition.

4 Decision Trees

The ACSys Data Mining Environment uses StarTree, for example, which is an implementation of the classification and regression tree software CART (Breiman, Friedman, Olshen and Stone 1984). This parallel implementation by Thinking Machines Corporation is part of the Darwin toolkit (Thinking Machines Corporation 1995). CART is similar in many ways to the traditional machine learning decision tree induction algorithm C4.5 (Quinlan 1993).

StarTree consists of three operations: *grow tree*; *evaluate tree*; and *apply tree*. The *grow tree* stage actually implements the usual divide-and-conquer strategy in a parallel architecture for building a full decision tree based on a training set. The *evaluate tree* stage evaluates the generated tree in the context of a test dataset. It is here that pruning is performed and a variety of test datasets can be used to explore the performance of the tree in the context of pruning. The *apply tree* stage simply applies any resulting trees to new unseen data. StarTree provides a variety of output options, including a LaTeX formatted decision tree or a collection of rules.

5 Risk Analysis

5.1 Methodology

The claim cost attribute in the final database was selected as the target attribute (or dependent variable) and the remainder (excluding exposure) as the independent variables. All policies were classified into two classes: those with claims and those with no claims—the claim cost was effectively reclassified as 1 or 0. The dataset was processed with *grow tree* to produce a complete decision tree.

Using StarTree a rule base can be built from a generated decision tree. A rule base consists of a set of rules, each corresponding to a different path through the tree representing a conjunction of the conditions associated with the nodes of the tree. An example rule might be:

If age ≤ 20
and sex $=$ Male
and insured_amount ≥ 5000
and insured_amount < 10000
Then insurance_claimed = 1, cost = 0, (0, 15)

This rule indicates that under the given conditions an insurance policy is claimed against (with a misclassification cost of 0). There are 15 examples from which this rule was generated.

The primary rule base generated by StarTree is now extended within the ACSys Data Mining Environment with new information that is derived from other sources, including the Source Dataset and original database. Significant areas of risk were explored by viewing those leaf nodes containing claims. The frequency of claims at any leaf together with the total of the claim costs at the leaf nodes provide important information. However, since claims cost had been factored out of the training set used by StarTree to build the decision tree, this information was incorporated as an extra step following on from the data mining. The rules are thus extended with information that records the total exposure and the total claim cost for each rule. Other information such as the total of the premiums associated with the "area of risk" could also have been determined.

The generation of this extra information is an important step in the KDD task which, as we emphasise above, is more than just the application of Data Mining tools. This post processing of the learned rules is an important part in the final stage of the KDD process—the evaluation of discovered patterns.

The additional information can be used to determine whether the information embodied as a rule is useful as knowledge. Such an exploration of the rules discovered turns data mining, in a sense, back onto itself. This is particularly the case where we are dealing with very large datasets with many attributes from which very large trees (and hence very many rules) are generated without pruning (and hence over fitting). The generated rules themselves need to be data mined (in the context of this evaluation stage).

5.2 Initial Results

Some initial results from the analysis of the NRMA Insurance Limited sample database are presented here. Due to the commercial nature of the data the actual variables will be referred to as `F01`, `F02`, . . ., `F23`.

For illustration, three decision trees were generated using different selection measures for the partitions—we will refer to the three trees as the entropy, gini, and error trees. (Refer to Mingers (1989) for a discussion of selection measures.) An idea of the size and organisation of the trees produced from the sample data is gained from the statistics of the various trees grown (Table 1).

Tree name:	Entropy	Gini	Error
Decrease Function	entropy	gini	error
Number of Nodes	10215	10887	8197
Depth	43	38	39
Total leaves	5062	5391	3331
Positive leaves (Pl)	1967	2188	1192
Negative leaves (Nl)	3095	3203	2139
Policies by Pl	3815	3795	1682
Policies by Nl	71993	72013	74126

Table 1: Using alternative decrease functions

It is clear from Table 1 that very large trees are generated given the large number of attributes (and possible values) and the large number of examples. (Refer to (Catlett 1991) for a discussion of using decision tree induction on large datasets.) The row in the table giving the total number of leaves corresponds also to the number of rules generated from the tree. Similarly the positive leaves and the negative leaves. The Entropy tree, for example, generates 1967 rules for claims.

Although 21 independent variables were used in the input data, only 19 were selected by the Entropy and Gini trees and 20 by the Error tree. Table 2 lists the selected variables and the frequencies with which the variables appear in each tree. Thus, the variable `F01` appears 111 times in the Error tree. Such information can be of use in gaining some insights into which variables appear to be important. (The location of the

Variable	Entropy	Gini	Error
F02	0	0	0
F01	0	0	111
F14	465	840	60
F21	668	711	52
F19	1756	1367	119
F22	2223	2177	85
F18	2258	2034	374
F12	2667	2589	289
F16	3013	2461	82
F07	3051	2947	1407
F15	3543	3122	189

Variable	Entropy	Gini	Error
F13	4629	4928	224
F20	4843	5175	19
F17	6164	6197	1023
F04	6773	6743	4041
F10	6799	6431	3975
F06	8760	9122	10195
F08	10771	11772	20842
F09	11362	10200	7895
F03	12073	12349	11464
F11	12329	12989	1480

Table 2: Attribute Frequencies

variable within the tree and the type of the variable—categorical versus continuous—are other important indicators.)

