

MineSet™ 2.5 Tutorial

Document Number 007-3573-003

CONTRIBUTORS

Written by Helen Vanderberg and Pam Sogard

Illustrated by Dany Galgani

Edited by Christina Cary

Production by Heather Hermstad

Engineering contributions by Dan Sommerfield, Barry Becker, Eric Eros, John Hawkes, Ron Kohavi, Clay Kunz, Brett Zane-Ulman, and the MineSet team.

St. Peter's Basilica image courtesy of ENEL SpA and InfoByte SpA. Disk Thrower image courtesy of Xavier Berenguer, Animatica.

© 1998 Silicon Graphics, Inc.— All Rights Reserved

The contents of this document may not be copied or duplicated in any form, in whole or in part, without the prior written permission of Silicon Graphics, Inc.

RESTRICTED RIGHTS LEGEND

Use, duplication, or disclosure of the technical data contained in this document by the Government is subject to restrictions as set forth in subdivision (c) (1) (ii) of the Rights in Technical Data and Computer Software clause at DFARS 52.227-7013 and/or in similar or successor clauses in the FAR, or in the DOD or NASA FAR Supplement. Unpublished rights reserved under the Copyright Laws of the United States. Contractor/manufacturer is Silicon Graphics, Inc., 2011 N. Shoreline Blvd., Mountain View, CA 94043-1389.

Silicon Graphics and the Silicon Graphics logo are registered trademarks, and MineSet is a trademark, of Silicon Graphics, Inc. Oracle is a registered trademark of Oracle Corporation. Sybase is a registered trademark of Sybase, Inc. Informix is a registered trademark of Informix Software, Inc. DBMS/COPY is a trademark of Conceptual Software, Inc.

The Tree Visualizer is patented under United States Patents No. 5,528,735 and 5,555,354 and 5,671,381.

Contents

| | | |
|-----------|---|-----------|
| | About This Tutorial | v |
| | Audience for This Tutorial | v |
| | Prerequisites for This Tutorial | v |
| | Structure of This Tutorial | vi |
| | Typographical Conventions | vi |
| 1. | Data Mining Fundamentals | 1 |
| | About Data Mining | 1 |
| | Data Mining Methods | 2 |
| | Analytical Data Mining Algorithms | 3 |
| | Supervised Modeling | 3 |
| | Unsupervised Modeling | 5 |
| | Data Visualization | 6 |
| | MineSet Tools for Data Mining Tasks | 6 |
| 2. | Data Mining Process | 9 |
| | Identifying the Data | 9 |
| | Preparing the Data | 10 |
| | Transforming the Data | 11 |
| | Building a Model | 12 |
| | Evaluating a Model | 12 |
| | Deploying a Model | 12 |
| | Applying the Process to a Specific Database | 12 |
| 3. | Churn Tutorial | 13 |
| | About the Raw Data | 13 |
| | Starting MineSet | 14 |
| | Viewing the Records | 16 |
| | Building an Evidence Classifier | 18 |

- Viewing Probabilities With Splat Visualizer 21
- Visualizing Geographic Distributions 24
- Creating a Decision Tree Classifier 29
- 4. Further Explorations 33**
 - Exploring Data Clusters 33
 - Relating the Columns and Axes in the Model 35
 - Finding Important Columns in the Clustered Model 37
 - Mapping to Scatter Visualizer 38
 - Invoking a Decision Table 40
 - Targeting Customers Using a Classifier 42
 - Creating a Training Sample 43
 - Applying a Model 44
 - Reducing Misclassification Costs 49
 - Displaying a Confusion Matrix 49
 - Defining a Loss Matrix 52
 - Viewing a Return on Investment Curve 53
 - Further Exploration of MineSet 55

About This Tutorial

The *MineSet Tutorial* introduces MineSet, an integrated suite of data mining and visualization tools, and provides a swift survey of the concepts and processes of data mining. This tutorial describes a few basic tasks to help you use MineSet immediately. Once you are familiar with the interface, refer to the *MineSet User's Guide* for a full description of other MineSet features. The Guide is delivered online as part of the MineSet product. See also <http://mineset.sgi.com> for more information.

Audience for This Tutorial

This tutorial is for end users. No experience in programming is required, nor is any previous knowledge of statistics, although a basic knowledge of UNIX is assumed.

Prerequisites for This Tutorial

To work with this tutorial, MineSet should be installed on your system, or you should have access to such a system. The examples depend on it. Instructions for installing MineSet are available in the *MineSet User's Guide* and on the MineSet Web page <http://mineset.sgi.com>, where MineSet itself can be downloaded for evaluation purposes.

For this tutorial you do not need access to a database. The data needed is included in the MineSet distribution.

Structure of This Tutorial

Chapter 1, “Data Mining Fundamentals,” introduces the concept of data mining and explains how it can be used to solve problems. Common data mining tasks are aligned with the various MineSet tools, although details are covered in later chapters.

Chapter 2, “Data Mining Process,” describes the tasks involved in the process of data mining. A case study of data mining using MineSet is provided.

Chapter 3, “Churn Tutorial,” provides a detailed tutorial for the process of data mining using MineSet. It begins from the initial screen and steps screen by screen through using MineSet tools on churn, a dataset provided with the MineSet distribution.

Chapter 4, “Further Explorations,” continues the exploration of MineSet with more complex variations of exploring data mining.

Typographical Conventions

This tutorial uses several font conventions:

- | | |
|----------------------|---|
| <i>italics</i> | Italics are used for command and reference page names, filenames, variables, hostnames, user IDs, and the first use of new terms. |
| <code>Courier</code> | Courier is used for examples of system output and for the contents of files. |
| Courier bold | Courier bold is used for commands and other text that you are to type literally. |

Data Mining Fundamentals

This chapter surveys data mining methods, model building and assessment, and the role of MineSet in connection with these topics:

- “About Data Mining” on page 1
- “Data Mining Methods” on page 2
- “Analytical Data Mining Algorithms” on page 3
- “Data Visualization” on page 6
- “MineSet Tools for Data Mining Tasks” on page 6

About Data Mining

The purpose of data mining is to discover patterns in data so that this knowledge can be applied to problem solving. Analytical data mining integrated with powerful visualizations present new pathways to knowledge discovery. The data mining system can automatically find and show you new patterns that will lead to fresh insight. Examples of this might be determining correlations among attributes, discriminating among subsets of the data with differing characteristics, and inferring probabilities of future events from historical data.

In ordinary database queries or online analytic processing (OLAP), the user must specify directly any relationships between data elements. Data mining can discover relationships that may be unknown or unseen by the user.

Data to be analyzed, or mined, is often initially retrieved when a business or scientific process is performed, such as acquiring data from customer billing, pharmaceutical testing, or point-of-sale transactions. The amount of data retrieved may be so large as to preclude analysis by means other than data mining. Such data, once properly transformed, is often stored in a data warehouse. See “Preparing the Data” on page 10 for further details.

Data Mining Methods

Data mining combines hypothesis testing and data-driven discovery. In hypothesis testing, the investigator tests an idea against a body of data to confirm or reject its validity. In some cases the data itself may drive discovery. In discovery, the investigator draws conclusions from the data, allowing the data itself to suggest conclusions. Often data mining problems are resolved by employing a blend of both methods. For example, conclusions may give rise to new hypotheses that can be tested, and confirmed or rejected. Data mining is where statistics and machine learning converge.

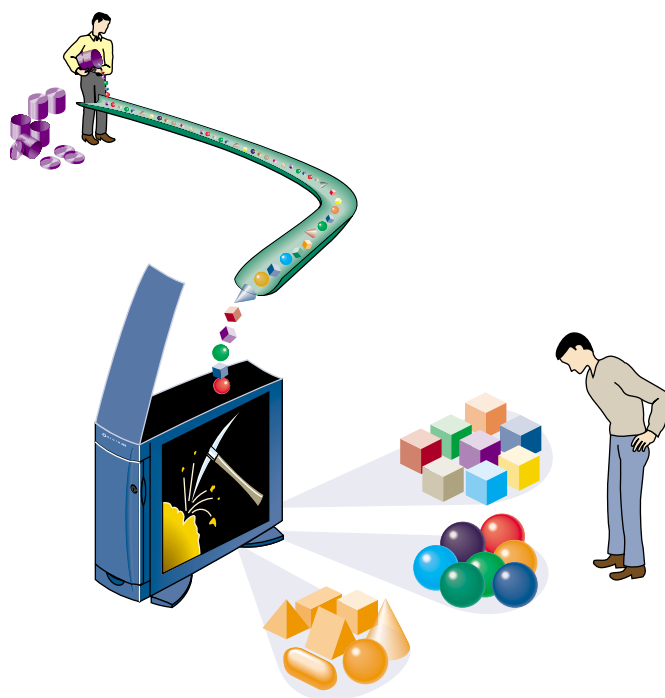


Figure 1-1 Analytical Data Mining Discovers Patterns in Data

The MineSet suite of tools lets you analyze, mine, and graphically display data so that you can visualize, explore, and understand your data. You can organize and examine your data in different ways. The mining tools automatically find patterns and build models that can be viewed using the visualization tools. When you apply the visualization tools directly to the data, you gain a deeper, intuitive understanding of your data, often discovering hidden patterns and important trends.

MineSet tools provide an interactive, three-dimensional (3D) visual interface that lets you manipulate visual objects on the screen as well as perform animations. This ability to visualize and survey complex data patterns can prove invaluable in making decisions.

The results of a typical analytical data mining operation in MineSet include both a model describing the data and a visualization of the model. The visualization allows you to understand the model, thus leading to greater insight. MineSet is an integrated system in which the analytical algorithms can generate the visualization, and users can select visualization elements for further mining.

Analytical Data Mining Algorithms

Analytical data mining algorithms automatically build models from the data. Two families of modeling algorithms are commonly used—supervised and unsupervised. Predictive modeling tasks, where the goal is to predict the value of one column based on the values of other columns, are called *supervised* tasks. These tasks are similar to the supervision of a teacher who gives you the correct answer for the question, to teach you.

The goal in descriptive modeling is to discover patterns and segments of the data. These are *unsupervised* tasks. There is no notion of a correct answer, nor any obvious agreed-upon measure of performance. Unsupervised tasks provide insight to the data as a whole by showing patterns and segments that behave similarly.

In the following discussion, the term *attribute*, as it applies to analytical data mining, may be thought of as a column.

Supervised Modeling

In supervised modeling, there is a special attribute called the “label” that you intend to predict. By encoding the relation between the label and the other attributes, the model can make predictions about new, unlabeled data. In addition, by visualizing the model itself, you can gain insight into the relationship between labels and other attributes. For example, if a customer has left your company (typically called attrition or churn), you can build a model that will not only predict which customers are likely to churn, but also help you understand the reasons and patterns that lead to this behavior.

