Applying Data Mining Techniques in Property/Casualty Insurance

Lijia Guo, Ph.D., ASA

Applying Data Mining Techniques in Property/Casualty Insurance

Lijia Guo, Ph.D., A.S.A.
University of Central Florida

Abstract

This paper addresses the issues and techniques for Property/Casualty actuaries using data mining techniques. Data mining means the efficient discovery of previously unknown patterns in large databases. It is an interactive information discovery process that includes data acquisition, data integration, data exploration, model building, and model validation. The paper provides an overview of the information discovery techniques and introduces some important data mining techniques for application to insurance including cluster discovery methods and decision tree analysis.

1. Introduction

Because of the rapid progress of information technology, the amount of information stored in insurance databases is rapidly increasing. These huge databases contain a wealth of data and constitute a potential goldmine of valuable business information. As new and evolving loss exposures emerge in the ever-changing insurance environment, the form and structure of insurance databases change. In addition, new applications such as dynamic financial analysis and catastrophe modeling require the storage, retrieval, and analysis of complex multimedia objects, which are often represented by high-dimensional feature vectors. Finding the valuable information hidden in those databases and identifying appropriate models is a difficult task.

Data mining (DM) is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules (Berry and Linoff, 2000). A typical data mining process includes data acquisition, data integration, data exploration, model building, and model validation. Both expert opinion and data mining techniques play an important role at each step of this information discovery process.

This paper introduces two important data mining techniques for application to insurance: cluster discovery methods and decision tree analysis.

Cluster analysis is one of the basic techniques that are often applied in analyzing large data sets. Originating from the area of statistics, most cluster analysis algorithms have originally been developed for relatively small data sets. In recent years, the clustering algorithms have been extended to efficiently work on large data sets, and some of them even allow the clustering of high-dimensional feature vectors (see Ester, Kriegel, Sander, and Xu, and Hinneburg, and Keim, 1998, for example).

Decision tree analysis is another popular data mining technique that can be used in many areas of actuarial practice. We discuss how to use decision trees to make important design decisions and explain the interdependencies among the properties of insurance data. We will also provide examples of how data mining techniques can be used to improve the effectiveness and efficiency of the modeling process.

The paper is organized as follows. Section 2 provides an overview of data mining and a list of potential DM applications to insurance. Section 3 demonstrates the cluster analysis data mining techniques. Section 4 presents application of predictive data mining process. This section identifies factors that influence auto insurance claims using decision tree techniques and quantifies the effects and interactions of these risk factors using logistic regression. Model assessment is also discussed in this section. Section 5 concludes the paper.

2. Data Mining

In this section, we will provide an overview of the data mining process (2.1), data mining operations (2.2), data mining techniques and algorithms (2.3), and their potential applications in the insurance industry (2.4).

2.1 Data Mining Process

Data mining combines techniques from machine learning, pattern recognition, statistics, database theory, and visualization to extract concepts, concept interrelations, and interesting patterns automatically from large corporate databases. Its primary goal is to extract knowledge from data to support the decision-making process. Two primary functions of data mining are: *prediction,* which involves finding unknown values/relationships/patterns from known values; and *description,* which provides interpretation of a large database.

A data mining process generally includes the following four steps.

STEP 1: Data acquisition. The first step is to select the types of data to be used. Although a target data set has been created for discovery in some applications, DM can be performed on a subset of variables or data samples in a larger database.

STEP 2: Preprocessing data. Once the target data is selected, the data is then preprocessed for cleaning, scrubbing, and transforming to improve the effectiveness of discovery. During this preprocessing step, developers remove the noise or outliers if necessary and decide on strategies for dealing with missing data fields and accounting for time sequence information or known changes. In addition, the data is often transformed to reduce the effective number of variables under consideration by either converting one type of data to another (e.g., categorical values into numeric ones) or deriving new attributes (by applying mathematical or logical operators).

STEP 3: Data exploration and model building. The third step of DM refers to a series of activities such as deciding on the type of DM operation; selecting the DM technique; choosing

the DM algorithm; and mining the data. First, the type of DM operation must be chosen. The DM operations can be classified as classification, regression, segmentation, link analysis, and deviation detection (see Section 2.2 for details). Based on the operation chosen for the application, an appropriate data-mining technique is then selected. Once a data-mining technique is chosen, the next step is to select a particular algorithm within the DM technique chosen. Choosing a data-mining algorithm includes a method to search for patterns in the data, such as deciding which models and parameters may be appropriate and matching a particular data-mining technique with the overall objective of data mining. After an appropriate algorithm is selected, the data is finally mined using the algorithm to extract novel patterns hidden in databases.

STEP 4: Interpretation and evaluation. The fourth step of the DM process is the interpretation and evaluation of discovered patterns. This task includes filtering the information to be presented by removing redundant or irrelevant patterns, visualizing graphically or logically the useful ones, and translating them into understandable terms by users. In the interpretation of results, we determine and resolve potential conflicts with previously found knowledge or decide to redo any of the previous steps. The extracted knowledge is also evaluated in terms of its usefulness to a decision maker and to a business goal. Then extracted knowledge is subsequently used to support human decision making such as prediction and to explain observed phenomena.

The four-step process of knowledge discovery should not be interpreted as linear, but as an interactive, iterative process through which discovery evolves.

2.2 Data Mining Operations

Assuming you have prepared a data set for mining, you then need to define the scope of your study and choose the subject of your study. This is referred as choosing a DM operation.

There are five types of DM operations: classification, regression, link analysis, segmentation, and deviation detection. Classification and regression are useful for prediction, whereas link analysis, segmentation, and deviation detection are for description of patterns in the data. A DM application typically requires the combination of two or more DM operations.

Classification
The goal of classification is to develop a model that maps a data item into one of several predefined classes. Once developed, the model is used to classify a new instance into one of the classes. Examples include the classification of bankruptcy patterns based on the financial ratios of a firm and of customer buying patterns based on demographic information to target the advertising and sales of a firm effectively toward the appropriate customer base.

Regression
This operation builds a model that maps data items into a real-valued prediction variable. Models have traditionally been developed using statistical methods such as linear and logistic regression. Both classification and regression are used for prediction. The distinction between these two models is that the output variable of classification is categorical, whereas that of

Table 1. DM Techniques for DM Operations

DM Technique	Induction	Neural Networks	Genetic Algorithms	Clustering	Logistic Regression	Association Discovery	Sequence Discovery	Visualization
DM Operation								
Classification	x	x	x					
Regression	x	x			x			
Link analysis		x	x			x	x	
Segmentation		x	x	x				
Deviation					x			x

Induction Techniques

Induction techniques develop a classification model from a set of records -- the training set of examples. The training set may be a sample database, a data mart, or an entire data warehouse. Each record in the training set belongs to one of many predefined classes, and an induction technique induces a general concept description that best represents the examples to develop a classification model. The induced model consists of patterns that distinguish each class. Once trained, a developed model can be used to predict the class of unclassified records automatically. Induction techniques represent a model in the form of either decision trees or decision rules. These representations are easier to understand, and their implementation is more efficient than those of neural network or genetic algorithms. A more detailed discussion on decision tree techniques and their applications will be presented in Section 4.

