Conceptual Mapping of Insurance Risk Management to Data Mining

Dilbag Singh
Deptt. of Comp. Science and Applications
Ch. Devi Lal University, Sirsa (Hr)-India

Pradeep Kumar
Deptt. of Comp. Science and Applications
Ch. Devi Lal University, Sirsa (Hr)-India

ABSTRACT

Insurance industry contributes largely to the economy therefore risk management in this industry is very much necessary. In the insurance parlance, the risk management is a tool identifying business opportunities to design and modify the insurance products. Risk can have severe impact in case not managed properly and timely. The mapping of risk management with data mining will help organizations to analyse risks and formulate risk mitigation and prevention techniques more efficiently and effectively. This paper aims to study the conceptual mapping of various task of insurance risk management to data mining. A new paradigm has been suggested for insurance risk management using the main attributes and key aspects of data mining.

Keywords

Insurance, Risk Management, Data mining

1. INTRODUCTION

Risk management (RM) is increasingly recognized as being concerned with both positive and negative aspects of risk. Risk management is a central part of any organization's strategic management. In the safety field, it is generally recognized that consequences are only negative and therefore the management of safety risk is focused on prevention and mitigation of harm. In the insurance parlance, the Risk Management is a tool identifying business opportunities. [1]

RM is a practice of systematically deciding cost effective approaches for minimizing the outcome of threat realization to any organization. RM is an attempt to minimize the chances of failure caused by unplanned events. [1] The aim of risk management is not to avoid getting into things that have risks but rather to minimize the impact of risks in the business task that are undertaken. Risk management consists of the processes, methodologies and tools that are used to deal with risks in insurance.

It is the process whereby organizations methodically address the risks attaching to their activities with the goal of achieving sustained benefit within each activity and across the portfolio of all activities. In Indian Industries, there are extensive uses of techniques like HAZOP, LOPA for risk assessment and logistic risk management with the use of sophisticated instruments like data logger. [2]

The insurance management can disperse the guarantee risk under massive the guarantee risk premise, each kind of risk data and the loss data are mathematical foundation of insurance management.[3] To a certain extent, the risk data and the loss data are insurance resources of insurance management. Many insurance companies in our country have not established the effective risk information system, which causes the decisions of insurance management to lack the reasonable basis, which is also one substantial clause of the low level of risk management in Indian insurance business. [1]

The data mining technology is the important mean that provides the scientific policy-making basis for insurance risk management. [3] Using data mining technology can filtrate and classify customer resources of insurance, divide credit customers into several grades, to predict the customer risk, thus investigating customer material of the low forecasted degrees of comparison can avoid deceiving policy effectively, and avoid service risk. Then can help the insurance company to play the role in the product fixed price in view of the different risk rank [3]. There are many influencing factors to the customer credits, such as the customer income level, the proportion of insurance premium to the income, the education level, area of residence, the credit history, the age, the occupation, and so on. The credit rating in fact is one method which can divide a collectively into different groups according to the different characteristic, that is, is one kind of methods to classify data and predict data [3].

This research paper aims to refine the boundaries of insurance risk management at the conceptual level by mapping the steps involved in performing risk management to the tools and techniques of data mining.

2. RISK MANAGEMENT IN INDIAN INSURANCE INDUSTRY

Insurance industry is keen in identifying the risks pertaining to their business. The property, interruption and liability related risks are keenly looked at by the insurers [5]. The risk management services like underwriting inspections or post loss inspections for settlement claims are used by them in India [12]. The support of public sector insurance companies sponsored organizations like Loss Prevention Association of India (LPA), Tariff Advisory Committee (TAC) was availed by the insurers in addition to the in-house services of individual companies.

In addition to their utility for insurance underwriting business risk management consultancy services are considered as the value addition to the clients of insurance companies. The advent of multinational companies and realization of importance of RM services by the clients have forced insurance companies to bring the innovative services of RM consultancy to the Indian markets [12]. Risk management suggests providing a framework for an organization that enables future activity to take place in a consistent and controlled manner

-Improving decision making, planning and prioritization by comprehensive and structured understanding of business activity, volatility and project opportunity/threat

-Contributing to more efficient use/allocation of capital and resources within the organization.