The two most common supervised modeling tasks are called classification and regression. If the label is discrete (that is, containing a fixed set of values), the task is called classification; if the label is a continuous value (that is, can take a value in a continuous range—for example, income, or stock price), the task is called regression.

Classification

Classification is the task of assigning a discrete label value to an unlabeled record. In doing so, records are divided into predefined groups. For example, a simple classification might group customer billing records into two specific classes: those who pay their bills within 60 days, and those who take longer than 60 days to pay. Further data classification examples might divide customers by sex or income. Classifiers can also predict the probability that the label will take on a specific value. For example, the probability that the person will pay their bill within 60 days can be computed.

A classifier is a model that predicts one attribute of a set of data when given other attributes. MineSet can induce (build) a classifier automatically from a training set. When a classifier is induced, MineSet also generates a visualization of the model that can help you understand how the classifier operates, thus providing valuable insight. Once a classifier is generated, it can be used to classify or predict class probabilities for unlabeled records (that is, for records that are missing the label attribute). This concept is explained further in Chapter 3.

MineSet has inducers for four classification models: Decision Trees, Option Trees, Evidence (Simple Bayes) and Decision Table Classifiers. Each model can be viewed using a visualizer: the Decision Tree models and Options Tree models can be viewed using the Tree Visualizer, the Evidence model can be viewed using the Evidence Visualizer, and Decision Tables can be viewed using the Decision Table Visualizer.

Regression

Regression is a supervised modeling task similar to classification, except that the label is not discrete. For example, predicting salary or the price of a stock is a regression, whereas predicting whether the salary is in a given range or whether a stock will go up or down is a classification task.

Assessing the Accuracy of Models

Predictive models are rarely perfect, therefore estimating their accuracy is an important part of the data mining process. The tool used to measure accuracy depends upon the model type. Classifiers are usually evaluated according to their error rate. The most common such measure is misclassification, or proportion of misclassified records. When assessing the accuracy of a model, it is important to test it on data that was not used in building the model. MineSet provides a number of methods for evaluating errors. See Chapter 4, “Further Explorations,” for details.

Unsupervised Modeling

In unsupervised modeling, the aim is to discover rules and segments of the data that behave similarly (clusters). Unsupervised modeling is a descriptive task, not a predictive task. The models cannot be used directly to make predictions, hence it is not necessary to set aside part of the data as a training set from which to build the classifier. The two most common unsupervised modeling tasks are associations and clustering.

Associations

To generate associations, the task is to determine rules of implication between data attributes so that A implies B. Associations are used to find affinity groupings that discover what items are usually purchased with others. The classic affinity grouping is market basket analysis, predicting the frequency with which certain items are purchased at the same time. For example, discovering that baby food implies a higher probability that a customer will buy low-tar cigarettes rather than regular cigarettes might help stores arrange their shelves differently. Associations can be viewed in MineSet using the Rules Visualizer.

Clustering

Clustering algorithms segment the data into groups of records, or clusters, that have similar characteristics. For instance, a health-insurance company may discover that these characteristics define a segment: 20-to-45 years old, technical worker, fewer than two children, television science-fiction fan, and a disposable income of \$5000 to \$10,000 per year.

The segment can then be targeted more effectively with a health insurance package well-suited for these people, by using television ads in new science-fiction episodes.

Data Visualization

An analytical data mining algorithm can be complemented with data visualization techniques taking advantage of the human brain's amazing pattern recognition capability. The following MineSet visualizers are available:

- Map Visualizer—Data is displayed on a map, commonly a geographical map.
- Scatter Visualizer—Data points are shown in one-, two-, or three-dimensions. Additional attributes can be mapped to color, size, and shape. Finally, two additional attributes may be mapped to sliders, allowing animation and fly-throughs, for a total of eight dimensions. The column importance operation in MineSet can help you identify the important dimensions to map for a given task.
- Splat Visualizer—Similar to Scatter Visualizer, with the distinction that data density is shown by opacity of color, which appears as a blurred translucent cloud. The result approximates the effect of rendering each data point individually.
- Tree Visualizer—Data is mapped to nodes in order to see hierarchical breakdowns of the data. Decision Trees, and Options Trees all show data in a variety of branching tree-like visualizations.

MineSet Tools for Data Mining Tasks

If you have data mining problems requiring classification, regression, and clustering, you will find these MineSet tools useful:

- Decision Tree Inducer and Classifier—Induces a classifier resulting in a decision tree visualization.
- Option Tree Inducer and Classifier—Induces a classifier similar to a decision tree inducer and classifier. However, it builds alternative options and averages them during classification, usually leading to improved accuracy.
- Evidence Inducer and Classifier—Creates its own classifier and produces a visualization to display evidence based on the data provided.
- Decision Table Inducer and Classifier—Creates a hierarchical visualization displaying pairs of dimensions at every level. You can drill-up and drill-down quickly, while maintaining context.

- Clustering Algorithm—Groups data according to similarity of characteristics, then displays it as a series of box plots and histograms, similar to the Statistics Visualizer. The clustering algorithm displays results using the Cluster Visualizer by default, but other visual tools may be used as an alternative.
- Regression Tree—Induces a regressor that predicts a real value, that is, results with gradations of value rather than specific predetermined limits.
- Column Importance—Determines the importance of specific columns in discriminating one label value from another. Used to observe the varying effects of changing variables, or to suggest columns to map to the axes of the Scatter and Splat Visualizers.

MineSet contains additional tools to aid the knowledge discovery process:

- Statistics Visualizer—Data is displayed in the form of box plots and histograms, one per column. Continuous columns are shown as box plots, discrete columns are shown as histograms.
- Record Viewer—The original data is displayed as a spreadsheet.

The next chapter, Chapter 2, describes a typical data mining process and how the tools are used.

Data Mining Process

This chapter introduces the specific tasks involved in the knowledge discovery process. The process is iterative, commonly going back to earlier stages once you discover new patterns and improve your understanding of the data, as shown in Figure 2-1.

This chapter describes a process that follows these steps:

1. Identify the source of the data—expanded in “Identifying the Data” on page 9.
2. Prepare the data—expanded in “Preparing the Data” on page 10.
3. Build a model—expanded in “Building a Model” on page 12.
4. Evaluate the model—expanded in “Evaluating a Model” on page 12.
5. Deploy the model—expanded in “Deploying a Model” on page 12.

Identifying the Data

The task of identifying the data begins by deciding what data is needed to solve a problem. For example, predictability about customer behavior is often a necessary goal, recast in terms of a problem. In defining the problem, the investigator must identify the data needed to solve that problem and explore other possible sources of data.

Data may be in a difficult location or in an obscure form. Sometimes there are several initial databases that may be incompatible with each other. Further, if data is scanty or incomplete, more data may be needed. The form in which new data is to be collected depends on the form of existing data. Finally, the data may exist but need to be extracted from a central data warehouse. MineSet accepts both flat files and binary files as well as data from several commercial database vendors such as Informix, Oracle, and Sybase. Tools such as DBMS/COPY from Conceptual Software, Inc. allow you to convert data from over 100 formats to MineSet.



Figure 2-1 Data Mining Process

Preparing the Data

Data may have to be loaded from legacy systems or external sources, stored, and cleaned. Specifically, the following problems are common:

- Data may be in a format incompatible with its end use (for example, EBCDIC format).
- Data may have many missing, incomplete, or erroneous values.
- Field descriptions may be unclear or confusing, or may mean different things depending on the source. For example, order date may mean the date that the order was sent, postmarked, received, or keyed in.
- Data may be stale. Customers may have moved, changed households, or changed spending patterns.

MineSet can help you discern data quality problems in the initial stages of building a data warehouse.

In spite of being clean, data may need to be transformed before it is suitable for mining and visualization. Specifically, the input to the algorithms and visualizations must be a single table. While SQL commands can be given to MineSet, it is recommended that database administrators create the appropriate views to simplify operations for end users. While MineSet can perform powerful data operations, in some cases certain transformations need to be done prior to using MineSet.

Transforming the Data

Considerable planning and knowledge of your data should go into data transformation decisions. Data transformations are at the heart of developing a sound model. You may even need to go back and transform the data differently:

- By adding columns, usually applying a mathematical formula to existing data.
- By removing columns which are not pertinent, are redundant, or contain obvious, uninteresting predictors.
- By filtering visualizations. For example, you may want to see only the strongest rules or the most profitable customer segments.
- By changing a column's name.
- By binning data—breaking up a continuous range of data into discrete segments.
- By aggregating data—grouping records together, and finding the sum, maximum, minimum, or average values.
- By sampling the data to get a random subset of the data (by percentage or count).
- By applying a classifier that you have previously created, to label new records with a class label, or to estimate the probability of a given label value.

In MineSet, most of these transformations take place using the Data Transformation pane in Tool Manager.

Building a Model

At the core of the knowledge discovery process is model building, automatically done by analytical data mining algorithms. This is clarified in Chapter 3.

Evaluating a Model

Evaluating the accuracy of a model refines your understanding of that model and its usefulness. Some models, notably the Decision Tree classifier and the Option Tree classifier, evaluate different parts of the model and display these models directly through visualization.

MineSet implements four model assessment methods: error estimation, confusion matrix, lift curve and ROI (return-on-investment) curve.

Deploying a Model

A model can be deployed by applying it to new data. New data can give rise to further questions, which may require further refinements.

In the telecommunications example in Chapter 3, a model can be created to determine which customers are likely to leave their phone carrier. Customer records can then be evaluated through the model to identify the specific customers most likely to leave. These customers can be given incentives to stay.

Applying the Process to a Specific Database

The next two chapters step you through the knowledge discovery process on the churn dataset—a prepared dataset of telecommunication customers. As you work through the examples, think of the process presented here and how your operations progress forward and loop back as shown in Figure 2-1.

Churn Tutorial

This chapter steps you through a possible knowledge discovery process using the churn dataset provided with MineSet. It is assumed that MineSet is installed on the system you use, together with all the sample files. Each step is explained in detail. Unless otherwise noted, each step builds on the step before. These steps are:

- “Starting MineSet” on page 14
- “Viewing the Records” on page 16
- “Building an Evidence Classifier” on page 18
- “Viewing Probabilities With Splat Visualizer” on page 21
- “Visualizing Geographic Distributions” on page 24
- “Creating a Decision Tree Classifier” on page 29

About the Raw Data

The churn dataset deals with telecommunications customers—people who use the phone regularly. Customers have a choice of carriers, or companies providing them with telephone service. When these customers change carriers they are said to “churn,” which results in a loss of revenue for the previous carrier. A telecommunications company is likely to have a database of call records containing call information (source, destination, date, duration), a billing database, a customer database, and a customer service database. Relevant information about the customer appears in all these databases. This information, when combined, yields a set of customer signatures. The churn dataset provided with MineSet is such a set; the step of identifying the data and creating customer signatures into records has already been done. This dataset, which is used in the rest of the chapter, contains one record per customer.