Neural Networks

Neural networks constitute the most widely used technique in data mining. They imitate the way the human brain learns and use rules inferred from data patterns to construct hidden layers of logic for analysis. Neural networks methods can be used to develop classification, regression, link analysis, and segmentation models. A neural net technique represents its model in the form of nodes arranged in layers with weighted links between the nodes. There are two general categories of neural net algorithms: supervised and unsupervised.

- Supervised neural net algorithms such as Back propagation (Rumelhart, Hinton, and Williams, 1986) and Perceptron require predefined output values to develop a classification model. Among the many algorithms, Back propagation is the most popular supervised neural net algorithm. Back propagation can be used to develop not only a classification model, but also a regression model.

- Unsupervised neural net algorithms such as ART (Carpenter and Grossberg, 1988) do not require predefined output values for input data in the training set and employ self-organizing learning schemes to segment the target data set. Such self-organizing networks divide input examples into clusters depending on similarity, each cluster representing an unlabeled category. Kohonen's Feature Map is a well-known method in self-organizing neural networks.

For organizations with a great depth of statistical information, neural networks are ideal because they can identify and analyze changes in patterns, situations, or tactics far more

regression is numeric and continuous. Examples of regression are the prediction of change between the yen and the Government Bond Market and of the crime rate of a city based on the description of various input variables such as populations, average income level and education.

Link Analysis

Link analysis is used to establish relevant connections between database records. Its typical application is market-basket analysis, where the technique is applied to analyze point-of-sales transaction data to identify product affinities. A retail store is usually interested in what items sell together -- such as baby's diapers and formula -- so it can determine what items to display together for effective marketing. Another application could find relationships among medical procedures by analyzing claim forms submitted to an insurance firm. Link analysis is often applied in conjunction with database segmentation.

Segmentation

The goal is to identify clusters of records that exhibit similar behaviors or characteristics hidden in the data. The clusters may be mutually exclusive and exhaustive or may consist of a richer representation such as hierarchical or overlapping categories. Examples include discovering homogenous groups of consumers in marketing databases and segmenting the records that describe sales during "Mother's Day" and "Father's Day." Once the database is segmented, link analysis is often performed on each segment to identify the association among the records in each cluster.

Deviation Detection

This operation focuses on discovering interesting deviations. There are four types of deviation:

- Unusual patterns that do not fit into previously measured or normative classes,
- Significant changes in the data from one time period to the next,
- Outlying points in a dataset -- records that do not belong to any particular cluster, and
- Discrepancies between an observation and a reference.

Deviation Detection is usually performed after a database is segmented to determine whether the deviations represent noisy data or unusual casualty. Deviation detection is often the source of true discovery since deviations represent anomaly from some known expectation or norm.

2.3 Data Mining Techniques and Algorithms

At the heart of DM is the process of building a model to represent the data set and to carry out the DM operation. A variety of DM techniques (tools) are available to support the five types of DM operations presented in the previous section. The most popular data mining techniques include Bayesian analysis (Cheeseman et al., 1988), neural networks (Bishop, 1995; Ripley, 1996), genetic algorithms (Goldberg, 1989), decision trees (Breiman et al., 1984), and logistic regression (Hosmer and Lemeshow, 1989), among others.

Table 1 summarizes the DM techniques used for DM operations. For each of the DM techniques listed in Table 1, there are many algorithms (approaches) to choose from. In the following, some of the most popular technologies are discussed.

quickly than any human mind. Although the neural net technique has strong representational power, interpreting the information encapsulated in the weighted links can be very difficult. One important characteristic of neural networks is that they are opaque, which means there is not much explanation of how the results come about and what rules are used. Therefore, some doubt is cast on the results of the data mining. Francis (2001) gives a discussion on Neural Network applications to insurance problems.

Genetic Algorithms

Genetic algorithms are a method of combinatorial optimization based on processes in biological evolution. The basic idea is that over time, evolution has selected the "fittest species." For a genetic algorithm, one can start with a random group of data. A *fitness function* can be defined to optimizing a model of the data to obtain "fittest" models. For example, in clustering analysis, a fitness function could be a function to determine the level of similarity between data sets within a group.

Genetic algorithms have often been used in conjunction with neural networks to model data. They have been used to solve complex problems that other technologies have a difficult time with. Michaelewicz (1994) introduced the concept of genetic algorithms and applying them with data mining.

Logistic Regression

Logistic regression is a special case of generalized linear modeling. It has been used to study odds ratios (e^{β_j}, $j = 1, 2, \cdots, k$ as defined in the following), which compares the odds of the event of one category to the odds of the event in another category, for a very long time and its properties have been well studied by the statistical community. Ease of interpretation is one advantage of modeling with logistic regression. Assume that the data set consist of i = 1, 2, ..., n records. Let $p_i, i = 1, 2, \cdots, n$ be the corresponding mortality rate for each record and $x_i = (x_{1i}, x_{2i}, \cdots, x_{ki})$ be a set of k variables associated with each record. A linear-additive logistic regression model can be expressed as

$$\text{logit} = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}, \; where \; i = 1, 2, \cdots, n.$$

If the model is correctly specified, each dependent variable affects logit linearly. Exponentiation of the parameter estimate of each slope, e^{β_j}, $j = 1, 2, \cdots, k$, can be interpreted as the odds ratio of the probability that p_i is associated with input variable x_{ji} (Kleinbaum, D., Kupper, L., and Muller, K., 1988). However, it poses several drawbacks especially with large data sets. The curse of dimensionality makes the detection of nonlinearities and interactions difficult. If the model is not correctly specified, the interpretation of the model parameter estimates becomes meaningless. In addition, the data might not be evenly distributed among the whole data space. It is very likely that some segments of the data space have more records than other segments. One model that fits the whole data space might not be the best choice depending on the intended application. Although there are many existing methods such as backward elimination and forward selection that can help data analyst to

build logistic regression model, judgment should be exercised regardless of the method selected.

Clustering

Clustering techniques are employed to segment a database into clusters, each of which shares common and interesting properties. The purpose of segmenting a database is often to summarize the contents of the target database by considering the common characteristics shared in a cluster. Clusters are also created to support the other types of DM operations, e.g. link analysis within a cluster. Section 3 will introduce more details of clustering and its application to insurance.

