Reduction of Risk Factors in Risk Assessment: An Identification Tree Approach

Ankit Agarwal, Jyotshna Dongardive, and Siby Abraham

***Abstract*—The paper proposes a unique procedure to reduce risk factors in risk assessment. It offers a variant of decision tree called Identification tree for reducing number of risk factors used in assessment. The model, which uses auto insurance as a case study, employs historical evidences of different vehicles as risk factors. The work offers reduction of original risk factors from a set of twenty three to a reduced set of nine risk factors. The model was validated using real time and industry specific data.**

***Index Terms*—Average disorder score, identification tree, risk assessment, risk factor**.

I. Introduction

We come across many uncertainties in our daily lives. Some of these uncertainties might result in financial losses. Insurance, which safeguards these financial losses to a minimum, has many forms like life, home, auto, property etc to name a few [1].

RISK ASSEMENT			
ASSET	THREAT	VULNERABILITY	MITIGATION
What are you trying to protect?	What are you afraid of happening?	How could the threat occur?	What is currently reducing the risk?
Impact/Severity What is the impact to the business? 1. Negligible 2.Minor 3. Moderate 4. Major 5. Critical 6.Catastrophic		Probability/Likelihood How likely is the threat? 1. Unforeseeable 2. Very unlikely 3. Possible 4. Likely 5. Very likely 6.Almost certain	
RISK LOG			
Priority hazard Occurrence of an event	Impact (1-6) Impact of event on	Probability (1-6) Probability of the event	Risk Rating = (Impact*probability) Risk Rating = Impact * Probability

Fig. 1. Risk assessment

Risk assessment in insurance is a practice wherein vulnerability of an asset against a risk is measured [2]. This involves finding likelihood of a financial loss caused by various risk factors. The likelihood and impact of these risk factors are collected into a risk log for assessment. The various stages involved in risk assessment are given in Fig. 1 [2]. Conventionally, it is done using various statistical and computational techniques [3]. Almost all these techniques assume availability of all risk factors.

The paper proposes a methodology to reduce risk factors. This uses an identification tree [4], [5] based approach to reduce all the risk factors to the most significant ones. The work, which uses auto insurance sector as a case study, proposes that risk assessment can be realized using this reduced set of risk factors. The paper is organized in six sections. Section II gives related work. Section III introduces the model proposed. Section IV provides the implementation. Section V gives experimental results and section VI offers conclusion and future work.

II. Related Works

There have been many attempts to discuss risk assessment computationally. Jianbing Xiahou and Yang Mu [6] used Decision Tree as a Data Mining technique to classify and select risk factors. They concluded that Data Mining approach gave better results than the conventionally used statistical technique called General Linear Model [7]. Yong Che *et al*. [8] offered an alternative model for risk assessment using Ubiquitous Computing. It was a three step process containing Clustering on input data using Adaptive Resonance Theory [9], a modification phase of the feature vectors and a Back Propagation Neural Network. Chin-sheng haung *et al*. [10] proposed an evaluation model for selecting insurance policy using Analytic Hierarchy Process [11] and Fuzzy Logic [12]. They used four variables to evaluate purchase of life insurance and annuity insurance including age, annual income, educational level and risk preference. Anna Jurek and Danuta Zakrzewska [13] used a Naive Bayes model along with Clustering technique for risk assessment in life insurance. The work involved classification of artificially generated data sets into three risk classes, which was further enhanced using Clustering. Arnold F. Shapiro [14] gave an overview of Soft Computing applications in Actuarial Science. The work covered important Soft Computing techniques like Neural Networks [15], Fuzzy Logic and Genetic Algorithms [16]. He [17] also provided a comprehensive overview of recent advances in the theory and implementation of intelligent and other computational techniques in insurance. Chin-Sheng-Huang [18] *et al*. used Decision Trees [19] to establish Decision models for five different insurance sectors in Taiwan. Almost all these methods assumed the existence of a set of exhaustive risk factors.

Manuscript received October 25, 2012; revise March 2, 2013.

Ankit Agarwal and Jyotsna Dongerdive are with the University Department of Computer Science, University of Mumbai, India (e-mail: ankit.g.agarwal@gmail.com, jyotss.d@gmail.com).

Siby Abraham is with the Department of Mathematics & Statistics, Guru Nanak Khalsa College, University of Mumbai, India (e-mail: sibyam@gmail.com).

