In the name of God

The most gracious the most merciful

E.C.O. College of Insurance

Allameh Tabatabaei University

Data Mining Rules and Classification Methods in Insurance: The Case of Collision Insurance

Submitted in partial fulfillment of the requirements

for the degree of

Master of Science

In the subject

Actuarial Science

By:

Zahra Orooji

Supervisor: Mrs. Atousa Goodarzi (Ph.D)

Advisor: Mr.Mohammad Reza Salehi Rad (Ph.D)

Tehran- Iran

March 2010

To my Family and Spouse

Abstract

In the present world, information explosion has been observed in the growth of databases. Volumes of data in databases increase year after year. Such volumes of data are not easy to interpret and also overwhelm the traditional manual methods of data analysis. Data Mining newly considered as the most efficient method to analysis data set and extracting many hidden information beyond the Dataset. In insurance, gathering proper information from aggregate dataset to confirm the best premium for the insurer, such a way, both the insurer and insured feel satisfied of the contract is one of the major goals of insurance company. In this project by using Data Mining methods and machine learning algorithms, running an organized process on aggregate dataset has been followed. As result, by setting proper algorithms, in the first step we can define risky insurer, and then identified the effective features in the value of risk loss and in the final step we are going to estimate the insurance loss. Finding insurance loss can be definitely used by insurance company to estimate the proper premium in each contract.

Keywords: Data Mining, Aggregate Dataset, machine learning algorithms, insurance loss

Acknowledgments

To God the father of all, we thank for the strength that keep me standing and for the hope that keep me believing that this affiliation would be possible and more interesting.

I am heartily thankful to my supervisor, professor Atousa Goodarzi, whose encouragement , guidance and support from the initial to the final level enabled me to develop an understanding of the subject. I would also like to thank my advisor Mr. Salehi Rad , which with his insightful comments guides me through this work.

A very special thanks goes to Mr. Ofoghi, which kindly reviewed this work. I want to express my gratitude to all the people who have given their heart whelming full support in making this compilation a magnificent experience.

I also wanted to thank my lovely family who inspired, encouraged and fully supported me for every trial that comes my way, in giving me not just financial, but morally and spiritually support.

Lastly, I want to have special thanks to my spouse who supported me in any respect of my life and especially during the completion of the project.

Table of Contents

Chapter 1

Introduction

1- Introduction

1-1 Data Mining

Data mining is the process of finding the correlations, patterns and trends by going through large amounts of data, using statistical and mathematical techniques. In the other words Data mining is the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner, data mining is an interconnected field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases.

Data mining can be defined as the process of selecting, exploring, and modeling large amounts of data to uncover previously unknown patterns. Most data-mining problems and corresponding solutions have roots in classical data analysis. Data mining has its origins in various fields, of which the two most important are *statistics* and *machine learning*.

Another point in data mining applications is in the relative emphasis they give to *models* and *algorithms*. Modern statistics is almost entirely driven by the notion of a model. This is an approximation to a structure, which could have led to the data. In place of the statistical emphasis on models, machine learning tends to emphasize algorithms. This is hardly surprising; the very word "learning" contains the notion of a process, an implicit algorithm.

In the other word, data mining is a process of discovering various models, summaries, and derived values from a given collection of data. The word "process" is very important here. Even in some professional environments there is a belief that data

mining simply consists of picking and applying a computer-based tool to match the presented problem and automatically obtaining a solution. There are several reasons why this is incorrect. One reason is that data mining is not simply a collection of isolated tools, each completely different from the other, and waiting to be matched to the problem.

A second reason lies in the notion of matching a problem to a technique. Only very rarely is a research question stated sufficiently precisely that a single and simple application of the method will suffice. In fact, what happens in practice is that data mining becomes an iterative process. One studies the data, examines it using some analytic technique, decides to look at it another way, perhaps modifying it, and then goes back to the beginning and applies another data-analysis tool, reaching either better or different results. This can go round and round many times; each technique is used to probe slightly different aspects of data to ask a slightly different question of the data. What is essentially being described here is a voyage of discovery that makes modern data mining exciting. Still, data mining is not a random application of statistical, machine learning, and other methods and tools. It is not a random walk through the space of analytic techniques but a carefully planned and considered process of deciding what will be most useful, promising, and revealing.[2]

1-2 Data Mining Steps

Estimating and finding the dependencies from data or discovering totally new data is only one part of the data analysis procedure used by scientists. The general process adapted to data-mining problems involves the following steps:

1-2-1 Clarifying the Problem

In most data base, specific knowledge and experience are usually necessary in order to come up with a meaningful problem statement. Unfortunately, many application studies tend to focus on the data-mining technique at the expense of a clear problem statement. In this step, a modeler usually specifies a set of variables for the unknown dependency and, if possible, a general form of this dependency as an initial hypothesis. There may be several hypotheses formulated for a single problem at this stage. The first step requires the combined expertise of an application domain and a data-mining model. In practice, it usually means a close interaction between the data-mining expert and the application expert. In successful data-mining applications, this cooperation does not stop in the initial phase; it continues during the entire data-mining process.

1-2-2 Data Collection

How the data are generated and collected have been investigated in this step. There are two distinct possibilities. The first is when the data-generation process is under the control of an expert (modeler): this approach is known as a *designed experiment*. The second possibility is when the expert cannot influence the data- generation process: this is known as the *observational approach*. An observational setting, namely, random data

generation, is assumed in most data-mining applications. Typically, the sampling distribution is completely unknown after data are collected, or it is partially and implicitly given in the data-collection procedure. It is very important, however, to understand how data collection affects its theoretical distribution, since such a priori knowledge can be very useful for modeling and, later, for the final interpretation of results. Also, it is important to make sure that the data used for estimating a model come from the same sampling distribution. If this is not, the estimated model cannot be successfully used in a final application.

1-2-3 Preparing the Data

Much of the raw data contained in databases is unpreprocessed, incomplete, and noisy.

For example, the databases may contain:

- Fields that are obsolete or redundant
- Missing values
- Outliers
- Data in a form not suitable for data mining models
- Values not consistent with policy or common sense.

To be useful for data mining purposes, the databases need to undergo preprocessing, in the form of data cleaning and data transformation. Data mining often deals with data that hasn't been looked at for years, so that much of the data contains field values that have expired, are no longer relevant, or are simply missing.

In the observational setting, data are usually "collected" from the existing databases, data warehouses, and data marts. Data preprocessing usually includes at least two common tasks:

1. *Outlier detection (and removal)* – Outliers are unusual data values that are not consistent with most observations. Commonly, outliers result from measurement errors, coding and recording errors, and, sometimes, are natural, abnormal values. Such nonrepresentative samples can seriously affect the model produced later. There are two strategies for dealing with outliers:
    a. Detect and eventually remove outliers as a part of the preprocessing phase, or
    b. Develop robust modeling methods that are insensitive to outliers.

2. *Scaling, encoding, and selecting features* – Data preprocessing includes several steps such as variable scaling and different types of encoding. For example, one feature with the range [0, 1] and the other with the range [−100, 1000] will not have the same weights in the applied technique; they will also influence the final data-mining results differently. Therefore, it is recommended to scale them and bring both features to the same weight for further analysis. Also, application-specific encoding methods usually achieve dimensionality reduction by providing a smaller number of informative features for subsequent data modeling.

3. These two classes of preprocessing tasks are only illustrative examples of a large spectrum of preprocessing activities in a data-mining process.

4. Data-preprocessing steps should not be considered completely independent from other data-mining phases. In every iteration of the data-mining process, all activities, together, could define new and improved data sets for subsequent iterations. Generally, a good preprocessing method provides an optimal representation for a data-mining technique by incorporating a priori knowledge in the form of application-specific scaling and encoding.

1-2-4 Model Estimation

In this Step selecting the appropriate data-mining technique is the main task. This process is not straightforward; usually, in practice, the implementation is based on several models, and selecting the best one is an additional task. The basic principles of learning and discovery from data will give to explain and analyze specific techniques that are applied to perform a successful learning process from data and to develop an appropriate model.

1-2-5 Deducing the Model and Conclusions

In most cases, data-mining models should help in decision making. Hence, such models need to be interpretable in order to be useful because humans are not likely to base their decisions on complex models. Note that the goals of accuracy of the model and

accuracy of its interpretation are somewhat contradictory. Usually, simple models are more interpretable, but they are also less accurate. Modern data-mining methods are expected to yield highly accurate results using high dimensional models. The problem of interpreting these models, also very important, is considered a separate task, with specific techniques to validate the results. A user does not want hundreds of pages of numeric results. He does not understand them; he cannot summarize, interpret, and use them for successful decision making.