Starting MineSet

1. Start MineSet by typing in a shell window:
mineset
or double-click the MineSet icon on your desktop. If you have previously executed MineSet, you may be presented with a restored session. If this happens, cancel the initial dialog box shown in Figure 3-1 and continue as described below.
2. Choose File > Connect to Server. At Server Name: enter **localhost** or the name of the server on which MineSet is installed. At Login name: enter your user name and password. Click OK.

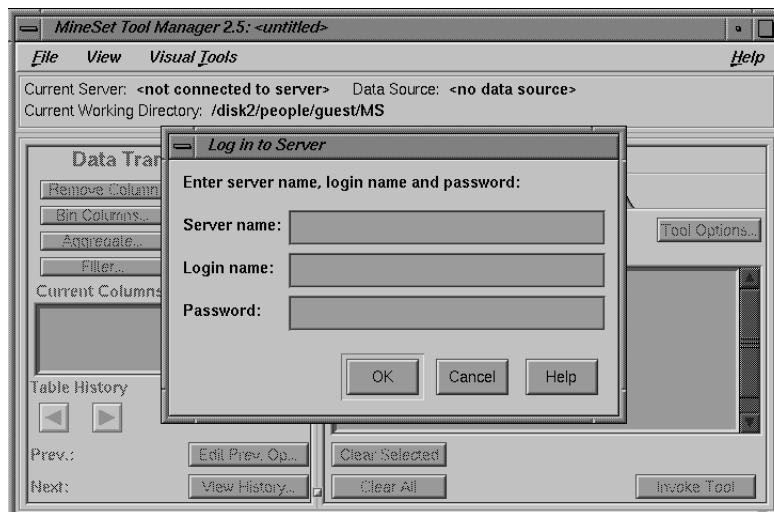


Figure 3-1 Tool Manager Log In Window

3. In the Tool Manager window, choose File > Open New Data File, change the directory pathname to `/usr/lib/mineset/data/`, and select `churn.schema`. A series of entries appears in the right-hand Preview Columns pane as shown in Figure 3-2. Click OK.

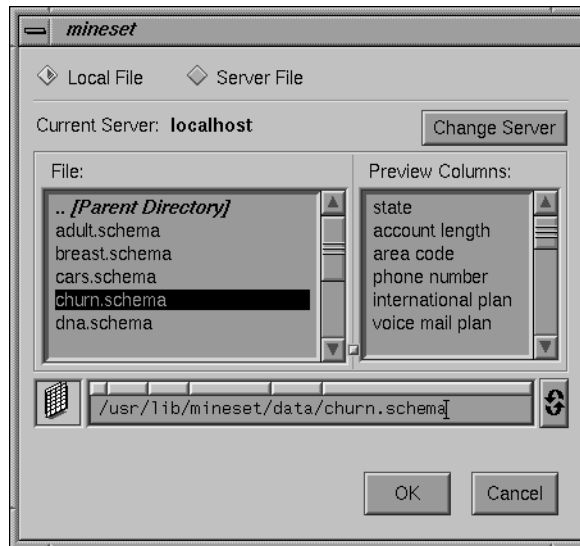


Figure 3-2 Open New Data File Window

This gives you access to a dataset of telecommunications customers. The next time you run MineSet, you are automatically returned to this position, and any option selections you make are saved.

Viewing the Records

You can see the records in spreadsheet form, after bringing up MineSet Tool Manager, by following these steps:

1. In the Data Destination pane of the Tool Manager window, click the Viz Tools tab; from the Tool popup menu, choose Record Viewer.
2. Click *Invoke Tool* at the lower right.

This churn dataset is used in the rest of the chapter and contains one record per customer. The Data Transformations pane on the left side of Tool Manager lists columns with their type: state (string), account length (double), and so forth. Columns are defined as double or float if they are numbers, or string if they are made up of characters.

The data appears as a spreadsheet. Some columns and their meanings are shown in Table 3-1.

Table 3-1 Details of Columns Shown in MineSet Record Viewer

| Column name | Value |
|-------------------------------|--|
| state | Two-letter abbreviation for the customer’s U.S. state of residence |
| account length | Numerical value indicating the number of months the customer has been with the long-distance carrier |
| area code | Typical three-digit telephone company designations |
| phone number | Typical three+four-digit telephone company designations |
| international plan | Special pricing package for international calls, expressed as a yes/no value |
| voice mail plan | Special pricing package for customers with voice mail provided by the carrier, expressed as a yes/no value |
| number of voice mail messages | Average number of voice mail messages per day |
| total day minutes | Number of minutes charged at the carrier’s day rate |
| number customer service calls | Number of calls this customer made for assistance to carrier customer support in the last six months |
| churned | Whether this customer changed long-distance carriers in the last six months, expressed as a yes/no value |

3. Choose File > Exit to close the Record Viewer.

You should see the Tool Manager window again, still using the churn data source.

4. In the Data Destination pane of the Tool Manager window, the Viz Tools tab is still displayed; from the Tool popup menu, choose Statistics Visualizer.
5. Click the *Invoke Tool* button.

A display appears consisting of a number of box plots and histograms. The box plots show summary statistics for continuous variables, the histograms show the distribution of values for discrete variables.

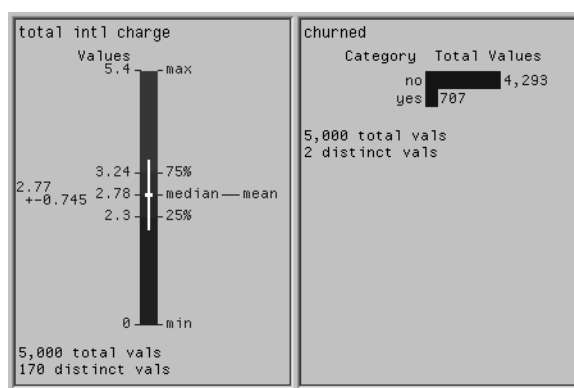


Figure 3-3 Representative Box Plot and Histogram Produced by Statistics Visualizer

Each box plot (on the left in Figure 3-3) shows statistics about data from a single column, including the minimum, maximum, mean, median, and two out of four quartiles (25th and 75th percentiles). These values are marked as lines, and the standard deviation is shown after the +/- sign.

The mean is the number found by adding the data in a column, then dividing by the number of records. The median is middle number when numbers in a given column are arranged in order of size. The standard deviation is a measure of the dispersion of the data in a column.

The histograms consist of specific discrete values: state names or yes/no values. Scroll down to find the churn histogram in the lower right of the display (see Figure 3-3, right). Notice the total values and their distribution, which shows that 707 customers out of 5,000 have left the carrier. The churn column is of greatest interest throughout this tutorial.

6. Choose File > Exit in the Statistics Visualizer window to close the window and return to the Tool Manager window.

Building an Evidence Classifier

You are now ready to perform analytical data mining. Verify that MineSet Tool Manager is connected to the appropriate server, and that the data source is `/usr/lib/mineset/data/churn.schema`. If you exited MineSet between sessions, the history file automatically returns to where you left off.

1. In the Data Destination pane of the Tool Manager window, click the Mining Tools tab.
2. Click the Classify tab, and make selections from these popup menus:

Mode: Classifier & Error

Inducer: Evidence

Discrete Label: churned

You are about to induce an evidence classifier to help characterize the customers who are likely to churn. The default mode, Classifier & Error, employs a holdout method on the data, inducing the classifier from two-thirds of the data and leaving the remainder as a test set to estimate the error rate.

3. Click *Go!*

The Status window on the bottom of Tool Manager shows progress and summary information about the induction process, including the estimated error rate of 12%. When the induction step is done, the Evidence Visualizer is automatically invoked, showing the model visually.

In the Evidence pane (on the left of Figure 3-4), rows of charts represent columns in the dataset, each showing the proportion of customers who churn. Switch cursor mode from grasp to pick, and click on the box titled Evidence. The pie charts show probable outcome, just as the cake charts showed evidence. For this tutorial these square cakes and round pies are termed "charts."

Attributes are sorted by discriminating power for the label "churned," starting from the top.

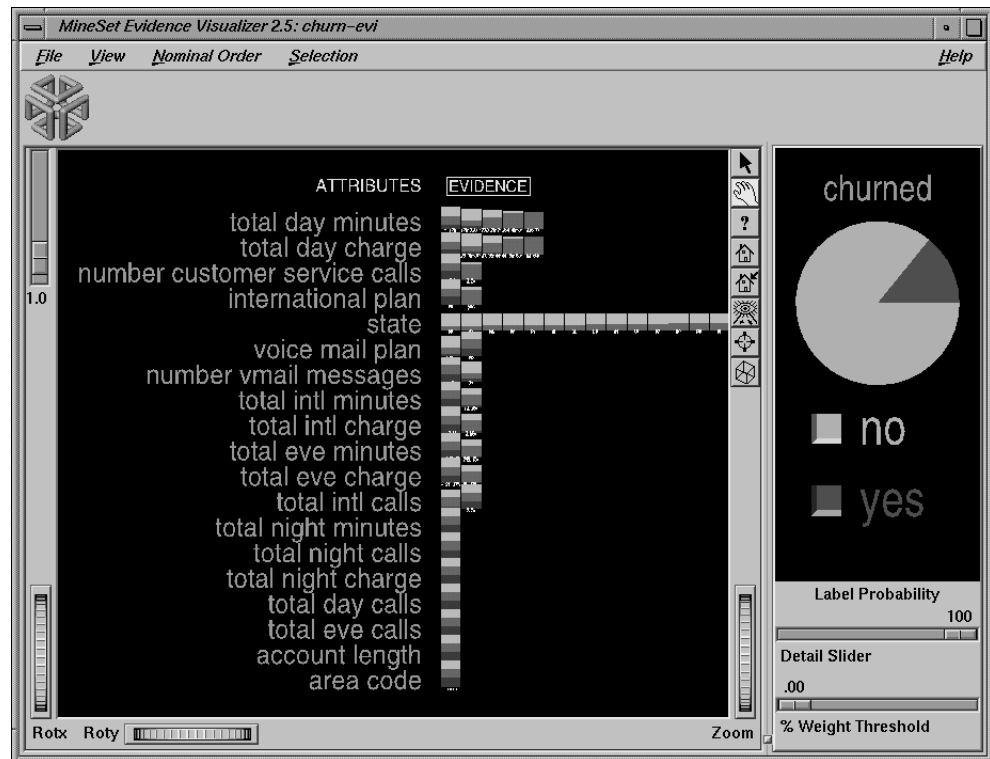


Figure 3-4 Evidence Visualizer Window

In the Label Probability pane on the right of the screen in Figure 3-4, the round pie chart reflects the probability of seeing this label in the data for a randomly chosen record, ignoring all attribute values. Mathematically, this is the number of records with the class label, divided by the total number of records.

Pass the pointer across a chart in the left pane, and the evidence or probability shows above the display window. Click on a chart to update right pane to show the expected probability according to the model.