TABLE I: RISK FACTORS (RF)

Risk Factors	Range
Fuel type	Diesel, gas.
Aspiration	std, turbo
No. of doors	four, two
Body-style	Hardtop, wagon, sedan, hatchback, convertible.
Drive wheels	4wd, fwd, rwd.
Engine location	Front, rear.
Wheel base	Continuous from 86.6 to 120.9
Length	Continuous from 141.1 to 208.1
Width	Continuous from 60.3 to 72.3
Height	Continuous from 47.8 to 59.8
Curb-weight	Continuous from 1488 to 4066
Engine type	dohc, dohcv, l, ohc, ohcf, ohcv, rotor
No. of cylinders	eight, five, four, six, three, twelve, two
Engine-size	Continuous from 61 to 326
Fuel-system	1bbl, 2bbl, 4bbl, idi, mfi, mfpi, spdi, spfi
Bore	Continuous from 2.54 to 3.94
Stroke	Continuous from 2.17 to 4.07
Compression ratio	Continuous from 7 to 23
Horsepower	Continuous from 48 to 288
Peak rpm	Continuous from 4150 to 6600
City-mpg	Continuous from 13 to 49
Highway-mpg	Continuous from 16 to 54
Price	Continuous from 5118 to 45400

III. RAIT

The proposed model is called Risk Assessment through Identification Tree (RAIT). It uses a classification technique called identification tree to convert risk factors into a relevant and reduced set.

The tree uses an Average Disorder Score (ADS) given by the following formula.

$$\sum_{b} \left(\frac{nb}{nt}\right) * \left(\sum_{c} -\frac{nbc}{nb} * \log_2 nbc / nb\right)$$

TABLE II: DIFFERENT CLASSES OF CONTINUOUS RF

Risk Factors	Range
Wheel base	(85-90), (90-95) ... (115,120).
Length	(140-145), (145-150) … (205,210).
Width	(60-62), (62-64 ... (70-72).
Height	(47-49), (49-51) ... (57-59).
Curb-weight	(1488-2488), (2488-3488), (3488-4488).
Engine type	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
No. of cylinders	Eight, five, four, six, three, twelve, two.
Engine-size	(61-71), (71-81)… (321-331).
Fuel-system	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi
Bore	(2.54-2.94), (2.94-3.34), (3.34-3.74), (3.74-4.14).
Stroke	(2.07-2.57), (2.57-3.07) ... (4.17-4.57).
Compression-ratio	(7-11), (11-15), (15-19), (19-23).
Horsepower	(48-58), (58-68), (68-78), (78-88)…. (278-288).
Peak rpm	(4150-4650), (4650-5150) ... (6150-6650)
City-mpg	(13-18), (18-23) ... (43-48).
Highway-mpg	(15-20), (20-25) ... (50-55).
Price	(5000-10000), (10000-15000) ... (45000-50000).

where ‘nb’ is the number of samples in branch ‘b’; ‘nt’ is the total number of samples in all branches and ‘nbc’ is the number of samples in branch b of class .

RAIT is constructed by taking car insurance data as an instance of auto insurance sector. The data consist of 23 different Risk Factors (RF) and are given in Table I. Of these, 15 RF have values on a continuous range. They are further partitioned into different classes as shown in Table II. Each of the 23 RF has a risk range between -3 and +3 as is the case with the data source [20].

1) Start
2) b= set of branches of RF;
   Reduced Risk Factors (RRF) = [];
   Ordered Branch (OB) = [];
   Disordered Branch (DB) = [].

3) Calculate ADS for each RF.
4) Choose RF with minimum ADS.
5) Insert RF with minimum ADS in RRF [].
6) For each RF in RRF [] Compute ADS for all b.
7) For each b perform following :
8) If (ADS! = 0)
   Label b as disordered and insert into DB [].
   Else
   Label b as ordered and insert into OB []
   End if.
9) End For.
10) For each b in DB [] perform following:
11) Calculate ADS.
12) Choose RF with minimum ADS.
13) Insert this RF in RRF [].
14) List values in RRF [] and its corresponding DB [] and OB [].
15) End for.
16) Go to 6.
17) End for.
18) Stop.

Fig. 2. Pseudo code for RAIT

The pseudo code used for the construction of RAIT, which uses the data given in Table I and Table II, is shown in Fig 2. It is illustrated in following steps.