-Reducing volatility in the non-essential areas of the business

-Protecting and enhancing assets and company image

-Developing and supporting people and the organization's knowledge base

-Optimizing operational efficiency

Risk management is the name given to a logical and systematic method of identifying, analyzing, treating and monitoring the risk involved in any activity or process [13]. Risk management is a methodology that helps managers to make best use of their available resources [13]. Risk management is intended to minimize the chance of unexpected events, or more specifically to keep all possible outcomes under tight management control.[1] Risk management is also concerned with making judgments about how risk events are to be treated, valued, compared and combined. Risk management helps an organization to target the creation and sustainability of competitive advantage by analyzing, mitigating and preventing risk as shown in fig 1[13]

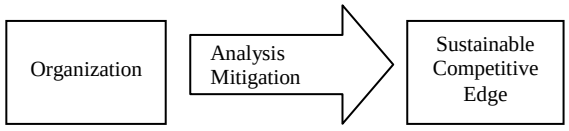

Figure-1: Risk Management in Organization

A formal risk management process is a continuous process for systematically addressing risk throughout the business process. Risks can be introduced at the very earliest stage of the business. The ability to identify risk earlier translates into earlier risk removal, at less cost, which promotes higher business success probability. The risk management process can be described as hierarchical nature of the business. [1,2]

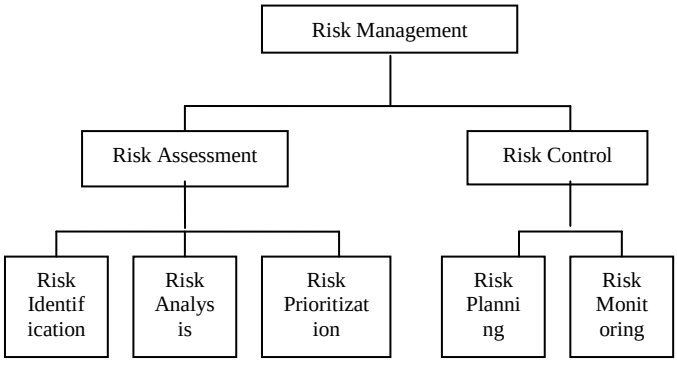

Figure-2: Risk Management Process

The most important part of the Risk Management process is the Risk Assessment. Risk Assessment – a process of evaluation including the identification of the attendant uncertainties, of the likelihood and severity of an adverse effect(s)/event(s) occurring to man or the environment following exposure under defined conditions to a risk source(s). In general the phases or steps of risk management process consist of the following five phases:

2.1 Risk Identification

Risk identification determines what might happen that could affect the objectives of the business and how those things could happen [13]. The output of risk identification is a comprehensive list of possible risk to the successful outcome of the business. Risk identification is a description of the perceived risk. The goal here is to understand the risks that could potentially influence the business.

2.2 Risk Analysis

A risk analysis is to know about frequency of occurrence and severity of impact. It can be a useful visual guide to address the risk exposure. One can achieve the same by multiplying frequency value to the impact value. This phase simply conveys how likely the risk event is likely to happen. It is the analysis of the probability and/or frequency of occurrence of risk event [13]. It also tells you what would be the impact of that event occurring at various stages.

2.3 Risk Prioritization

On the basis of risk analysis, the things can be prioritize by risk category and rated by likelihood and possible cost or consequence [13]. Risk prioritization is a critical characteristic of the formal risk management process, as it provides the opportunity to apply what are typically limited resources to those risks having the largest potential impact on business. The objective here is to short through a large number of risks and determine which are most important and to separate the risk which should be dealt with first while allocating resources. Normally high probability-high severity tasks are given the highest priorities. While low probability-low severities are kept at low priority.

2.4 Risk Planning

After prioritization it becomes clear which risk should be handled. To manage the risks, proper planning is essential. The main task is to identify the action needed to minimize the risk consequences, generally called risk mitigation steps. At this stage we develop and implement a plan with specific counter measures to address the identified risk [13].

2.5 Risk Monitoring

There are a number of measurements and tools that can be used to monitor and report status of risk. As risks are probabilistic events, frequently dependent on external factor, the threat due to risk may change with time as factor changes. Then the risk perception may also change with time. Risks in business should be treated as static and must be reevaluated periodically [13].

The various tools with respect to the phases and their respective application area are listed below.