Use the thumbwheels at the screen's border to dolly back and forth, and tilt either the X or Y axis to get a closer view of any chart. In grasp mode, the middle mouse button moves your view in the display. In pick mode, select multiple charts by holding down the Shift key and pressing the middle mouse button.

You can see the factors that affect churning, because the slice representing churn increases from left to right on the first and second rows, so a serious problem is evident. Customers that use the company's service the most, also churn at a higher rate. The company is not just losing customers; the lost customers are its most valuable.

To find out about a class label (for example, `churned =yes`), select a value in the Label Probability pane on the right. Click on the button near the label `yes` in the right pane, and the evidence is shown as bars. Pointing to bars will show you the estimated probabilities.

The discriminating attributes shown here can be used to choose axes for a scatterplot visualization. The state attribute, shown relatively high on the list of attributes, hints at a possible geographical relationship.

Evidence models show attributes independently; however in many datasets a combination of attributes determines the label. The error estimation shown in the status window, indicates that the classifier is expected to have about a 12% error rate. Later we generate a decision tree that is significantly more accurate. Because total day minutes and total day charge are related, only total day charge is used for the next step. Close the Evidence Visualizer (File > Exit) before moving on to work with the Splat Visualizer.

Viewing Probabilities With Splat Visualizer

The Splat Visualizer requires that the column mapped to color must have a numerical value. The churned column is a string that must be converted to a number (p_churned, indicating the probability of churning), before mapping it in Splat Visualizer. The next procedure shows how this is accomplished.

1. In the Data Destination pane of Tool Manager, select the Viz Tools tab, then choose Splat Visualizer from the Tools popup menu.
2. In the Data Transformations pane of Tool Manager, click Add Column.

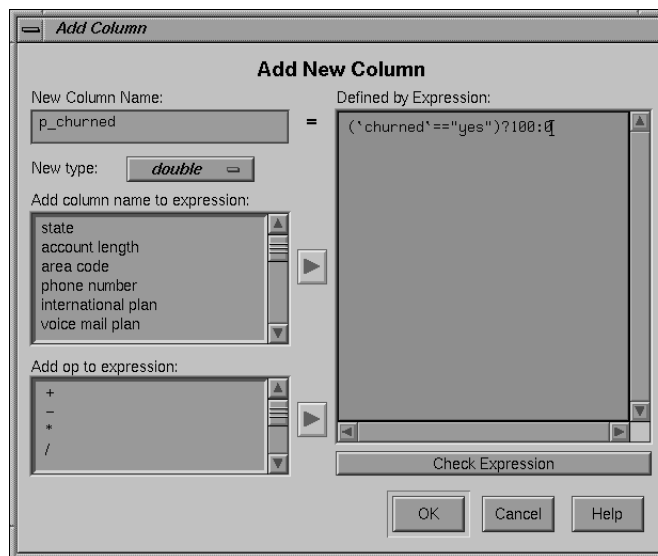


Figure 3-5 Adding a New Column

- From the Add Column dialog box, in the New Column Name text field, enter the new name, `p_churned` (see Figure 3-5). The intention is to make a column of numbers, based on the churned column.

From the Add Column dialog box, in the Defined By Expression text field, create the expression `(`churned`=="yes")? 100:0`. You can create this expression from the building block in the two scrolled lists "Add column name to expression" and "Add op to expression", or you can type it in directly. This expression translates to "for all the values in the churned column that are yes, give them the value of 100, otherwise give them the value of 0." The purpose of this is to translate a string (yes or no) into a numerical value. Verify that the *New type* button is set to `double`.

Click *Check Expression* to ensure there are no syntax errors. Click *OK* to add the column.

- In the Data Transformations pane, map items in Current columns to items in Visual Elements by clicking first the left pane, then the right. For this tutorial, map `total day charge` to Axis 1, `number customer service calls` to Axis 2, `international plan` to Axis 3, and `p_churned` to `color`, shown in Figure 3-6. Only those entities without an asterisk require mapping.

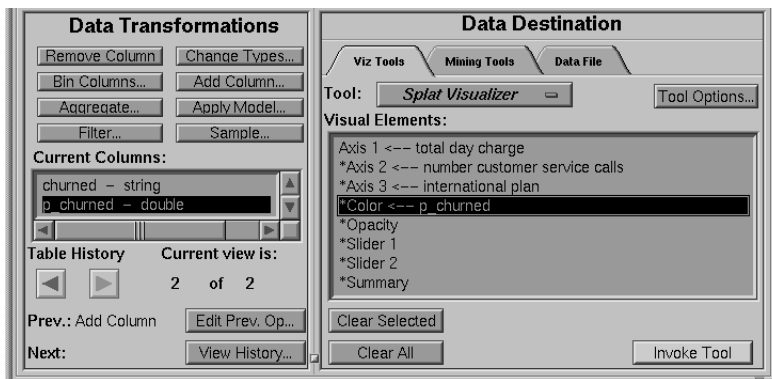


Figure 3-6 Mapping Columns to Elements for Splat Visualizer

5. Click *Invoke Tool*.

The data is plotted on the Splat Visualizer window, shown in Figure 3-7. The slider bar in the upper left varies the color density. Use the question mark in the right toolbar for help on window manipulation. Rotate the splat plot with the grasping hand until any trends stand out. Splat Visualizer allows you analyze complex data by using the varying behavior in several dimensions.

You can save the current state of the Tool Manager including special options by choosing Save Current Session As from the File menu, and specifying *churn1.mineset*. To save a snapshot of the current visualization, choose Save As from the File menu, and specify *churn1-out.rgb*.

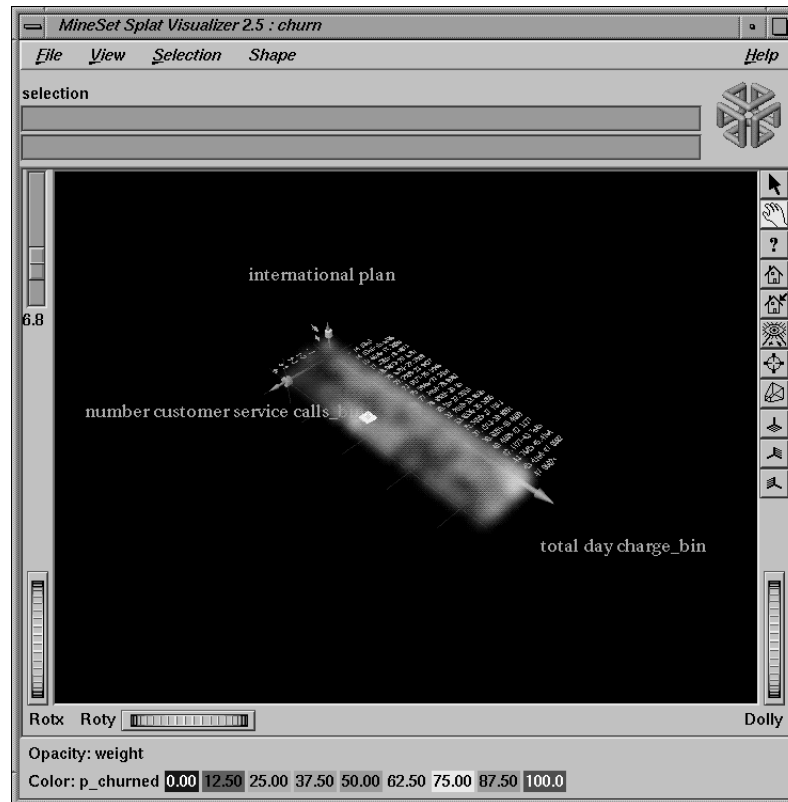


Figure 3-7 Splat Visualizer Window

In the visualization shown in Figure 3-7, the highest probability of churn occurs in two places: in the yellow to red areas when total day charge is high, shown in the bottom of this figure; and when the total day charge is low and customer service calls are high (near the upper left of this figure). Low-paying customers who make many customer service calls leave. These are customers you may not want to stay, because they cost you money and bring little reward. The high-paying customers at the bottom of the figure are a better target.

6. Close Splat Visualizer and return to the Tool Manager Window.

Visualizing Geographic Distributions

As shown in Figure 3-4 on page 19, the Evidence model indicated that state was a powerful discriminating attribute. This section builds on previous computations to display data geographically to illustrate how churn varies by state. You have already added the column (`p_churned`) from existing columns in the dataset.

You can now transform the data into a smaller dataset that contains the average churn per state. Such a transformation is called aggregation.

1. In the Data Transformations pane of the Tool Manager window click *Aggregate*.

In the Aggregate dialog box move `p_churned` into the left column, click *Average* and *Count* on, and turn off *Sum*. Leave `state` in the central column and move all the rest to the right column. (Hold down the Shift key to gather multiple columns.) Make sure your screen looks like Figure 3-8. Click *OK* to apply your choices.

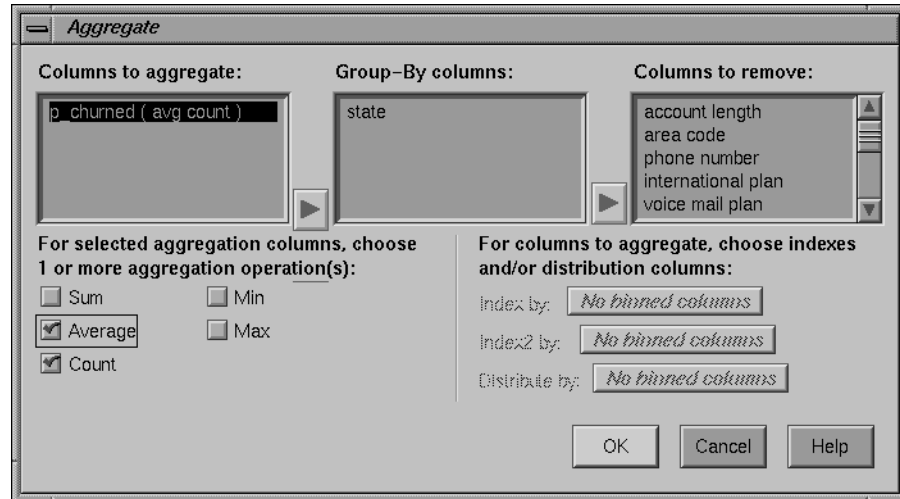


Figure 3-8 Aggregate Dialog Box

2. Click the Viz Tools tab in the Data Destination pane of the Tool Manager window; choose Record Viewer from the Tool popup menu; click *Invoke Tool*. You should see a record for each state with the average churn and number of customers for that state.
3. Close the Record Viewer window and return your focus to the Tool Manager window. You will now link this data onto a map of the United States.
4. Choose Map Visualizer from the Tool popup menu, and click the *Tool Options* button on the Viz Tools tab of the Tool Manager window.