A. Calculation of ADS

RAIT calculates the ADS for all the 23 RF mentioned in Table 1.The calculation of ADS of one of the RF namely 'Engine size' is explained in the following lines. This RF is further divided into different classes, each of which is called a branch denoted by 'b'. The different branches of this RF are labelled as b1, b2… b27, which are shown in Fig. 3, where each of the summation corresponds to ADS of branch of RF.

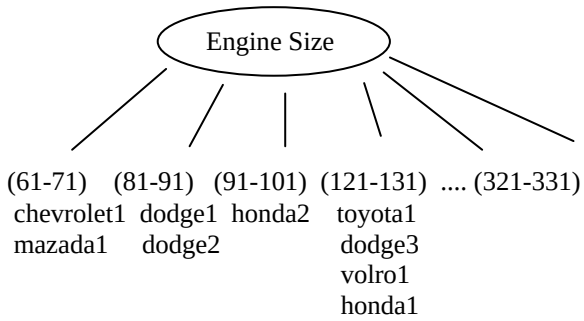

Fig. 3. Classification as per engine size

Hence the ADS of 'Engine size' is given by

ADS=

$$\frac{2}{159}\left[-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right]+\frac{2}{159}\left[-\frac{2}{2}\log_2\frac{2}{2}\right]+$$

$$\frac{1}{159}\left[-\frac{1}{1}\log_2\frac{1}{1}\right]+\frac{4}{159}\left[-\frac{2}{4}\log_2\frac{2}{4}-\frac{1}{4}\log_2\frac{1}{4}-\frac{1}{4}\log_2\frac{1}{4}\right]$$

$$+\ldots.+\frac{0}{159}\left[-\frac{0}{0}\log_2\frac{0}{0}-\frac{0}{0}\log_2\frac{0}{0}\right]$$

'Engine size'. For instance, one of these branches namely b1 (61-71) has parameters as nb = 2, nt = 159, nbc = 1, c = 2. Similar is the case with other branches. Thus ADS for 'Engine size' is returned as 1.4615. Likewise, the ADS for each of the RF is calculated which, are shown in Table III.

B. Selection of RF with Minimum ADS

The RF 'Engine size' which has minimum ADS is selected from Table III. This RF is then placed in RRF [] and hence RRF [] now becomes {Engine Size}. This RF has branches b1, b2... b27. Then, ADS of each of these branches are calculated. The branches of RF are segregated into ordered and disordered. The ordered branches of RF 'Engine size' are b2 (81-91) and b3 (91-101), which are listed in OB []. Hence OB [] becomes {b2, b3}. The disordered branches b1 (61-71) and b4 (121-131) are listed in DB [].

Now RAIT consider the elements of DB [], which consists of {b1, b4}. The ADS of b1 and b4 are now calculated. Since b1 (61-71) has the minimum ADS, its RF with minimum ADS is selected. In this case it is 'hp'. Similarly, the branch b4 (121-131) also has the minimum RF as 'hp'. Hence 'hp' is added to RRF [] and now it becomes {Engine Size, hp}.

Now the RF 'hp' is considered. The branch b2 (91-101) has ADS zero and hence is ordered. The branch b1 (71-81) has non-zero ADS and hence is disordered. The RF with minimum ADS for this branch is 'body style', which has ADS zero. This is shown in Fig 4. The same procedure is continued until all the branches of each element in DB [] becomes a ordered branch. The RAIT terminates by listing the elements of RRF [] with their respective elements of DB [] and OB []. Each iteration returns an element of RRF [] and the same is chosen as root node. Consequently, elements of DB [] become the intermediate nodes and elements of OB [] become leaf nodes. RAIT is executed until the termination condition is met wherein, it returns a sub – tree. The union of each of these sub-trees generates a tree, a snapshot of which is shown in Fig. 5. It was observed that the tree contains only nine distinct types of nodes in the entire tree.

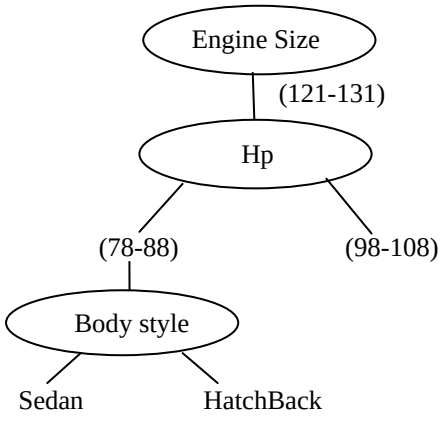

Fig. 4. An identification tree obtained in step 2.