Associated Discovery

Given a collection of items and a set of records containing some of these items, association discovery techniques discover the rules to identify affinities among the collection of items as reflected in the examined records. For example, 65 percent of records that contain item A also contain item B. An association rule uses measures called "support" and "confidence" to represent the strength of association. The percentage of occurrences, 65 percent in this case, is the confidence factor of the association. The algorithms find the affinity rules by sorting the data while counting occurrences to calculate confidence. The efficiency with which association discovery algorithms can organize the events that make up an association or transaction is one of the differentiators among the association discovery algorithms. There are a variety of algorithms to identify association rules such as Apriori algorithm and using random sampling. Bayesian Net can also be used to identify distinctions and relationships between variables (Fayyad et al., 1996).

Sequence Discovery

Sequence discovery is very similar to association discovery except that the collection of items occurs over a period of time. A sequence is treated as an association in which the items are linked by time. When customer names are available, their purchase patterns over time can be analyzed. For example, it could be found that, if a customer buys a tie, he will buy men's shoes within one month 25 percent of the time. A dynamic programming approach based on the dynamic time warping technique used in the speech recognition area is available to identify the patterns in temporal databases (Fayyad et al., 1996).

Visualization

A picture is worth thousands of numbers! Visual DM techniques have proven the value in exploratory data analysis, and they also have a good potential for mining large databases. Visualizations are particularly useful for detecting phenomena hidden in a relatively small subset of the data. This technique is often used in conjunction with other DM techniques: features that are difficult to detect by scanning numbers may become obvious when the summary of data is graphically presented. Visualization techniques can also guide users when they do not know what to look for to discover the feature. Also, this technique helps end users comprehend information extracted by other DM techniques. Specific visualization techniques include projection pursuit and parallel coordinates. Tufte (1983, 1990) provided many examples of visualization techniques that have been extended to work on large data sets and produce interactive displays.

2.4 Using Data Mining in the Insurance Industry

Data mining methodology can often improve existing actuarial models by finding additional important variables, by identifying interactions, and by detecting nonlinear relationships. DM can help insurance firms make crucial business decisions and turn the new found knowledge into actionable results in business practices such as product development, marketing, claim distribution analysis, asset liability management and solvency analysis. An example of how data mining has been used in health insurance can be found in Borok, 1997. To be more specific, data mining can perform the following tasks.

Identify Risk Factors that Predict Profits, Claims and Losses

One critical question in ratemaking 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 such as decision trees and Neural Networks can more accurately predict risk than current actuarial models, therefore insurance companies can set rates more accurately, which in turn can result in more accurate pricing and hence a better competitive position.

Customer Level Analysis

Successfully retaining customers requires analyzing data at the most appropriate level, the customer level, instead of across aggregated collections of customers. Using the Associated Discovery DM technique, insurance firms can more accurately select which policies and services to offer to which customers. With this technique insurance companies can:

- Segment the customer database to create customer profiles.
- Conduct rate and claim analyses on a single customer segment for a single product. For example, companies can perform an in-depth analysis of a potential new product for a particular customer segment.
- Analyze customer segments for multiple products using group processing and multiple target variables. For example, how profitable are bundles of policies (auto, home, and life) for certain customer segments of interest?
- Perform sequential (over time) market basket analyses on customer segments. For example, what percentage of new policyholders of auto insurance also purchases a life insurance policy within five years?

Database segmentation and more advanced modeling techniques enable analysts to more accurately choose whom to target for retention campaigns. Current policyholders that are likely to switch can be identified through predictive modeling. A logistic regression model is a traditional approach to predict those policyholders who have larger probabilities of switching. Identifying the target group for retention campaigns may be improved by modeling the behavior of policyholders.

Developing New Product Lines

Insurance firms can increase profitability by identifying the most lucrative customer segments and then prioritize marketing campaigns accordingly. Problems with profitability can occur if

firms do not offer the "right" policy or the "right" rate to the "right" customer segment at the "right" time. For example, for an insurer or reinsurer to use the log normal distribution for rating when the Pareto distribution is the true distribution would likely prove to be an expensive blunder, which illustrates the importance of having the right tool to identify and estimate the underlying loss distribution. With DM operations such as segmentation or association analysis, insurance firms can now utilize all of their available information to better develop new products and marketing campaigns.

Reinsurance

DM can be used to structure reinsurance more effectively than the using traditional methods. Data mining technology is commonly used for segmentation clarity. In the case of reinsurance, a group of paid claims would be used to model the expected claims experience of another group of policies. With more granular segmentation, analysts can expect higher levels of confidence in the model's outcome. The selection of policies for reinsurance can be based upon the model of experienced risk and not just the generalization that it is a long tailed book of business.

Estimating Outstanding Claims Provision

The settlement of claims is often subject to delay, so an estimate of the claim severity is often used until the actual value of the settled claim is available. The estimate can depend on the following:

- Severity of the claim.
- Likely amount of time before settlement.
- Effects of financial variables such as inflation and interest rates.
- Effects of changing social mores. For example, the tobacco industry has been greatly affected by the changing views toward smoking.

DM operations such as Link Analysis and Deviation Detection can be used to improve the claim estimation.

The estimate of the claims provision generated from a predictive model is based on the assumption that the future will be much like the past. If the model is not updated, then over time, the assumption becomes that the future will be much like the distant past. However, as more data become available, the predictive DM model can be updated, and the assumption becomes that the future will be much like the recent past. Data mining technology enables insurance analysts to compare old and new models and to assess them based on their performance. When the newly updated model outperforms the old model, it is time to switch to the new model. Given the new technologies, analysts can now monitor predictive models and update as needed.

An important general difference in the focus between existing actuarial techniques and DM is that DM is more oriented towards applications than towards describing the basic nature of the underlying phenomena. For example, uncovering the nature of the underlying individual claim distribution or the specific relation between drivers' age and auto type are not the main goal of Data Mining. Instead, the focus is on producing a solution that can improve the predictions for future premiums. DM is very effective in determining how the premiums related to

multidimensional risk factors such as drivers' age and type of automobile. Two examples of applying data mining techniques in insurance actuarial practice will be presented in the next two sections.

3. Clustering - Descriptive Data Mining

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster is more similar to each other than points in different clusters (Guha et al., 1998). For example, segmenting existing policyholders into groups and associating a distinct profile with each group can help future rate making strategies.

Clustering methods perform disjoint cluster analysis on the basis of Euclidean distances computed from one or more quantitative variables and seeds that are generated and updated by the algorithm. You can specify the clustering criterion that is used to measure the distance between data observations and seeds. The observations are divided into clusters such that every observation belongs to at most one cluster.