5. Click the *Find File* button to the right of the Entities File text field and select `usa.state.hierarchy`. The status of the dialog box is shown in Figure 3-9. The pathname is `/usr/lib/mineset/mapviz/gfx_files`. Click *OK* to retrieve that file, and *OK* to dismiss the Map Viz Options dialog box.

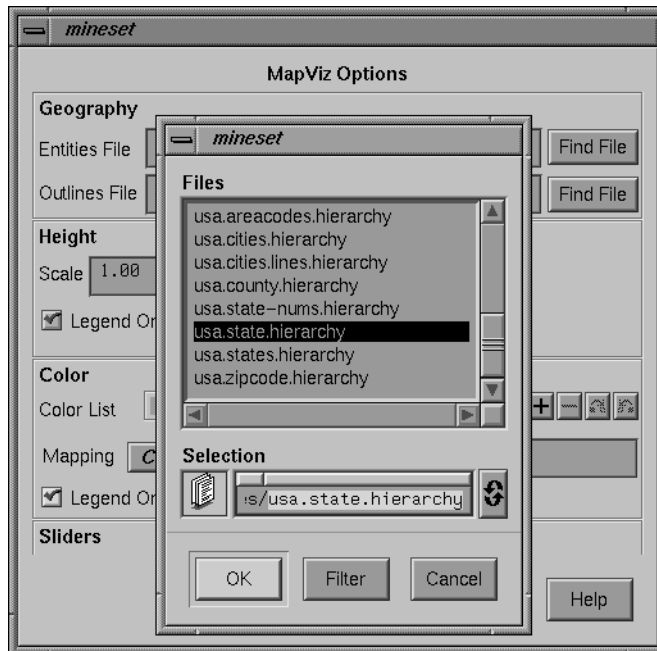


Figure 3-9 Map Viz Options Dialog Box

The next step is to link the visual elements to the columns.

- From the Data Transformations pane of Tool Manager, map the columns to entities in the Data Destination pane. Map items in *Current Columns* to items in *Visual Elements* by clicking on first the left pane, then the right.

Map *state* to *Entity-Bars* and *avg_p_churned* to *Color-Bars*, and *count_p_churned* to *Height-Bars*.

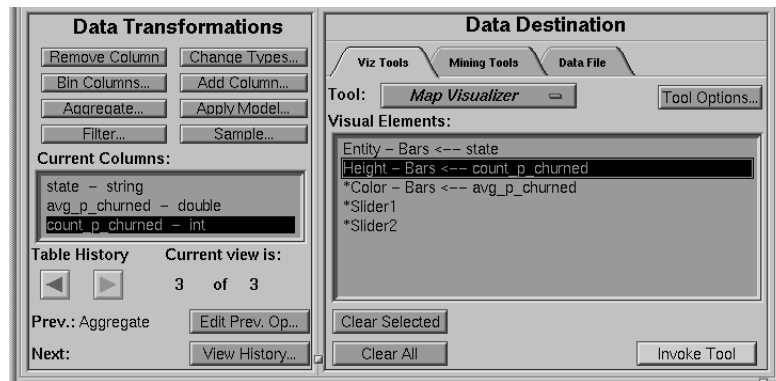


Figure 3-10 Mapping Columns to Visual Entities for Map Visualizer

- Click *Invoke Tool* to view the map distribution of churned customers according to state (Figure 3-11).

The tool shows the distribution of churned customers across the United States. For each state, the color indicates the probability of churn and the height indicates the number of customers in that state. For example, in Figure 3-11, Maine is chosen, showing an average churn rate of 18.4466%, but based on the churn count of 103. In other words, the average is based on only 103 customers. West Virginia shows the greatest height, with a probability of churn based on 158 customers. States showing the clearest, brightest colors calculate an average churn rate over 21% (Texas, Montana, Washington, California, and New Jersey). This visualization indicates that there is no obvious relationship between churn and geography, although different states do have different churn rates.

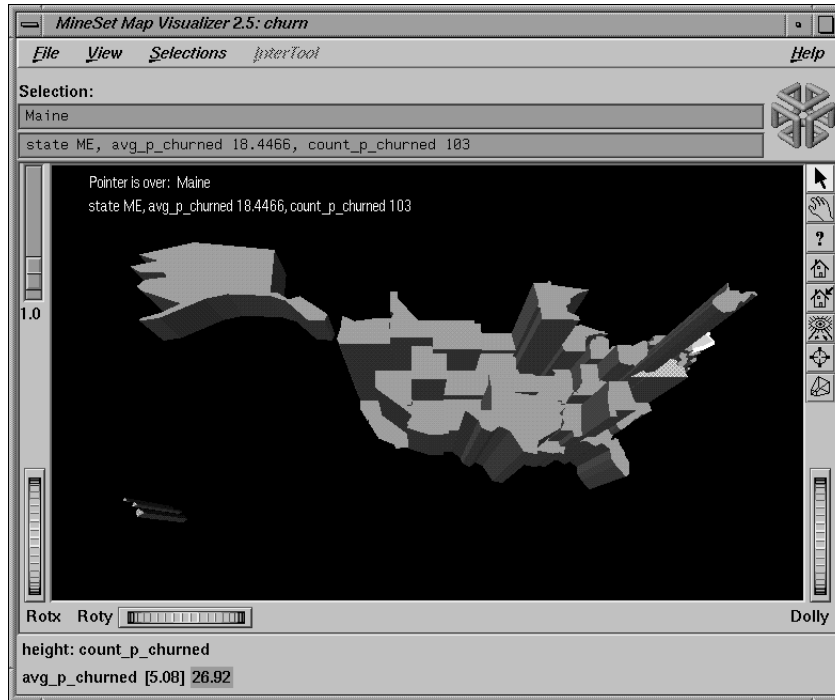


Figure 3-11 Map Visualizer Window With Average Churn Distribution

Close Map Visualizer using File > Exit. The next example explores the Decision Tree classifier, using the same dataset to produce a different visualization.

Creating a Decision Tree Classifier

Unlike the Evidence classifier, the Decision Tree classifier can show attribute interactions, that is, combinations of attribute values that affect the label. For this section, start with a fresh file by selecting File > Open New Data File, and entering `/usr/lib/mineset/data/churn.schema`. You will now build a decision tree classifier and visualize it.

1. Click the Mining Tools tab from the Data Destination pane.
2. Click the Classify tab, and make selections from these popup menus:

Mode: Classifier & Error

Inducer: Decision Tree

Discrete Label: churned

3. Click *Go!*

MineSet classifies and creates the Decision Tree model as shown in Figure 3-12. Notice that the estimated error rate is significantly improved (5.40%) over the Evidence Visualizer in Figure 3-4, confirming the earlier hypothesis that interactions between attributes are significant. In Figure 3-12, every node in the decision tree has two bars on it, one for each label value. Pointing to a bar will show the record count and percentage for that label value. Every node has a base, indicating the number of records that reach it, and a color, indicating the estimated error rate for the subtree, (see legend on bottom of the visualization).

In this example the root of the decision tree is marked with total day minutes, indicating that this is the single most important factor—how long these customers talked, with a dividing threshold of 264.45 minutes.

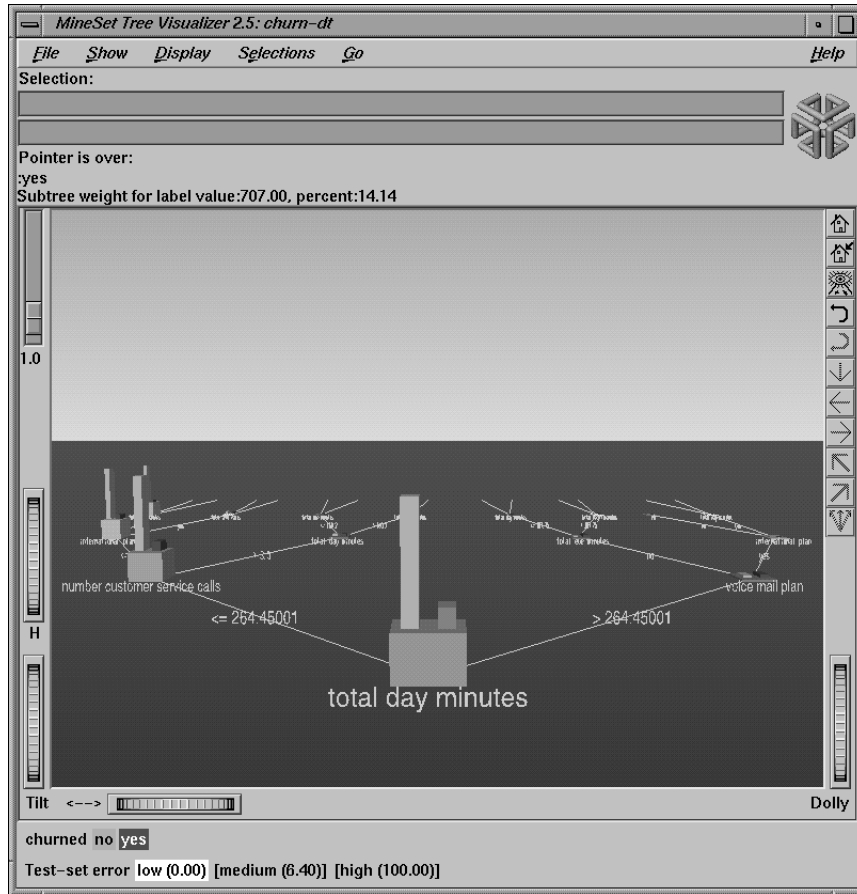


Figure 3-12 Tree Visualizer Window

You can virtually fly through the Tree Visualizer landscape using the middle mouse button. By pointing to the red bar (churned yes) at the root, you can see that 14.14% of the customers churn. Follow the right line (total day minutes > 264.45) to the child node, which contains customers that talk frequently on the phone, with 59.31% of the customers churning. Again follow the right line from the child node, which shows that of those customers who talk frequently on the phone, those with voice mail churn at a much lower rate of 9.33%. Perhaps offering voice mail to customers can help reduce churning.

It is important to understand that this tree was automatically induced from the data. The attributes chosen for nodes and the thresholds are determined by the process of induction.

To drill-through and see the original data, select a node base or a bar and choose Selections > Show Original Data. A Record Viewer will show the records matching the node you selected.

If you would like to explore MineSet further, and discover more about applying a classifier, continue to the next chapter, Chapter 4, "Further Explorations."

Further Explorations

This chapter continues to explore the MineSet tools. It assumes that you have worked through Chapter 3, “Churn Tutorial,” and prepares you to use other aspects of MineSet:

- “Exploring Data Clusters” on page 33
- “Invoking a Decision Table” on page 40
- “Targeting Customers Using a Classifier” on page 42
- “Reducing Misclassification Costs” on page 49
- “Further Exploration of MineSet” on page 55

Exploring Data Clusters

When confronted with an unfamiliar dataset, you can discover evocative attributes or characteristics using the clustering algorithm. This algorithm segments records into clusters that are similar in several ways. For this example, return to the Tool Manager window, and begin a new history by returning to `/usr/lib/mineset/data/churn.schema`.