1-3 Data Warehouses and Data Marts

A primary goal of a data warehouse is to increase the "intelligence" of a decision process and the knowledge of the people involved in the process.. Note that the existence of a data warehouse is not a prerequisite for data mining, in practice, the task of data mining, become easier by having access to a data warehouses.

A data warehouse means different things to different people definitions are limited to data; some refer to people, processes, software, tools, and data. One of the global definitions is that:

- ✓ The data warehouse is a collection of integrated, subject-oriented databases designed to support the decision-support functions (DSF), where each unit of data is relevant to some moment in time.

Based on this definition, a data warehouse can be viewed as an organization's repository of data, set up to support strategic decision-making. The function of the data warehouse is to store the historical data of an organization in an integrated manner that reflects the various facets of the organization and business. The data in a warehouse are never updated but used only to respond to queries from end users who are generally decision-makers. Typically, data warehouses are huge, storing billions of records. In many instances, an organization may have several local or departmental data warehouses often called data marts. A data mart is a data warehouse that has been designed to meet the needs of a specific group of users. It may be large or small, depending on the subject areas.[2]

Today's computers and corresponding software tools support the processing of data sets with millions of samples and hundreds of features. Large data sets, including those with mixed data types, are a typical initial environment for application of data-mining techniques. When a large amount of data is stored in a computer, one cannot rush into data-mining techniques, because the important problem of data quality has first to be resolved. Also, it is obvious that a manual quality analysis is not possible at that stage. Therefore, it is necessary to prepare a data-quality analysis in the earliest phases of the data-mining process; usually it is a task to be undertaken in the data-preprocessing phase. The quality of data has a profound effect on the image of the system and determines the corresponding model that is implicitly described; it could also limit the ability of end users to make informed decisions. Using the available data-mining techniques, it will be difficult to undertake major qualitative changes in an organization if the data is of a poor

quality; similarly, to make new sound discoveries from poor quality scientific data will be almost impossible. There are a number of indicators of data quality:

1. The data should be accurate. The analyst has to check that the name is spelled correctly, the code is in a given range, the value is complete, etc.

2. The data should be stored according to data type. The analyst must ensure that the numeric value is not presented in character form, that integers are not in the form of real numbers, etc.

3. The data should have integrity. Updates should not be lost because of conflicts among different users; robust backup and recovery procedures should be implemented if they are not already part of the Data Base Management System (DBMS).

4. The data should be consistent. The form and the content should be the same after integration of large data sets from different sources.

5. The data should not be rebundant. In practice, redundant data should be minimized and reasoned duplication should be controlled. Duplicated records should be eliminated.

6. The data should be timely. The time component of data should be recognized explicitly from the data or implicitly from the manner of its organization.

7. The data should be well understood. Naming standards are a necessary but not the only condition for data to be well understood. The user should know that the data corresponds to an established domain.

8. The data set should be complete. Missing data, which occurs in reality, should be minimized. Missing data could reduce the quality of a global model. On the other hand, some data-mining techniques are robust enough to support analyses of data sets with missing values.

1-4 Forms of Inputs

With any software system, understanding what the inputs are is very important. The input takes the form of *concepts, instances,* and *attributes.*

1-4-1 Concept

The idea of a concept, like the very idea of learning in the first place, is hard to explain precisely, and we better not to spend time philosophizing about just what it is and isn't. In a sense, what we are trying to find—the result of the learning process—is a description of the concept that is *intelligible* in that it can be understood, discussed, and disputed, and *operational* in that it can be applied to actual examples.

There are some distinctions among different kinds of learning problems, that are very concrete and very important in practical data mining.

Four basically different styles of learning appear in data mining applications. In *classification learning,* the learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples.

In *association learning,* any association among features is sought, not just ones that predict a particular *class* value.

In *clustering,* groups of examples that belong together are sought. In *numeric prediction,* the outcome to be predicted is not a discrete class but a numeric quantity. Regardless of the type of learning involved, we call the thing to be learned the *concept* and the output produced by a learning scheme the *concept description.*

1-4-2 Instances

The information that the learner is given takes the form of a set of *instances.* Each instance is an individual, independent example of the concept to be learned. The input to a machine learning scheme is a set of instances. These instances are the things that are to be classified, associated, or clustered. Although until now we have called them *examples,* henceforth we will use the more specific term *instances* to refer to the input. Each