TABLE III: ADS FOR RISK FACTORS

Risk Factors	ADS value
Fuel type	2.2133
Aspiration	2.2121
No. of doors	1.8681
Body-style	1.9095
Drive wheels	2.0899
Engine location	2.2667
Wheel base	1.5525
Length	1.5902
Width	1.7615
Height	1.8096
Curb-weight	2.0150
No. of cylinders	2.1283
Engine-size	**1.4615**
Fuel-system	1.9294
Bore	1.9835
Stroke	2.0927
Compression-ratio	2.1799
Horsepower	1.6061
Peak rpm	1.9823
City-mpg	1.9272
Highway-mpg	1.8602
Price	1.9767

Since each of these nodes correspond to a risk factor, the original problem of risk assessment involving all the 23 risk factors is reduced to a much lesser problem involving only 9 risk factors. Thus the assessment of risk can be focused on this reduced set of 9 risk factors instead of original 23.

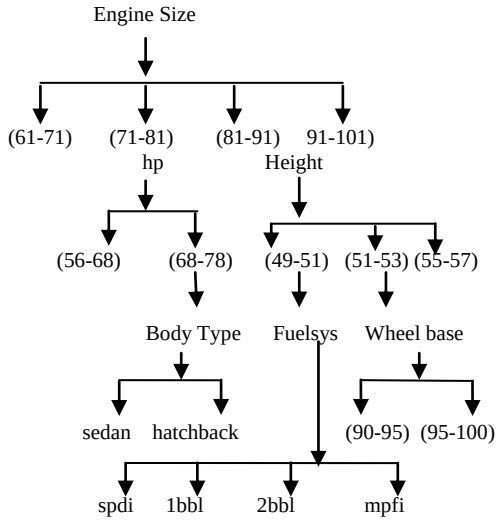

Fig. 5. Snapshot of identification tree evolved.

IV. IMPLEMENTATION

The system is implemented by storing data in Excel sheets. The values were retrieved and processed using VBA functions, a VBA macros utility of MS Excel 2007. The experiments were conducted on a workstation with an Intel Pentium(R) 4 CPU, 2.56GHz, 1GB of RAM, running on Microsoft Windows XP Home Edition, Version 2000.

V. EXPERIMENTAL RESULTS

The RAIT was implemented using 1985 auto import database [20]. The database consists of two types of entities: (a) specification of a car in terms of various characteristics; and (b) its assigned insurance risk rating ranging from -3 to +3. A data set of 199 samples was used. Of which, 80% was used for training and the remaining 20% for testing.

A. Generation of a Sub Tree

Table IV shows the content of RRF [], OB [] and DB [] in the generation of a sub-tree. The procedure was executed until DB [] becomes empty. It demonstrates that the tree was evolved iteratively resulting in the creation of the reduced Risk factors.

TABLE IV: ELEMENTS OF SUB-TREE

Level of sub tree	RRF []	OB []	DB []
1	{**Engine Size**}	{(71-81),(91-101)}	{(61-71), (121-131)}.
2	{Engine size, **hp**}	{(98-108)}	{(78-88)}.
3	{Engine size, hp, **body style**}	{(sedan, hatchback)}	{}.

B. Reduced Risk Factors

Table V presents the Reduced Risk Factors (RRF []) offered by RAIT at each level in the tree. One can conclude that the 'Engine size' is the most significant risk factor in the assessment. The order in which the risk factors are evolved in the tree shows the decreasing order of preference of these risk factors in the assessment.

TABLE V: REDUCED RISK FACTORS (RRF)

Level	Risk Factors
1	Engine size.
2	Hp, Body style.
3	Height, Stroke, Bore.
4	Length
5	Aspiration
6	Number of doors

C. Efficiency of RAIT

The RAIT was tested using 20% of the original data set. The system, which was built on 9 risk factors instead of the 23 risk factors, offered an accuracy of 82.5%.

VI. CONCLUSION AND FUTURE WORK

The paper proposes an identification tree based approach to reduce risk factors in risk assessment. It offers a novel and unique classification procedure using auto insurance sector as a case study. The work provides a substantial reduction of risk factors, which helps in the easy and early assessment of risk from minimal information.

Further enhancements include calculation of risk rate which will quantify the extent of risk. This would eventually lead to the generation of a deterministic model which would offer a collection of antecedent – consequent rules for more accurate and effective assessment.

REFERENCES

[1] General Insurance. [Online]. Available: http://www.irdaindia.org/iac/whatisgeneralinsurance.htm.