1. In the Data Destination pane of the Tool Manager window, click the Mining Tools tab.
2. Click the Cluster tab, and make these selections:

Method: Single k-Means

Number of Clusters: 3

3. In the Data Transformations pane of Tool Manager, select and remove these columns from the Current Columns pane: `state`, `account length`, `area code`, `phone number`, `international plan` (because it correlates to `total intl charge`), and `voice mail plan` (because it correlates to `number vmail messages`.) See Figure 4-1. Remove the selected columns using the *Remove column* button. Multiple selections can be made using the Control key.

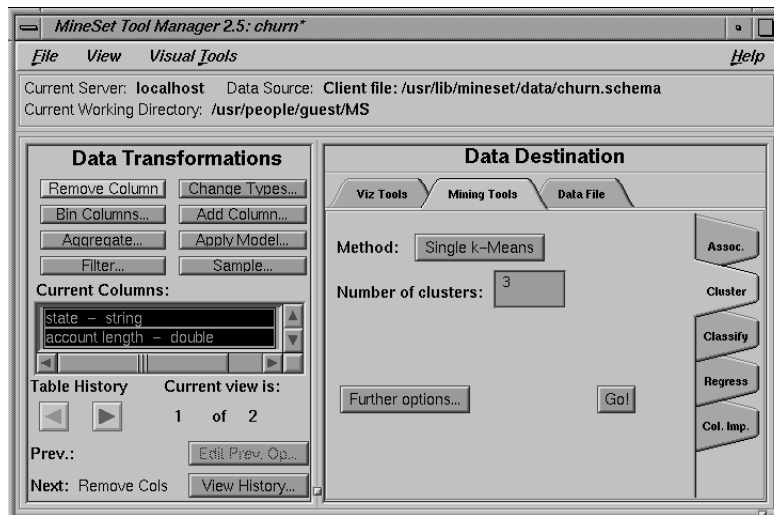


Figure 4-1 Removing Columns to Prepare for Clustering

The columns that are removed are those that are not likely to influence clustering productively. The column `churned` is retained, to help explain results. You can experiment with which columns to remove as you explore the dataset further.

4. Click *Further Options* to set the weight of the attributes.

By default, the weight of each column is set to one (1), which means each column is given equal importance. Set the `churned` column to 0 for this example, to see if this attribute is generated spontaneously as the dataset is clustered. Click *Set* then *OK*.

5. Click *Go!* on the right side of the Tool Manager window.

The Status window on the bottom of Tool Manager shows the progress of the clustering operation, as the algorithm selects significant characteristics by which to group the records. The model is saved as `churn.cluster`. This particular clustering takes the algorithm about 10 iterations.

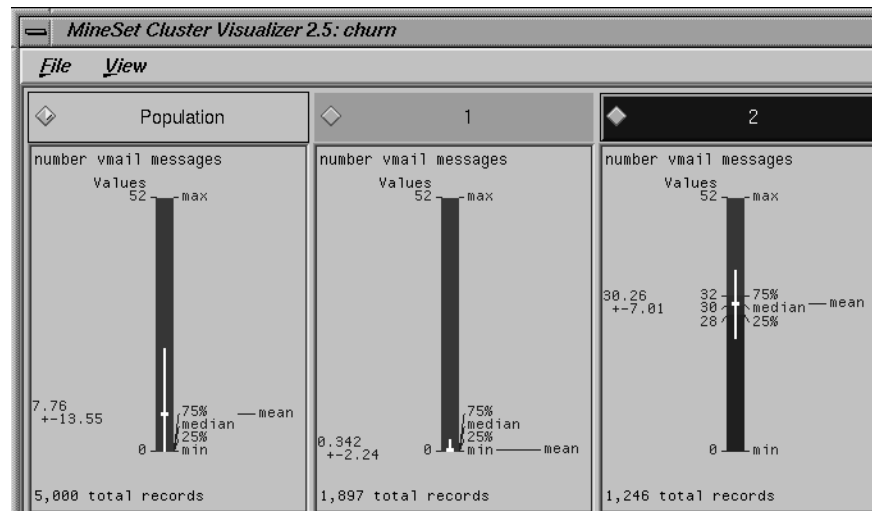


Figure 4-2 Box Plots Produced by Cluster Visualizer

Figure 4-2 shows a partial view of the result of Cluster Visualizer. The columns are sorted by their power in discriminating how one cluster differs from another. It is clear that the number of voice mail messages, total day minutes and total day charge are the most important columns. Color and means are quite different between clusters at the top of columns, yet as you scroll down the display, the differences are minimal.

6. Click the button near the cluster number to change the attribute ordering, so that attributes that are important in discriminating this cluster from the others show at the top of the display.
7. Choose File > Exit in the Cluster Visualizer window to close the window and return to the Tool Manager window.

Relating the Columns and Axes in the Model

With Cluster Visualizer you can look at independent attributes in a dataset, examine the most prominent, and see how each differs. However, to see how attributes relate to each other between clusters, Scatter Visualizer provides a clearer view. To apply the clustered model to Scatter Visualizer, you need to determine which columns should be mapped to the various axes.

1. In the Data Transformations pane of Tool Manager, click on *Apply Model* and select *churn.cluster* from the list of available models. Click *OK*.

Although Cluster Visualizer indicated three columns as the most important, each cluster's order of importance was independent, with no indication of interactions between attributes. At this point, the Column Importance tool is useful.

2. In the Data Destinations pane of Tool Manager, select the Data File tab, and click the *...Server, and named:* button. In the text field, type the filename **churn-crop**, and click *Create File*. This saves the abbreviated version of the churn dataset that is used later in this tutorial.



Figure 4-3 Saving Data File to Server

Finding Important Columns in the Clustered Model

1. In the Data Destination pane of Tool Manager, with Mining Tools selected, click on the tab labeled Col. Imp. (abbreviation for Column Importance). By default the tool selects the top three columns in terms of importance. The discrete label is Cluster.

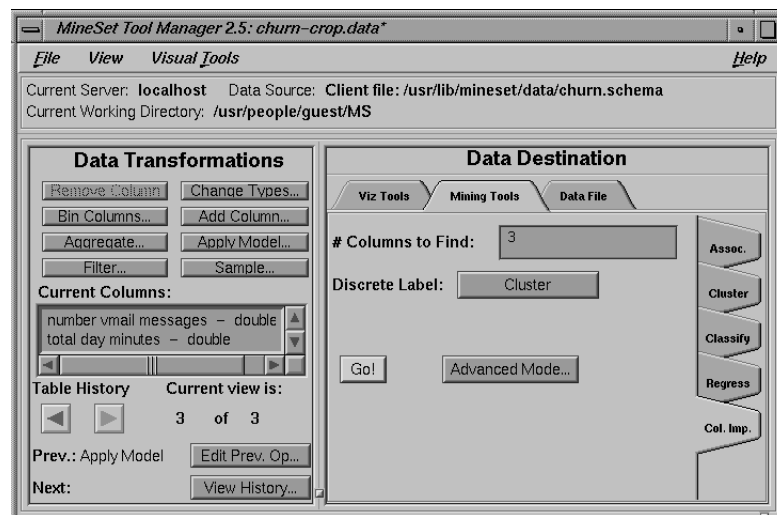


Figure 4-4 Column Importance Selections for Cluster

2. Click *Go!*

The displayed panel shows:

1. number vmail messages
2. total day charge
3. total eve charge

It seems that time spent on the phone during the day is a factor, with all other charges showing a correlation. The next step is to map these columns to axes in Scatter Visualizer.

Mapping to Scatter Visualizer

1. In the Data Destination pane of Tool Manager, select the Viz Tools tab; then choose Scatter Visualizer from the Tools popup menu.

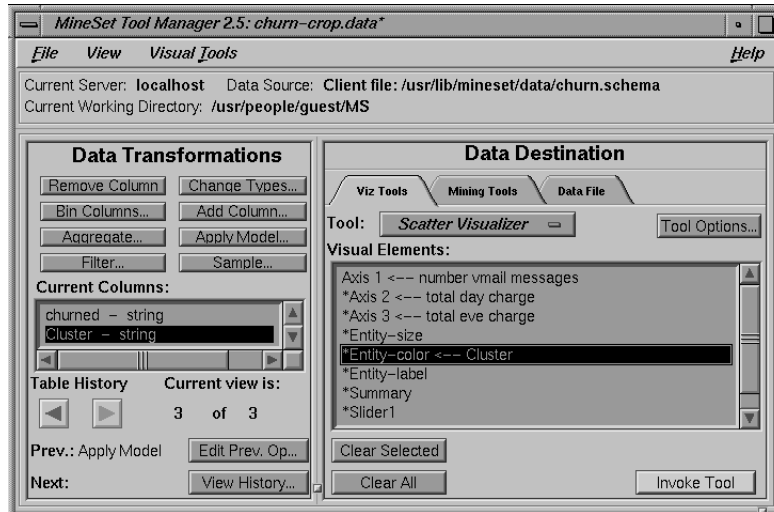


Figure 4-5 Mapping Columns to Axes for Scatter Visualizer

2. In the Data Transformations pane, map items in Current columns to items in Visual Elements by clicking first on the left pane, then the right. For this tutorial, map number vmail messages to Axis 1, total day charge to Axis 2, total eve charge to Axis 3, and Cluster to Entity-color. See Figure 4-5. When you applied the model, the column named Cluster was created.

3. Click *Invoke Tool*.

The Scatter Visualizer window in Figure 4-6 shows the clusters clearly differentiated in color. The blue cloud of scatterpoints represents cluster 2, and the flat pancake shape is split evenly between red and green—clusters 1 and 3. This pancake indicates very low numbers of voice mail messages. Clearly, total day charge and total evening charge are interdependent. If you click on an interesting visual point, the supporting data is displayed as an independent record. Dismiss the display with File > Exit and return to Tool Manager for the next step.