Table-1: Risk Tools and their Applications

S. No.	Tools/Phases	Application Areas
1	Risk Identification	Stock investment, Insurance product designing, Medical analysis, Power restoration, Postal service etc.
2	Risk Analysis	Market segmentation, Insurance comparisons
3	Risk Prioritization	Classification to quantify risk in purchasing insurance policies/stocks as low/high etc.
4	Risk Management and planning	Insurance analysis, Stock market analysis, Yield projections, Process and quality control.
5	Risk Monitoring	Fraud management in insurance,

3. DATA MINING

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. [7] These tools can include statistical models, mathematical algorithm and machine learning methods. Consequently, data mining consists of more than collection and managing data, it also includes analysis and prediction. In general data mining task can be classified into two categories: Descriptive mining and predictive mining. [3] Descriptive mining is the process of drawing the essential characteristics or general properties of the data in the database. Clustering, association rules are the example of descriptive mining. Predictive mining is the process of inferring patterns from data to make predictions. Predictive mining techniques involves task like classification, time series analysis.

3.1 Cluster Analysis

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering .A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the object in other clusters. A cluster of data objects can be treated collectively as one group in many applications.[4] Cluster analysis tools based on *k*-means, *k*-medians, and several other methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS, and SAS. In machine learning, clustering is an example of unsupervised learning.

In clustering, *k*-means clustering [4] is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data as well as in the iterative refinement approach employed by both algorithms.

3.2 Classifications

Classification is a data mining (machine learning) technique used to predict group membership for data instances. [11] It is a supervised learning technique. Supervised machine learning is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances. In other words, the goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown. [4, 9]

Decision Tree/C4.5 [14] is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

Support vector machine (SVM): A support vector machine (SVM) [14] is a concept in computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes the input is a member of, which makes the SVM a non-probabilistic binary linear classifier.

k-nearest neighbor algorithm **:** In pattern recognition, the *k*-nearest neighbor algorithm (*k*-NN) [14] is a method for classifying objects based on closest training examples in the feature space. *k*-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The *k*-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms.

3.3 Association Rule

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time [10].

Apriori [14] is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps (DNA sequencing).As is common in association rule mining, given a set of *itemsets* (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number C of the itemsets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

3.4 Time Series Analysis

A time series is a collection of observations made sequentially in time. Examples are daily mortality counts, particulate air pollution measurements, and temperature data. The public health question is whether daily mortality is associated with particle levels, controlling for temperature. Time series may be viewed as finding patterns in the data and predicting future values. With time series analysis, the value of an attribute is examined as it varies over times. The values are usually obtained as evenly spaced time points (daily, weekly, hourly etc.) [8]

3.5 Multidimensional – On Line Analytical Processing (M-OLAP)

OLAP is a category of s/w technology that enables analyst/manager/executives to gain insight into data through fast, consistent, interactive access in a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user. In MOLAP Data are stored in specialized multidimensional databases (MDDBs). MDDBs store data in the form of multidimensional hypercube. You have to run special extraction and aggregation job to create cube in the MDDBs from data warehouse. [11]

The core data mining/OLAP techniques and application are as follows:

Table-2: Mining Tools & their Applications

S. No.	Tools/Techniques	Application area
1	Association	Market Basket Analysis, Intrusion Detection, Postal Service, Medical Analysis, Power Restoration etc.
2	Clustering	Market segmentation, Gene expression Analysis,
3	Classification	Classification of customers/ stocks, fraud detection etc.
4	Time Series Analysis	Sale forecasting, Stock market analysis, process and quality control etc.
5	MOLAP	Business multi-dimensional analysis

4. CONCEPTUAL MAPPING

To indicate the relationship between risk management and data mining, the conceptual mapping of two is desired. This is developed to unearth the inherent mapping of the concepts and techniques of the two fields. This can help in enhancing meaningful learning and provide a basis for defining a new paradigm by relating the relationship and idea of risk management and data mining.

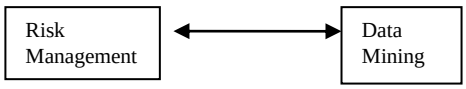

Figure-3: Mapping concepts of Risk Management and data mining

4.1 Risk Identification vs Association

In risk identification, the various tasks that are used to identify whether a set of circumstances represents a risk to the project can be represented in the form of If-"condition" Then "action" For example If customer is asking for high sum assured then proper verification if required.