[2] D. Cooper, S. Grey, G. Raymond, and P. Walker, *Project Risk Management Guidelines*, New Jersey: John Wiley and Sons, Inc. 2004.
The Risk Management Guide. (2006). [Online]. Available: http://www.ruleworks.co.uk/riskguide/risk-evaluation.htm.

[3] Casualty Actuarial Society. (2008). [Online]. Available: http://www.casact.org/admissions/syllabus/ch3.pdf.

[4] A. Narayanan, E. C. Keedwell, and B. Olsson, "Artificial Intelligence Techniques for Bioinformatics," *Appl Bioinformatics*, vol. 1, no. 4, pp. 191- 222, 2002.

[5] E. Keedwell and A. Narayan, "Intelligent Bioinformatics: the application of artificial intelligence techniques to bioinformatics problem" Wiley, 2005.

[6] J. Xiahou and Y. Mu, "Analysis on application of data mining upon rate making and risk analysis insurance," in *Proc. 2nd International conference Computer Engineering and Technology (ICCET)*, Chengdu, 2010.

[7] D. Anderson, S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, and N. Thandi, *A Practicioner's Guide To Generalized Linear Models*, 2nd Edition, Casualty Actuarial Society - Arlington, Virginia 2004.

[8] Yong Chen *et al*., "Risk Probability Estimation Based on Clustering," in *Proc. Workshop of Information assurance United States military academy*, West point, NY June 2003.

[9] Neuro AI. Intelligent systems and Neural Networks: Neural networks: A requirement for intelligent systems. [Online]. Available: http://www.learnartificialneuralnetworks.com/art.html.

[10] C. S. Haung *et al*., "An Evaluation Model for Determining Insurance Policy Using AHP and Fuzzy Logic, Case Studies of Life and Annuity Insurances," in *Proc. 8th WSEAS International*, Vancouver, Canada, 2010.

[11] ISAHP. (2011). The International Symposium on the Analytical Hierarchy Process. [Online]. Available: http://www.isahp.org/italy2010/index.php

[12] T. J. Ross, *Fuzzy Logic with Engineering Applications*, McGraw Hill, May 1995.

[13] A. Jurek and D. Zakrzewska, "Improving Naïve Bayes models of insurance risk by unsupervised classification," in *Proceedings of the International Multiconference on Computer Science and Information Technology, IEEE,* IMCSIT 2008, Wisla, Poland, pp. 137-144, 2008.

[14] A. F. Shapiro. Soft Computing Applications In Actuarial Science. [Online]. Available: www.soa.org/library/research/actuarial...house/.../arch01v113.pdf.

[15] S. Haykin, *Neural Network a- comprehensive Foundation*, Pearson Education, July 1998.

[16] D. E. Goldberg, Genetic Algorithms in search, Optimization & Machine Learning," Addsion Wesley, 1989.

[17] A. F. Shapir and L. C. Jain, *Intelligent and other computational techniques in Insurance Theory and applications*, December 2003.

[18] C. S. Haung *et al*., "Implementation of Classifiers for Choosing Insurance Policy Using Decision Trees: A Case Study," in *Proc. WSEAS transactions on computers*, Stevens Point, Wisconsin, USA.

[19] Decision Tree Entropy Calculation. [Online]. Available: http://decisiontrees.net/?q=node/26.

[20] Auto MPG Data Set. UCI Machine Learning Repository, University of California, Irvine. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Automobile

Ankit Agarwal received the B.Sc. degree in Computer Science and the M.Sc. degree in Computer Science; both from the University of Mumbai, India. His research interests include artificial intelligence, computational biology. Presently he is working as a software developer in a reputed firm in Mumbai, India.

Dongardive Jyotshna is an active researcher in the area of Bioinformatics. She has experience of teaching bioinformatics at Masters Level for more than six years. She did her Masters in Computer Science and Currently Pursing PhD in Computer Science. Presently she is working as an Assistant Professor in University Department of Computer Science, University of Mumbai, India.

Siby Abraham has a multidisciplinary research background with special interests in Machine Intelligence. He has experience in successfully applying different techniques of Machine Intelligence in Mathematical, Biological, Social, Health and Computational Sciences. He did his Masters in Mathematics and Ph.D. in Computer Science. Presently he is working as an Associate Professor and Head of the Department of Mathematics and Statistics at Guru Nanak Khalsa College, Mumbai, India. He is also a visiting faculty at the Department of Computer Science, University of Mumbai, India.