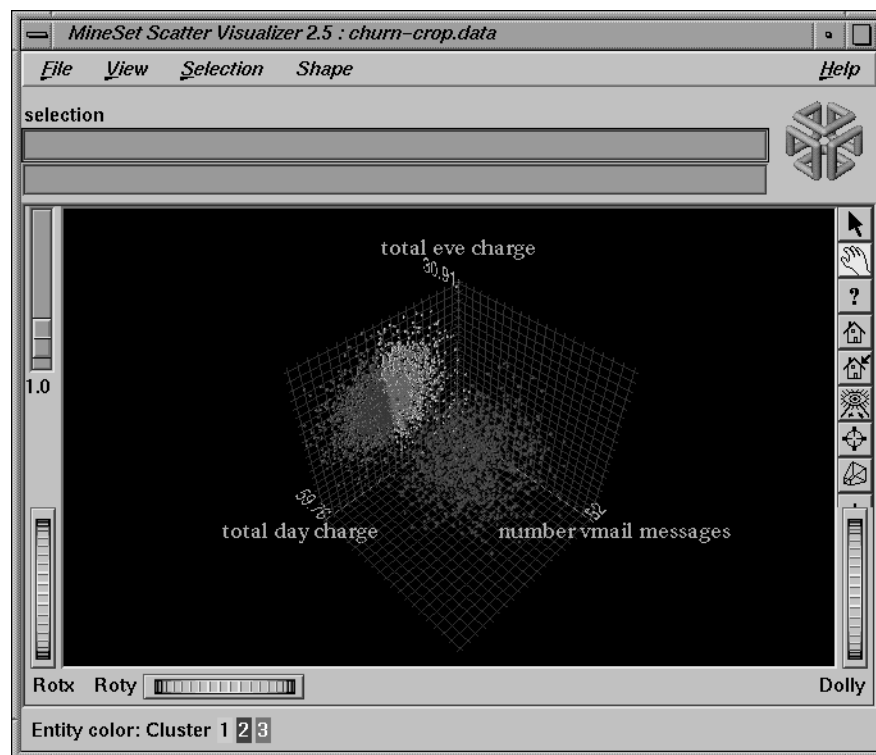


Figure 4-6 Scatter Visualization Plotted from Clustering

Invoking a Decision Table

For this example, instead of using Scatter Visualizer to see your clustered data, you can visualize the same data as a Decision Table.

1. In the Tool Manager window, choose File > Open New Data File.
2. Click the *Server File* button, and select *churn-crop.schema*. This is the file saved earlier. If you exited MineSet between sessions you are automatically returned to where you left off.
3. Click *OK*.
4. In the Data Destination pane of the Tool Manager window, click the Mining Tools tab.
5. Click the Classify tab, and make selections from these popup menus:

Mode: Classifier & Error

Inducer: Decision Table

Discrete Label: churned

Make sure you have the correct discrete label. You are about to induce a decision table and allow the algorithm to suggest which columns are most important to map to the X and Y axes.

6. Verify the *Suggest* checkbox is checked on, then click *Go!*

You can see columns being mapped to axes as the tool suggests appropriate mappings. The Status window on the bottom of Tool Manager shows progress and summary information about the induction process, including the classification error rate. When the induction step is done, the Decision Table Visualizer is automatically invoked, showing the model visually.



Figure 4-7 Decision Table Showing Clustered Churn Results

Figure 4-7 shows a round pie in the right pane, like Evidence Visualizer, indicating the overall percentage of churn. In the left pane, the data is shown as cake charts or bars, that tell you, within this subset of the data, how much churn exists. Clearly, customers with high total day charge always churn.

Center the display using the middle mouse button, and use the slider in the upper left of the window to control the differences between segments. Use the middle and left mouse buttons together to enlarge or reduce the display size.

The Decision Table shows you data at different levels of detail, taking only a few columns into consideration at first, adding more detail as you examine further. Change the cursor mode from grasp to pick, and pass the cursor across the scene to display data above the window. Notice the bar that falls out of the expected pattern—total day charge less than 29.75, and customer service calls over 3.5. Click with right mouse button on that bar to drill down for details, click with middle mouse button to drill up. Dismiss the display using File > Exit when you are finished examining the Decision Table, and return to Tool Manager.

Targeting Customers Using a Classifier

Previously, you created classifiers to predict which customers are likely to churn. Now that you have such a model, you may want to target customers who are likely to churn *before* they churn. The lift curve helps accomplish this goal.

A lift curve is a plot in which the X axis shows the number of records from 0 to 100% and the Y axis shows the number of records corresponding to customers who have a given label value (Churn=yes in our case). Two curves are shown on the graph in Figure 4-10. The lower curve (red) shows the number of customers expected to churn given a random ordering of the records. The upper curve (white) shows the percentage of customers who churn when placed in order according to the classifier's score (probability estimate) for each record. Records representing customers that the classifier identifies as most likely to churn appear first; those less likely to churn appear last. The advantage that the classifier ordering provides can be seen by the difference between the classifier curve and the random curve.

In building this lift curve, a selected classifier is applied to the test set. In the example below, a specified segment of the dataset is used for training. Then the induced classifier run on the remainder of the dataset. Although lift curves can be generated easily by selecting Lift Curve from the Tool Options for classifiers, in this tutorial a more complex scenario is shown, one that involves sampling and application of a classifier to a dataset.

Creating a Training Sample

For this example, return to the Tool Manager base window, and begin a new history by using File > Open New Data File and returning to the Local File `/usr/lib/mineset/churn.schema`.

1. In the Data Transformations pane, click *Sample*. In the Sampling dialog box type 40 for the percentage of sampling, and click *OK*.

This choice simply samples a random 40% of the total dataset, from which the classifier is induced.

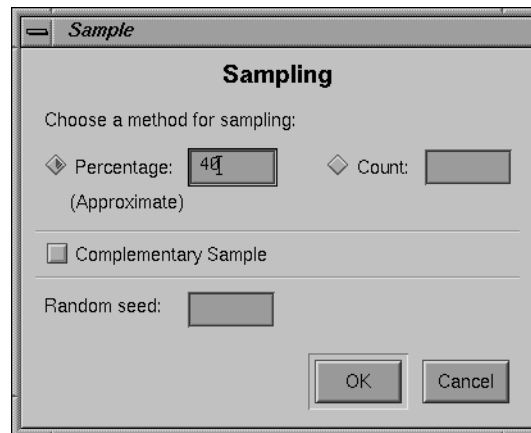


Figure 4-8 Selecting a Sampling for Testing

2. In the Data Destination pane of Tool Manager, select the Mining Tools tab; then choose the Classify tab and make selections from these popup menus:

Mode: Classifier only

Inducer: Decision Tree

Discrete Label: churned

You are inducing a decision tree classifier based on the random 40% sampling, and choosing classifier only because this is the training set. The test set will be the remainder of the dataset (excluding the 40% sampled records).

3. Click *Go!*

The resulting decision tree demonstrates the classifier, which is required in the next stage. Note that the root weight is substantially diminished, because the size of the sample is less than the complete dataset. Also note that no color appears at the base of each node, indicating that no error estimation is available.

You can see in the status field that the classifier is automatically saved under the name `churn-dt.class`. The next step is to use this classifier on the remainder of the churn dataset.

Applying a Model

Dismiss the Decision Tree window and return to Tool Manager window. Because you have used the first 40% of the dataset to build the model, you have the remaining 60% to use as a test set.

1. Click *Edit Prev. Op.* in the Data Transformations pane and you should be presented with the Sampling Dialog box again.
2. In the Sampling Dialog box, enter 40 in the Percentage text field again, but this time click the *Complementary Sample* button to indicate you want the other part of the sample.
3. Click *OK*.
4. Click on the *Apply Model* button in the Data Transformations pane.
5. From the Test and Apply Model window choose `churn-dt.class`.

- Click the Test Model tab, turn off *Show viz*, turn on *Show lift curve*, and set the *ROI/Lift label* popup menu to yes.

Having built a classifier based on the random sample, you are now planning to apply it to the remainder of the churn dataset.

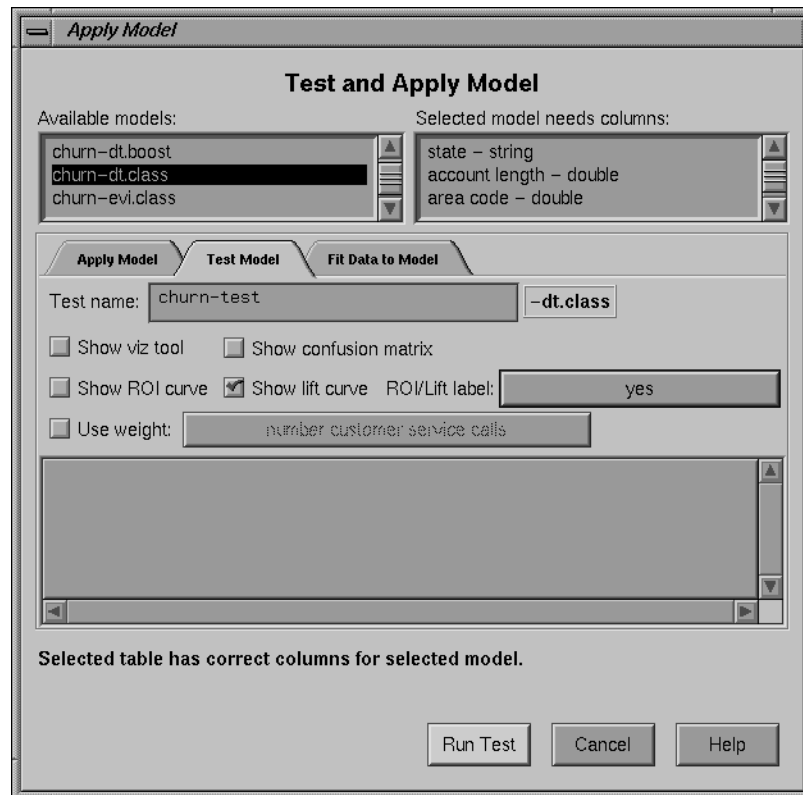


Figure 4-9 Preparing to Test Classifier on Full Dataset

- Click *Run Test*. The process takes some time. The resulting lift curve is shown in Figure 4-10, with the details of any selected point shown in the upper banner.

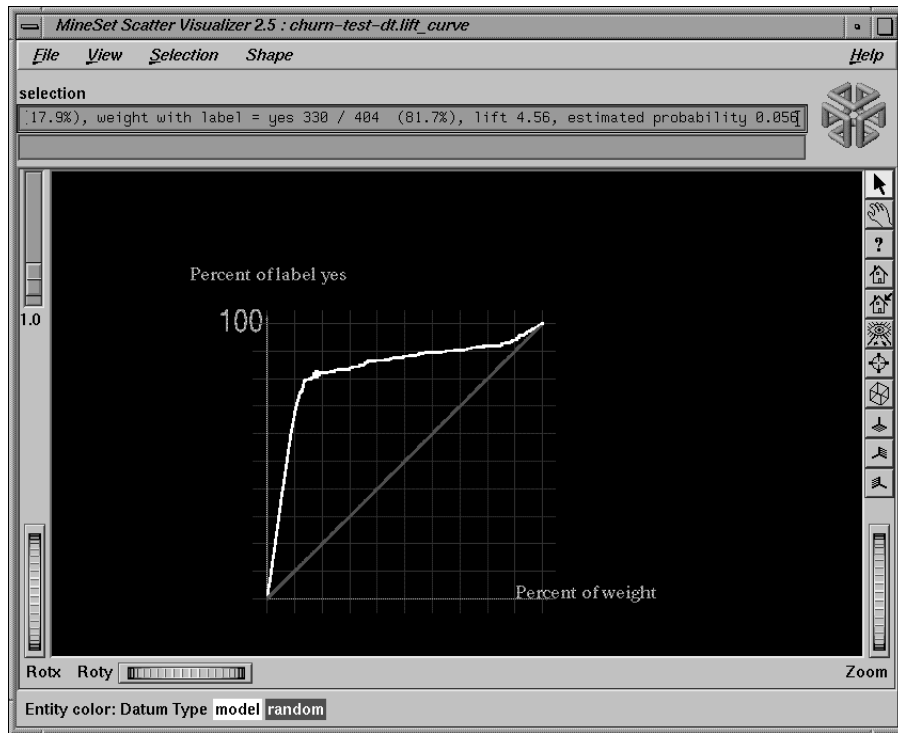


Figure 4-10 Lift Curve

Move the pointer along the white (model) line, clicking at various points to see the lift and percentage of customers with *churn=yes*. Look for the knee of the curve, in this example where the estimated probability of the classifier is 0.056.