Clustering studies are also referred to as unsupervised learning and/or segmentation. Unsupervised learning is a process of classification with an unknown target, that is, the class of each case is unknown. The aim is to segment the cases into disjoint classes that are homogenous with respect to the inputs. Clustering studies have no dependent variables. You are not profiling a specific trait as in classification studies.

A database can be segmented by:

- Traditional methods of pattern recognition techniques,
- Unsupervised neural nets such as ART and Kohonen's Feature Map,
- Conceptual clustering techniques such as COBWEB (Fisher, Pazzani and Langley, 1991) and UNIMEM, or
- A Bayesian approach like AutoClass (Chessman, 1996).

Conceptual clustering algorithms consider all the attributes that characterize each record and identify the subset of the attributes that will describe each created cluster to form concepts. The concepts in a conceptual clustering algorithm can be represented as conjunctions of attributes and their values. Bayesian clustering algorithms automatically discover a clustering that is maximally probable with respect to the data using a Bayesian approach. The various clustering algorithms can be characterized by the type of acceptable attribute values such as continuous, discrete or qualitative; by the presentation methods of each cluster; and by the methods of organizing the set of clusters, either hierarchically or into flat files. K-mean clustering, a basic clustering algorithm is introduced in the following.

3.1 K-means clustering

Problem Description:

Given a data set with N n-dimensional data points x^n, the goal is to determine a natural partitioning of the data set into a number of clusters (k) and noise. We know there are k disjoint clusters containing N_j data points with representative vector μ_j where j=1, ..., k. The K-means algorithm attempts to minimize the sum-of-squares clustering function given by

$$J = \sum_{j=1}^{k} \sum_{n \in S_j} \left\| x^n - \mu_j \right\|^2$$

where μ_j is the mean of the data points in cluster S_j and is given by

$$\mu_j = \frac{1}{N_j} \sum_{n \in S_j} x^n .$$

The training is carried out by assigning the points at random to k clusters and then computing the mean vectors μ_j of the N_j points in each cluster. Each point is re-assigned to a new cluster according to which is the nearest mean vector. The mean vectors are then recomputed.

K-means clustering proceeds as follows:
1. Specify the number of clusters (classes) k.
2. Choose k initial cluster seeds.
3. Assign cases closest to seed j as belonging to cluster j, j=1,..., k.
4. Calculate the mean of the cases in each cluster, and move the k cluster seeds to the mean of their cluster.
5. Reassign cases closest to the new seed j as belonging to cluster j.
6. Take the mean of the cases in each cluster as the new cluster seed.

This procedure is repeated until there is no further change in clustering.

K-means clustering is an unsupervised classification method. It is computationally efficient provided the initial cluster seeds are intelligently placed. Clustering methods depend on a measure of distance or similarity between points. Different distance metrics used in k-means clustering can result in different clusters.

3.2. Example: Clustering Automobile Drivers

The ABC Insurance Company periodically purchases lists of drivers from outside sources. Actuaries at ABC want to evaluate the potential claim frequency for underwriting purposes. Based on their experience, they know that driver claim frequency depends on geographic and demographic factors. Consequently, they want to segment the drivers into groups that are similar to each other with respect to these attributes. After the drivers have been segmented, a random sample of prospects within each segment will be used to estimate the frequency. The results of this test estimate will allow the actuaries to evaluate the potential profit of prospects from the list, both overall as well as for specific segments.

The synthetic data that was obtained from the vendor is given in Table 2.

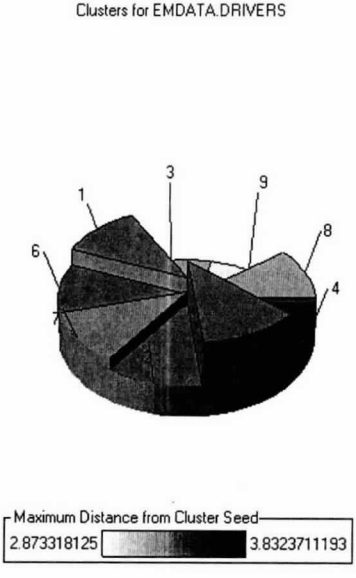
After preprocessing the data, which might include selecting a random sample of the data for initial analysis, filtering the outlying observations, and standardizing the variables in some way, we use the K-means clustering to form the clusters.

Table 2. Automobile Drivers Data

Variable	*Variable Type*	*Measurement Level*	*Description*
Age	Continuous	Interval	Driver's age in years
Car age	Continuous	Interval	Age of the car in years
Car type	Categorical	Nominal	Type of the car
Gender	Categorical	Binary	F=female, M=male
Coverage level	Categorical	Nominal	Policy coverage
Education	Categorical	Nominal	Education level of the drive
Location	Categorical	Nominal	Location of residence
Climate	Categorical	Nominal	Climate code for residence
Credit rating	Continuous	Interval	Credit score of the driver
ID	Input	Nominal	Driver's identification number

The following pie chart provides a graphical representation of key characteristics of the clusters.

Figure 1 Clusters Pie Chart

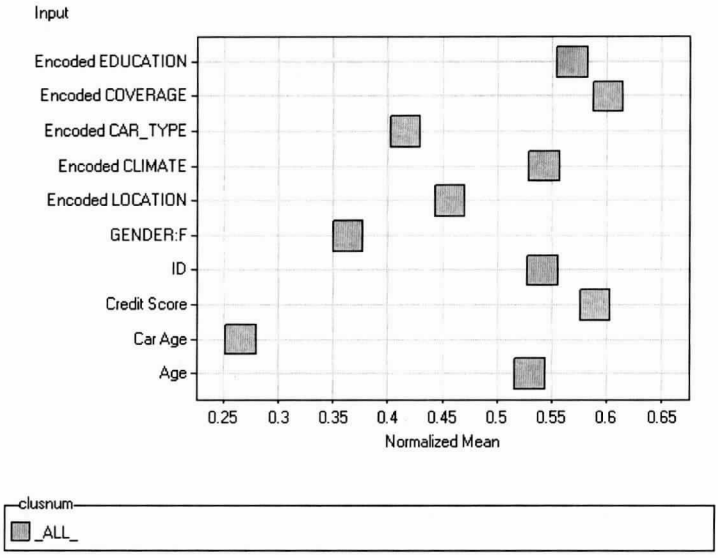


In the pie chart, slice width is the root-mean-square distance (root-mean-square standard deviation) between cases in the cluster; the height means the frequency and the color represents the distance of the farthest cluster member from the cluster. Cluster 5 contains the most cases while cluster 9 has the fewest.