The association technique of data mining involves discovery of relationship between frequently occurring item sets. Association rule mining is a popular data mining technique because of its wide application in marketing and retail. Association rule mining is a method of finding relationship of the form X-Y amongst item set that occurs together in a database where X and Y are disjoint item set. Finding association rule is defined as the data mining task of finding frequently co-occurring items in a large transactional database. Essentially, association rules discover associations between two set of items/products, such that the presence of one set in a particular traction implies the presence of another in the same transaction.

At the conceptual level, Risk identification and association rule mining map onto each other as both are aimed at developing and understanding relations between different organizational parameter.

Table-3: Risk Identification Vs Association

Risk Identification	Association
Risk Factor A Effect business objective B If factor A then B	Item A Item B If A then B

4.2 Risk Analysis Vs Clustering

In risk analysis, risks are partitioned into groups or clusters such as technical, cost, schedule, management etc. Some risks may fall into multiple categories. Risk needs to be partitioned as some as some risks are more important than the others. Also different stake holders may be concerned about different risks, or different personal may bear responsibility for tracking/monitoring different risk.

Clustering is a technique of dividing data into groups of similar. Each group, called cluster, consists of objects that are similar between them and dissimilar to objects of other groups. Clustering is aimed at grouping of items by maximizing the intra-class similarity and minimizing the interclass similarity.

At the conceptual level, risk analysis task can be solved by using clustering. In risk analysis, there is need of dividing the things into meaningful clusters.

Table-4: Risk Analysis Vs Clustering

Risk Analysis	Clustering
frequency of occurrence x severity of impact The outcome is partitioned into group	Attribute A, B, C, D ….. Grouped into cluster C1, C2 C3

4.3 Risk Prioritization Vs Classification

Risk prioritization provides the opportunity to apply what are typically limited business resources to those risks having the largest potential impact on the business. Risks are ranked and prioritized based on some combination of probability and impact. This can be done qualitatively in a risk prioritization matrix or quantitatively using some type of composite probability-impact score. Based on this score the risk may be classified as tolerable, low, medium, high, or intolerable. In data mining, classification refer to the prediction of the category of categorical data by building a model based on some predictor variable(s)

A classification technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. The model generated by a learning algorithm should both fit the input data well and correctly predict the class label of the record it has never seen before. All approaches to classification however assume some knowledge of the data. Often a training set is used to develop the specific parameter required by the techniques.

Hence, risk prioritization can be viewed as a classification problem as shown below

Table-5: Risk Prioritization Vs Classification

Risk Prioritization	Classification
Category X:	Attribute X:
Yes → High Risk	Yes → Class B
No →category Y:	No →Attribute Y:
Yes → Medium Risk	Yes → Class A
No → Low risk	No → Class B

4.4 Risk Planning Vs Time Series Analysis

Risk planning provide the basis for the identification of the monitoring procedures that should be put in place for each risk, including how to tell if a risk is going to manifest as a real problem , and how frequently each identified risk should be monitored. Risk planning also takes into account risk aversion planning and contingency planning.

The time series analysis in data mining involves forecasting of future values of a time series based on past values. There are three basic functions performed in time series analysis. In one case, distance measures are used to determine the similarity between different time series. In second case, the structure of the line is examined to determine and perhaps classify its behavior. A third application would be to use the historical time series plot to predict future values. It models the relationship between one or more independent or predictor variables and a dependent or response variable. In the context of data mining, the predictor variables are the attributes of interest describing the tuple which make up the attribute vector. In general, the values of the predictor variables are known. Techniques also exist for handling cases where such values may be missing. The response variable is what we want to predict. It is what is referring to as the predicted attribute. Given a tuple described by predictor variable, the goal is to predict the associated value of the response variable.

Risk planning are educated assumptions about future trends and events based on the historical available data and are widely affected by technological innovation, cultural changes, new product, improved services, strong competitors, and shifts in the government priorities, changing social values, unstable economic conditions, and unforeseen events. The mapping of risk planning to time series is shown below.