Table 3 records the number of claims associated with the rules generated from each of the three trees, aggregated by the number of claims. There are 1090 rules from the Entropy tree, for example, which have just a single claim. On the other hand, there are 36 claims associated with one of the rules from the Gini tree. Those rules associated with a higher number of claims would tend to indicate areas of high risk.

Claims	Number of Rules		
	Entropy	Gini	Error
1	1090	1493	941
2	494	404	150
3	192	135	48
4	82	51	27
5	38	31	7
6	15	16	7
7	18	17	3
8	9	7	3
9	5	6	2
10	5	5	1
11	4	5	1

Claims	Number of Rules		
	Entropy	Gini	Error
12	2	4	
13	4	5	
14	4		2
15	2	1	
16	2	1	
17		2	
18		1	
22	1	2	
24		1	
36		1	

Table 3: Claims associated with each risk area.

Tables 4 summarise the rules from each tree with more than 14 claims against them or with a high average claim cost with respect to the exposure. For each such rule the tables record the number of claims, the total amount of exposure, and the sum of the claim costs. Each row represents a single rule. In the Gini sub-table, for example, there are 36 claims associated with a single rule. These claims cover 442 days of exposure and have a total claim cost of $159,472. Such information can again be used to target areas described by the rules for further explorations.

Armed with such "post data mining" analyses areas of significant insurance risk can be identified and further investigated. Such investigations could be expected to lead to a better understanding of insurance risk and to a finer tuning of insurance premium setting.

6 Summary

In this paper we have followed the KDD process and highlighted a novel approach to its application to the domain of insurance risk assessment. A particular focus has been on the interaction between the data mining stage and the evaluation stage of the KDD process. StarTree was used to analyse the large insurance dataset, generally building very large decision trees (and hence large rule sets). The rules were used only as

Gini

Claims	Exposure	Cost
11	221	139673
15	1868	74367
16	2513	54750
17	2635	33152
17	2637	52721
22	2305	70558
22	3839	86988
24	3399	98918
36	442	159472

Entropy

Claims	Exposure	Cost
15	1745	43308
15	197	55921
16	2198	50213
16	2471	52201
22	265	85678

Error

Claims	Exposure	Cost
10	349	27827
14	2183	49883
14	2023	49889

Table 4: High Claim Risk Areas

an aid in the "discovery" of interesting areas of the data. By understanding such hot spots, better insights into policy premium setting can be gained.

Acknowledgements

This work has been performed within the Divisions of Information Technology (DIT) and Mathematics and Statistics (DMS) of the Australian Governments' Commonwealth Scientific and Industrial Research Organisation (CSIRO). The KDD team also includes Peter Milne of DIT and Murray Cameron, Glenn Stone, Petra Kuhnert, and David Chan of DMS. Industry collaborators from NRMA Insurance Limited include Keith Forster, Philip Woods, and Hong Ooi. Thanks to all for their comments on this paper.

References

Brachman, R. J. and Anand, T.: 1996, The process of knowledge discovery in databases: A human-centered approach, *in* Fayyad, Piatetsky-Shapiro, Smyth and Uthurusamy (1996), chapter 2.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J.: 1984, *Classification and regression trees*, Wadsworth, Belmont, CA.

Brockman, M. J. and Wright, T. S.: 1992, Statistical motor rating: Making effective use of your data, *Journal of Institute of Acturaries* **119**, 457–543.

Catlett, J.: 1991, *Megainduction: Machine learning over very large databases*, PhD thesis, Basser Department of Computer Science, University of Sydney.

Coutts, S. M.: 1984, Motor insurance rating: An acturarial approach, *Journal of Institute of Acturaries* **111**, 87–148.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P.: 1996, From data mining to knowledge discovery: An overview, *in* Fayyad, Piatetsky-Shapiro, Smyth and Uthurusamy (1996), chapter 1.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds): 1996, *Advances in Knowledge Discovery and Data Mining*, AAAI Press.

Mingers, J.: 1989, An empirical comparison of selection measures for decision-tree induction, *Machne Learning* **3**(4), 319–342.

Quinlan, J. R.: 1993, *C4.5: Programs for machine learning*, Morgan Kaufmann.

Siebes, A.: 1994, Homogeneous discoveries contain no surprises: Inferering risk-profiles from large databases, *Technical Report CS-R9430*, CWI.

Thinking Machines Corporation: 1995, The Darwin solution: A family of prediction and classification tools for large databases, *Technical report*, Thinking Machines Corporation, available from http://www.think.com/.

Williams, G. J. and Huang, Z.: 1996a, The ACSys data mining environment, *Technical Report TR-DM-96001*, CSIRO Division of Information Technology.

Williams, G. J. and Huang, Z.: 1996b, Modelling the KDD process: A four stage process and four element model, *Technical Report TR-DM-96013*, CSIRO Division of Information Technology, available from Graham.Williams@dit.csiro.au.