Figure 2 below displays the input means for the entire data set over all of the clusters. The input means are normalized using a scale transformation

$$y = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

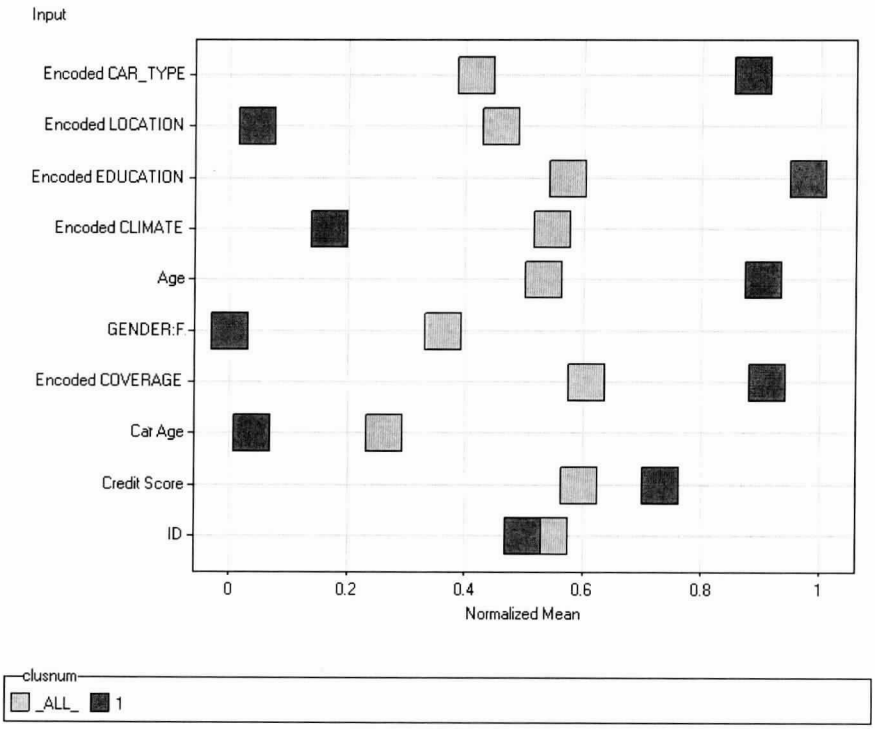
Figure 2. Overall Input Means



The Normalized Mean Plot can be used to compare the overall normalized means with the normalized means in each cluster. Figure 3 compare the input means from cluster 1 (red blocks) to the overall input means (blue blocks). You want to identify the input means for clusters that differ substantially from the overall input means. The plot ranks the input based on how spread out the input means are for the selected cluster relative to the overall input means. The input that has the biggest spread is listed at the top and the input with the smallest spread is listed at the bottom. The input with the biggest spread typically best characterizes the selected cluster (Cluster 1 in Figure 3). Figure 3 shows that the variable "Car-Type" and "Location" are key inputs that help differentiate drivers in Cluster 1 from all of the drivers in

the data set. Drivers in Cluster 1 tend to have higher than average education levels than average drivers in the data set.

Figure 3. Comparing the Input Means for Cluster 1 to the Overall Means



Cluster 5, as shown in Figure 4, has higher than average education and better than average credit scores. Most drivers in Cluster 5 live in location zone 4 and they drive newer car than average drivers. These characteristics can also be observed from Table 3.

Figure 4. Comparing the Input Means for Cluster 5 to the Overall Means

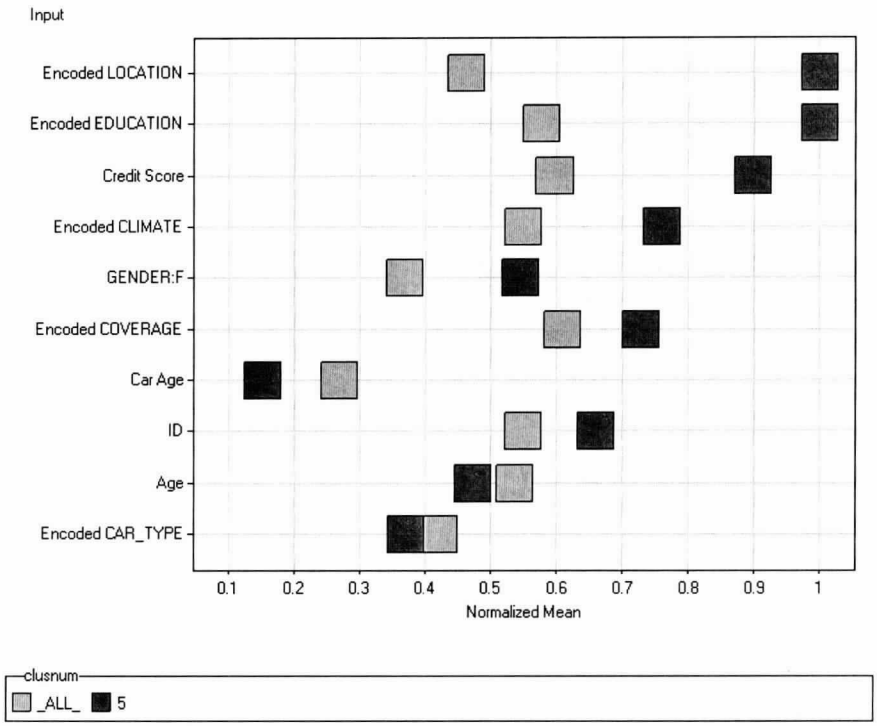


Table 3 displays information about each cluster. The statistics Root-Mean-Square Standard Deviation means the root-mean-square error across variables of the cluster standard deviations, which is equal to the root-mean-square distance between cases in the cluster.

Table 3. Clustering Summary Statistics

Cluster	Frequency of Cluster	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Credit Score	Car Age	Age	Gender	Location	Climate	Car Type	Coverage	Education
9	7	2.87	5	2.82	0.86	3.29	35.57	1.00	3.43	1.29	3.57	2.43	1.86
8	20	3.22	7	2.40	0.62	2.15	46.65	0.65	2.80	2.55	2.25	2.85	1.85
7	22	3.25	2	2.25	0.65	2.73	24.59	0.27	1.95	2.09	1.45	2.36	2.27
6	21	3.38	4	2.41	0.81	6.52	35.19	0.43	2.00	1.48	1.67	1.19	1.76
5	33	3.41	4	2.37	0.82	3.00	32.79	0.58	3.82	2.33	2.03	2.39	3.03
4	18	3.83	5	2.37	0.59	5.17	34.44	0.39	3.50	1.83	2.72	1.44	2.56
3	7	3.21	7	3.14	0.46	8.00	20.57	0.43	3.57	2.43	1.14	1.43	2.00
2	18	3.38	7	2.25	0.56	3.56	26.00	0.28	2.89	2.67	1.28	2.28	1.39
1	27	3.40	5	2.55	0.75	2.37	44.15	0.07	2.04	1.52	3.30	2.70	3.00

During the clustering process, an importance value is computed as a value between 0 and 1 for each variable. Importance is a measure of worth of the given variable to the formation of the clusters. As shown in Table 4, variable "Gender" has an importance of 0, which means that the variable was not used as a splitting variable in developing the clusters. The measure of "importance" indicates how well the variable divides the data into classes. Variables with zero importance should not necessary be dropped.