Table-6: Risk Planning Vs Time Series Analysis [13]

Risk Planning	Time Series Analysis
Risk Var X	Response Var X
Contingency Var Y	Predictor Var Y
Coefficients α,β	Coefficients α,β
$Y=\alpha+\beta X$	$Y=\alpha+\beta X$

4.5 Risk Monitoring Vs MOLAP

There are a number of measurements and tools that can be used to monitor and report status of risk. As risks are probabilistic events, frequently dependent on external factor, the threat due to risk may change with time as factor changes. Then the risk perception may also change with time. Risks in business should be treated as static and must be reevaluated periodically.

In business intelligence, OLAP and dimensional analysis are such facilities that can be used to monitor the risk in insurance business. The critical question in rate making is the following: "What are the risk factors or variables that are important for predicting the likelihood of claims and the size of a claim?" Although many risk factors that affect rates are obvious, subtle and non-intuitive relationships can exist among variables that are difficult, if not impossible, to identify without applying more sophisticated analyses. Modern data mining models can more accurately predict risk, therefore insurance companies can set rates more accurately, which in turn results in lower costs and greater profits.

Table-7: Risk Monitoring Vs MOLAP

Risk Monitoring	MOLAP
Factor A, B	Dimension A, B, C
Outcome X, Y	Attributes of A, B , C
Report R1, R2	View V1, V2

5. CONCLUSION

The conceptual mapping of the insurance risk management to the concepts and techniques used by data mining has been discussed. Firstly, the tools/tasks of risk management are conceptually explained. On the other side various techniques of data mining like association, clustering, classification, time series analysis and MOLAP are conceptually explained. Secondly, the mapping of the insurance risk management to the concepts and tools used by data mining was done. Conceptual mapping of risk identification with association has been shown in table-3. Conceptual mapping of risk analysis with clustering has

been shown in table-4. Conceptual mapping of risk prioritization with classification has been shown in table-5. Conceptual mapping of risk planning with time series analysis has been shown in table-6. Conceptual mapping of risk monitoring with MOLAP has been shown in table-7.It has been shown that there is 1:1 relationship between these two. It means the risks can be managed with the help of various techniques of data mining.

6. REFERENCES

[1] Christopher L Culp 2001. The Risk Management Process- Business Strategy and Tactics. John Wiley & Sons, Inc., New York.

[2] Chen Gang 2009. Mathematics and applications of Risk Management in E-commerce. ISECS International Colloquium on computing, communication, Control, and Management

[3] Yu Yan, Haiying Xie 2009. Research on the Application of Data Mining Technology in Insurance Informatization' In Proceedings of the International conference on Hybrid Intelligent Systems(HIS), PP. 202-205 © IEEE

[4] Jianxin Bi, 2010. Research for Customer Segmentation of Medical Insurance Based on K-means and C&R Tree Algorithms" In Proceedings of the International conference on semantics knowledge and grids (SKG),PP. 359-362 © IEEE

[5] E.B. Belhadji, G. Dionne, and F. Tarkhani 2000. A model for the detection of insurance fraud. Geneva Papers on Risk and Insurance-Issues and Practice, vol. 25, pp. 517-539.

[6] M. Artis, A. Mercedes, and M. Guillen 2002 Detection of automobile insurance fraud with discrete choice models and misclassified claims. The Journal of Risk and Insurance, vol. 69, no. 3, pp. 325-340

[7] Williams, Graham J. and Simoff, Simeon J. 2006. Data Mining Theory, Methodology, Techniques, and Applications. Lecture Note in Computer Science/ Lecture Note in Artificial Intelligence, Springer

[8] Yanghu. 2004. Linear Mining Algorithms Design for Outliers in Financial Time Series and its Authentic Proofs. Chinese Journal of Management Science,12(6):7-11

[9] Hand D. J., Mannil H. & Smyth P. 2001. Principles of Data Mining Cambridge MA: MIT Press,

[10] Jeffrey W. Seifert 2004. Data Mining – An Overview" in proceedings of CRS report for congress

[11] Gupta G.K 2008. Introduction to data mining with case studies", PHI private ltd.

[12] Insurance Regulatory and Development Authority (IRDA) India, Annual Report 2009-10.

[13] Johnson Terence 2010. Conceptual Mapping of Risk Management to Data Mining. In Proceedings of the ICETET, PP. 636-641, © IEEE

[14] XindongWu , Vipin Kumar and others 2008. Top 10 algorithms in data mining. Knowledge and Information System, vol- 14 pp 1–37