Table 4. Variable Importance

Name	Importance
GENDER	0
ID	0
LOCATION	0
CLIMATE	0
CAR_TYPE	0.529939
COVERAGE	0.363972
CREDIT_SCORE	0.343488
CAR_AGE	0.941952
AGE	1
EDUCATION	0.751203

Clustering analysis can be used by the property/casualty insurance industry to improve predictive accuracy by segmenting databases into more homogeneous groups. Then the data of each group can be explored, analyzed, and modeled. Segments based on types of variables that associate with risk factors, profits, or behaviors often provide sharp contrasts, which can

be interpreted more easily. As a result, actuaries can more accurately predict the likelihood of a claim and the amount of the claim. For example, one insurance company found that a segment of the 18- to 20-year old male drivers had a noticeably lower accident rate than the entire group of 18- to 20-year old males. What variable did this subgroup share that could explain the difference? Investigation of the data revealed that the members of the lower risk subgroup drove cars that were significantly older than the average and that the drivers of the older cars spent time customizing their "vintage autos." As a result, members of the subgroup were likely to be more cautious driving their customized automobiles than others in their age group.

Lastly, the cluster identifier for each observation can be passed to other nodes for use as an input, id, group, or target variable. For example, you could form clusters based on different age groups you want to target. Then you could build predictive models for each age group by passing the cluster variable as a group variable to a modeling node.

4. Predictive Data Mining

This section introduces data mining models for prediction (as opposed to description, such as in Section 3). Section 4.1 gives an overview of the Decision Tree DM algorithm. Section 4.2 presents a claim frequency model using Decision Trees and Logistic Regression.

4.1 Decision Trees

Decision trees are part of the Induction class of DM techniques. An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. One rule is applied after another, resulting in a hierarchy of segments within segments. The hierarchy is called a tree, and each segment is called a node. The original segment contains the entire data set and is called the root node of the tree. A node with all its successors forms a branch of the node that created it. The final nodes are called leaves. For each leaf, a decision is made and applied to all observations in the leaf. The type of decision depends on the context. In predictive modeling, the decision is simply the predicted value.

The decision tree DM technique enables you to create decision trees that:

- Classify observations based on the values of nominal, binary, or ordinal targets,
- Predict outcomes for interval targets, or
- Predict the appropriate decision when you specify decision alternatives.

Specific decision tree methods include Classification and Regression Trees (CART; Breiman et. al., 1984) and the count or Chi-squared Automatic Interaction Detection (CHAID; Kass, 1980) algorithm. CART and CHAID are decision tree techniques used to classify a data set.

The following discussion provides a brief description of the CHAID algorithm for building decision trees. For CHAID, the inputs are either nominal or ordinal. Many software packages accept interval inputs and automatically group the values into ranges before growing the tree. For nodes with many observations, the algorithm uses a sample for the split search, for computing the worth (measure of worth indicates how well a variable divides the data into

each class), and for observing the limit on the minimum size of a branch. The samples in different nodes are taken independently. For binary splits on binary or interval targets, the optimal split is always found. For other situations, the data is first consolidated, and then either all possible splits are evaluated or else a heuristic search is used.
The consolidation phase searches for groups of values of the input that seem likely to be assigned the same branch in the best split. The split search regards observations in the same consolidation group as having the same input value. The split search is faster because fewer candidate splits need evaluating. A primary consideration when developing a tree for prediction is deciding how large to grow the tree or, what comes to the same end, what nodes to prune off the tree. The CHAID method of tree construction specifies a significance level of a Chi-square test to stop tree growth. The splitting criteria are based on p-values from the F-distribution (interval targets) or Chi-square distribution (nominal targets). For these criteria, the best split is the one with the smallest p-value. By default, the p-values are adjusted to take into account multiple testing.

A missing value may be treated as a separate value. For nominal inputs, a missing value constitutes a new category. For ordinal inputs, a missing value is free of any order restrictions.

The search for a split on an input proceeds stepwise. Initially, a branch is allocated for each value of the input. Branches are alternately merged and re-split as seems warranted by the p-values. The original CHAID algorithm by Kass stops when no merge or re-splitting operation creates an adequate p-value. The final split is adopted. A common alternative, sometimes called the exhaustive method, continues merging to a binary split and then adopts the split with the most favorable p-value among all splits the algorithm considered.

After a split is adopted for an input, its p-value is adjusted, and the input with the best-adjusted p-value is selected as the splitting variable. If the adjusted p-value is smaller than a threshold you specified, then the node is split. Tree construction ends when all the adjusted p-values of the splitting variables in the unsplit nodes are above the user-specified threshold.

Tree techniques provide insights into the decision-making process, which explains how the results come about. The decision tree is efficient and is thus suitable for large data sets. Decision trees are perhaps the most successful exploratory method for uncovering deviant data structure. Trees recursively partition the input data space in order to identify segments where the records are homogeneous. Although decision trees can split the data into several homogeneous segments and the rules produced by the tree can be used to detect interaction among variables, it is relatively unstable and it is difficult to detect linear or quadratic relationships between the response variable and the dependent variables.

4.2 Modeling claim frequency

We now start the modeling process by studying the relationship between claim frequency and underlying risk factors including age, gender, credit score, location, education level, coverage, ..., and car age. Again, the synthetic data is used. A hybrid method is developed for this study – the modeling process is a combination of the decision tree techniques and logistic regression.

First, we use the decision tree algorithm to identify the factors that influence claim frequency. After the factors are identified, the logistic regression technique is used to quantify the claim frequency and the effect of each risk factor.

The data for the study has the following variables as shown in Table 5:

Table 5. Automobile Driver's Claim Information

Variable	*Variable Type*	*Measurement Level*	*Description*
Age	Continuous	Interval	Driver's age in years
Car age	Continuous	Interval	Age of the car
Car type	Categorical	Nominal	Type of the car
Gender	Categorical	Binary	F=female, M=male
Coverage level	Categorical	Nominal	Policy coverage
Education	Categorical	Nominal	Education level of the drive
Location	Categorical	Nominal	Location of residence
Climate	Categorical	Nominal	Climate code for residence
Credit rating	Continuous	Interval	Credit score of the driver
ID	Input	Nominal	Driver's identification number
No. of claims	Categorical	Nominal	Number of claims

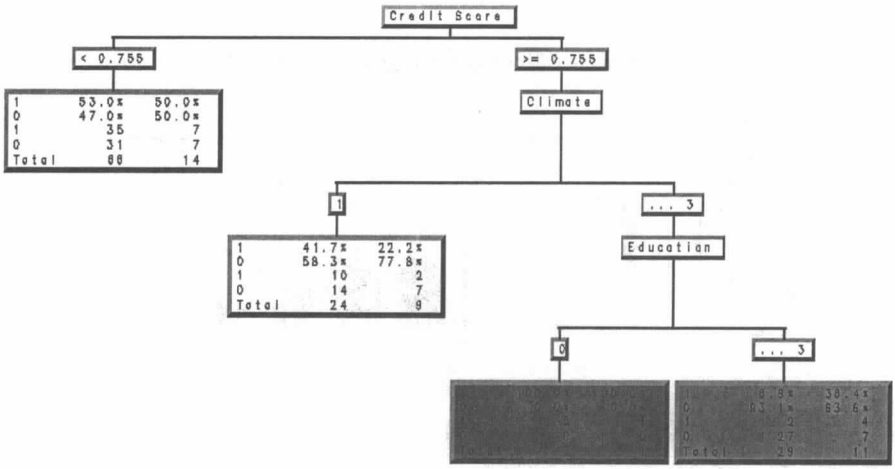
We now use the decision tree algorithm to analyze the influences and the importance of the claim frequency risk factors. The tree algorithm used in this research is SAS/Enterprise Miner Version 4.2 (2002). We built 100 binary regression trees and 100 CHAID-like trees for optimal decision tree. Our decision tree analysis reveals that the credit score has the greatest impact on the claim frequency. The claim frequency, and the interaction among different factors that affect the claim frequency, vary as the credit score status changes. Furthermore, there is a significant climate influence within the "higher credit score" status.

A Tree diagram contains the following items:

- Root node -- top node in the tree that contains all observations.
- Internal nodes – non-terminal nodes (including the root node) that contain the splitting rules.
- Leaf nodes -- terminal nodes that contain the final classification for a set of observations.

The tree diagram displays node (segment) statistics, the names of variables used to split the data into nodes, and the variable values for several levels of nodes in the tree. Figure 5 shows a partial profile of the tree diagram for our analysis:

Figure 5. Tree Diagram



In Figure 5, each leaf node displays the percentage and n-count of the values that were used to determine the branching. The second column contains the learning from the training data including the percentage for each target level, the count for each target level, and the total count. The third column contains the learning from the validation data including the percentage for each target level, the count for each target level, and the total count. For example, among these drivers with credit score below 75.5%, 53% of them submitted a claim from the training data.

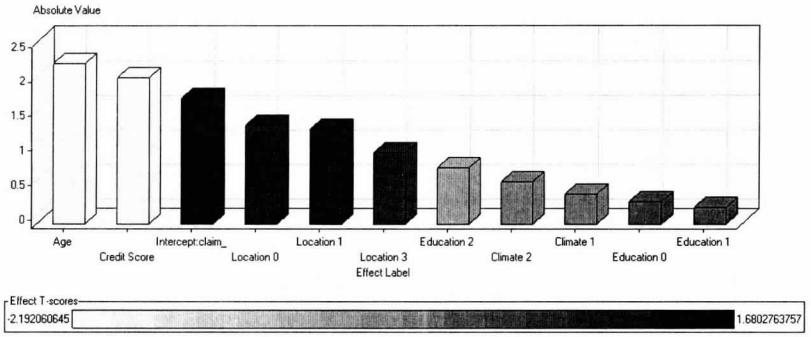
The assessment values are used to recursively partition the data in homogenous subgroups. The method is recursive because each subgroup results from splitting a subgroup from a previous split. The numeric labels directly above each node indicate at which point the tree algorithm found significant splits in interval level variable distributions or in categorical splits for nominal or ordinal level distributions. The character labels positioned central to each split are the variable names. You can trace the paths from the root to each leaf and express the results as a rule.

As shown in Figure 5, the claim frequency varies with the most important risk factor (the credit score status, in this study) among all the other variables. Based on tree analysis, the car age, coverage, and car-type are the irrelevant factors. They should not be included in the claim frequency model.

Based on the tree analysis, we now use logistic regression to estimate the probability of claim occurrence for each driver based on the factors under consideration. As discussed in Section 2,

logistic regression attempts to predict the probability of a claim as a function of one or more independent inputs. Figure 6 shows a bar chart of the effect T-scores from the logistic regression analysis. An effect T-score is equal to the parameter estimate divided by its standard error.

Figure 6. Effect T-scores from the logistic regression

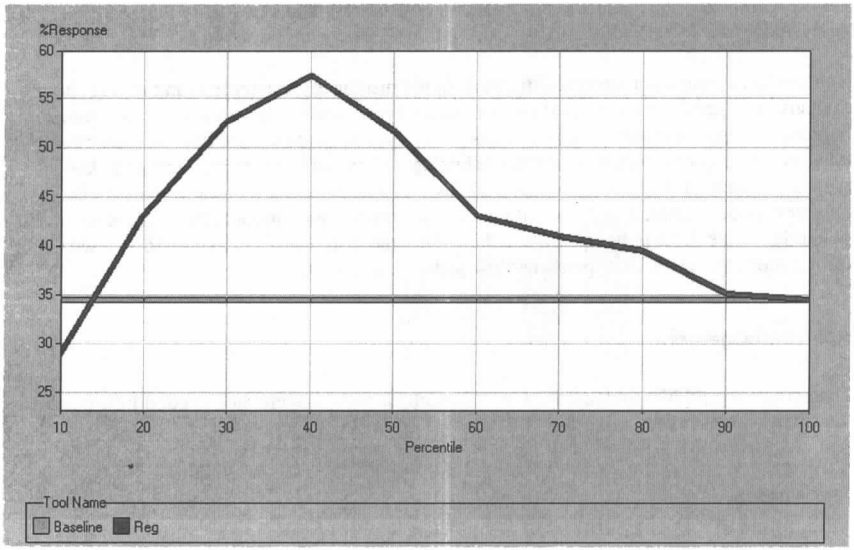


The scores are ordered by decreasing absolute value in the chart. The color density legend indicates the size of the score for a bar. The legend also displays the minimum and maximum score to the left and right of the legend, respectively. The vertical axis represents the absolute value for the effect. In this example, the first variable, Age has the largest absolute value, Credit Score has the second largest absolute value, and so on. The estimates for Location and Education are positive, so their bar values is colored a shade of orange. The estimates for Age and Credit Score have negative values, so their bars are displayed in yellow.

Assessment is the final part of the data mining process. The Assessment criterion is a comparison of the expected to actual profits or losses obtained from model results. This criterion enables you to make cross-model comparisons and assessments, independent of all other factors (such as sample size, modeling node, and so on).

Figure 7 is a cumulative % claim-occurrence lift chart for the logistic regression model. Lift charts show the percent of captured claim-occurrence (a.k.a. the lift value) on the vertical axis. In this chart the target drivers are sorted from left to right by individuals most likely to have an accident, as predicted by each model. The sorted group is lumped into ten percentiles along the X-axis; the left-most percentile is the 10% of the target predicted most likely to have an accident. The vertical axis represents the predicted cumulative % claim-occurrence if the driver from that percentile on down submitted a claim.

Figure 7. Lift Chart for Logistic Regression



The lift chart displays the cumulative % claim-occurrence value for a random baseline model, which represents the claim rate if you chose a driver at random, given the logistic regression model.

The performance quality of a model is demonstrated by the degree the lift chart curve pushes upward and to the left. For this example, the logistic regression model captured about 30% of the drivers in the 10th percentile. The logistic regression model does have better predictive power from about the 20th to the 80th percentiles. At about the 90th percentile, the cumulative % claim-occurrence values for the predictive model are about the same as the random baseline model.

5. Conclusions

This paper introduced the data mining approach to modeling insurance risk and some implementation of the approach. In this paper, we provide an overview of data mining operations and techniques and demonstrate two potential applications to property/casualty actuarial practice. In section 3.2, we used k-means clustering to better describe a group of drivers by segmentation. In section 4.2, we examined several risk factors for automobile drivers with the goal of predicting their claim frequency. The influences and the correlations of these factors on auto claim distribution were identified with exploratory data analysis and decision tree algorithm. Logistic regression is then applied to model claim frequency.

Due to our use of synthetic data, however, the examples show limited advantages of DM over traditional actuarial analysis. The great significance of the data mining, however, can only be shown with huge, messy databases. Issues on how to improve data quality through data acquisition, data integration, and data exploration will to be discussed in the future study.

The key to gaining a competitive advantage in the insurance industry is found in recognizing that customer databases, if properly managed, analyzed, and exploited, are unique, valuable corporate assets. Insurance firms can unlock the intelligence contained in their customer databases through modern data mining technology. Data mining uses predictive modeling, database segmentation, market basket analysis, and combinations thereof to more quickly answer crucial business questions with greater accuracy. New products can be developed and marketing strategies can be implemented enabling the insurance firm to transform a wealth of information into a wealth of predictability, stability, and profits.

Acknowledgement

The authors would like to thank the CAS Committee on Management Data and Information for reviewing this paper and providing many constructive suggestions.

References

Adya, M. 1998. "How Effective Are Neural Networks at Forecasting and Prediction? A Review and Evaluation." *Journal of Forecasting* 17(5-6, Sep-Nov): 481-495.

Berry, M. A., AND G. S. Linoff. 2000. *Mastering Data Mining*. New York, N.Y.: Wiley.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*. New York: Oxford University Press.

Breiman, L., J. H. Friedman, R. A. Olshen, AND C. J. Stone. 1984. *Classification and Regression Trees*. New York, N.Y.: Chapman & Hall.

Borok, L.S. 1997. "Data mining: Sophisticated forms of managed care modeling through artificial intelligence." *Journal of Health Care Finance*. 23(3), 20-36.

Carpenter, G., AND S. Grossberg. 1988. "The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network." *IEEE Computer*, 21(3): 77-88.

Cheesman, P. 1996. "Bayesian Classification (AutoClass): Theory and Results," in *Advances in Knowledge Discovery and Data Mining*, ed. by Fayyad, U., G. Piatetsky-Shapiro, P. Smyth, AND R. Uthurusamy. Menlo Park, CA: AAAI Press/The MIT Press: 153-180.

Chessman, P., J. Kelly, M. Self, J. Stutz, W. Taylor, AND D. Freeman. 1988. "Auto Class: A Bayesian classification system." *5th Int'l Conf. on Machine Learning*, Morgan Kaufman.

Goldberg, D.E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning.* Morgan Kaufmann.

Guha, S., R. Rastogi, AND K. Shim K. 1998. "CURE: An Efficient Clustering Algorithm for Large Databases." *Proceedings of the ACM SIGMOD Conference.*

Ester, M., H. Kriegel, J. Sander, AND X. Xu. 1998. "Cluttering for Mining in Large Spatial Databases." *Special Issue on Data mining, KI-Journal,* 1. Scien Tec Publishing.

Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, AND R. Uthurusamy(Eds.). 1996. *Advances in Knowledge Discovery and Data Mining.* Cambridge, MA: The MIT Press.

Fisher, D., M. Pazzani, AND P. Langley. 1991. *Concept Formation: Knowledge and Experience in Unsupervised Learning.* San Mateo, CA: Kaufmann.

Francis, L. 2001. "Neural Networks Demystified." *CAS Forum* 253-319.

Hand, D., H. Mannila, AND P. Smyth. 2001. *Principles of Data Mining.* Cambridge, Massachusetts: MIT Press.

Hinneburg, A., AND D.A. Keim. 1998. "An Efficient Approach to Clustering in Large Multimedia Databases with Noise." *Proceeding 4th Int. Conf. on Knowledge Discovery and Data Mining.*

Hosmer, D. W., AND S. Lemeshow. 1989. *Applied Logistic Regression.* New York, N. Y.: John Wiley & Sons.

Kass, G. V. 1980. *"An Exploratory Technique for Investigating Large Quantities of Categorical Data,"* Applied Statistics, 29:119-127.

Kleinbaum, D. G., L. Kupper, AND K. Muller. 1988. *Applied Regression Analysis and other Multivariable Methods, 2nd edition.* PWS-KENT Publishing Company, Boston.

Michaelewicz, Z. 1994. *Genetic Algorithms + Data Structures = Evolution Programs.* New York: Springer-Verlag.

Quinlin, J.R. 1983. "Induction of Decision Trees." *Machine Learning* 1: 81-106.

D.E. Rumelhart, G.E. Hinton, AND R.J. Williams. 1986. "Learning Internal Representation by Error Propagation." *Parallel Distributed Processing,* ed. by Rumelhart, D.E., J.L. McClelland, AND the PDP Research Group. Cambridge, MA: The MIT Press: 318-362.

SAS Institute. 2000. *Enterprise Miner,* Cary, N.C.: SAS Institute.

Tufte, E.R. 1983. *The Visual Display of Quantitative Information,* Graphics Press, Cheshire, CN.

Tufte, E.R. 1990. *Envisioning Information,* Graphics Press, Cheshire, CN.