This is the point at which the return on investment in sending incentives to customers that may churn diminishes rapidly. The next step is to apply the classifier to the full dataset.

- Return to the Test and Apply Model dialog box; click the Apply Model tab and make these selections:

Estimated probability values for label yes

*New column name: **p_churned*** (You must type this in.)

When you click *Estimated probability values for label, yes* is chosen to match the corresponding selection in the Test Model step. This process adds a new column representing the likelihood that certain people will churn (**p_churned**.) Click *OK*.

- On the Data Transformations pane of Tool Manager click *Filter*; in the Filter by Expression dialog box text field create the expression **p_churned > 0.056**. Check expression before clicking *OK*.

This is the estimated probability figure retrieved from Step 8 shown in Figure 4-10. The intention is to select only those customers with the greatest likelihood of churning. In real-life, this step would be executed against unlabeled data in order to predict which of the existing customers are likely to churn.

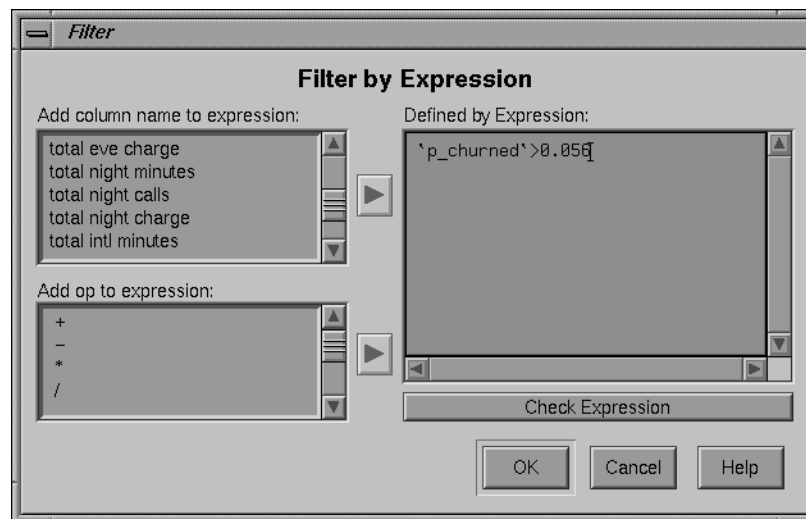
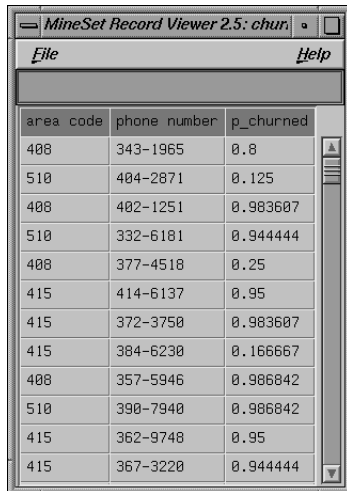


Figure 4-11 Filtering for the Probability of Churn

The final step is to see the results in Record Viewer, eliminating unnecessary columns for easier reference, as follows:

10. In the Data Destination pane of Tool Manager, click the Viz Tools tab, and select Record Viewer. In the Data Transformations pane, select all columns except `area code`, `phone number` and `p_churned`, then click the *Remove Column* button. Select multiple columns by pressing the Shift key for a range, or the Control key for specific selections.
11. Click on Invoke Tool.

The result is a useful phone list of those customers shown in Figure 4-12, who have the greatest likelihood of churning based on the model.



| area code | phone number | p_churned |
|-----------|--------------|-----------|
| 408 | 343-1965 | 0.8 |
| 510 | 404-2871 | 0.125 |
| 408 | 402-1251 | 0.983607 |
| 510 | 332-6181 | 0.944444 |
| 408 | 377-4518 | 0.25 |
| 415 | 414-6137 | 0.95 |
| 415 | 372-3750 | 0.983607 |
| 415 | 384-6230 | 0.166667 |
| 408 | 357-5946 | 0.986842 |
| 510 | 390-7940 | 0.986842 |
| 415 | 362-9748 | 0.95 |
| 415 | 367-3220 | 0.944444 |

Figure 4-12 Record Viewer Results

In Record Viewer, for every record there will be a number estimating the probability that they will churn. Filtering has retained those customers with highest numbers. That provides the list of only those potential churn customers you should give incentives to (for example, solicit by phone, send mail, and so forth.)

Reducing Misclassification Costs

You can reduce the cost of making mistakes in building the model using three important tools in MineSet: confusion matrix to give a detailed picture of errors and incorrect predictions, loss matrix to take into account that some mistakes are worse than others, and return-on-investment curve to show when investing more time or money is fruitless.

Displaying a Confusion Matrix

Return to the Tool Manager window, and reopen `/usr/lib/mineset/churn.schema`.

1. In the Data Destination pane, select the Mining Tools tab; then choose the Classify tab and make selections from these popup menus:
 - Mode:* Classifier & Error
 - Inducer:* Decision Tree
 - Discrete Label:* churned
2. Click *Further inducer options* and the Classifier options pane shown in Figure 4-13 appears.



Figure 4-13 Classifier Options Pane Showing Confusion Matrix Checked On

3. Click on *Display Confusion Matrix* in the lower right, then click *OK*.
4. In the Tool Manager Data Destination pane click *Go!*

The Confusion Matrix displays where the classifier makes mistakes in classifying. Dismiss the Tree Visualizer and examine the Confusion Matrix. From this you can construct a Loss Matrix based on what you now know about the data, to make certain kinds of errors less tolerable than others.

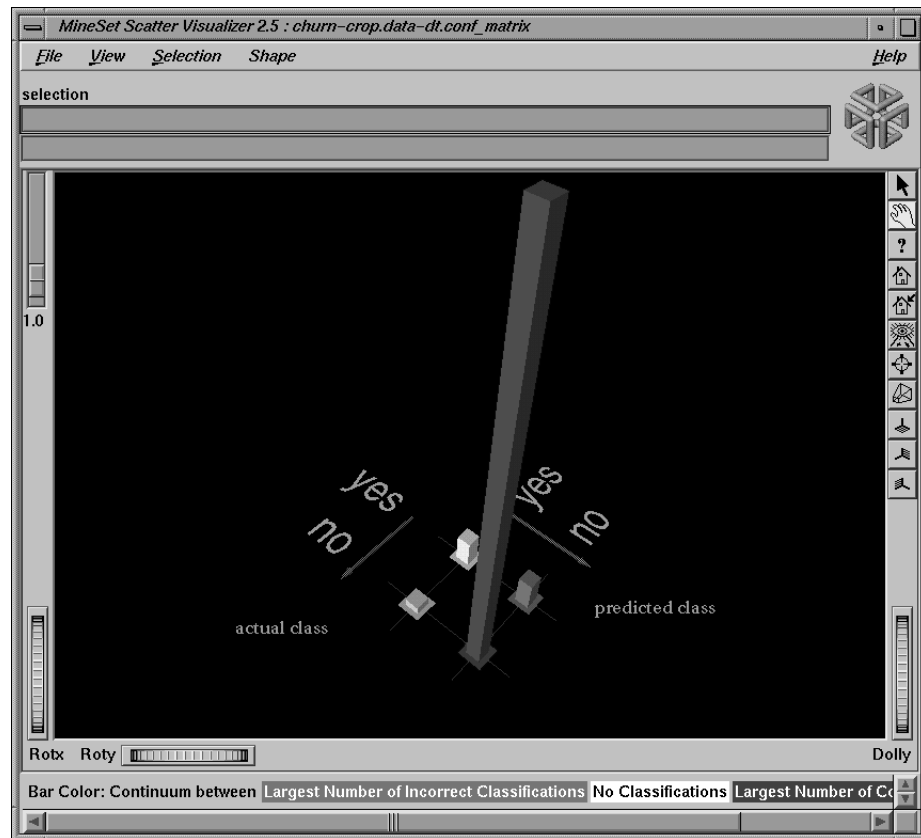


Figure 4-14 Confusion Matrix Showing Correct and Incorrect Classifications

In the window shown in Figure 4-14, the tall blue bar and the short white bar represent correct classifications; but substantial misclassification occurs in the category represented by the red bar—Predicted Class: no, Actual Class: yes. These customers were predicted not to churn, but actually did so, a costly mistake even at 4.6%. You can try to reduce that error by constructing a Loss Matrix based on what you now know about the data, and weight the errors represented by the red bar more heavily.

Defining a Loss Matrix

The purpose of this process is to control which errors the classifier will favor and which it will avoid.

1. Dismiss the Confusion Matrix display with File > Exit and return to the Tool Manager window.
2. Click *Further inducer options* to return to the Classifier options pane.
3. In the second section of the upper left of the pane, click first *Use Loss Matrix*, wait briefly.
4. Click *Edit Matrix* to weight the cost of making errors. The Loss Matrix pane similar to that shown in Figure 4-15 appears.

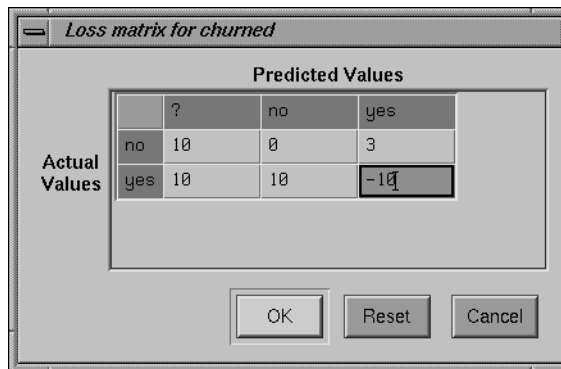


Figure 4-15 Loss Matrix Showing Weighting

5. Set these values across the rows of the Loss Matrix, reading from left to right:

Actual Values: no: 10—0—3

Actual Values: yes: 10—10—(-)10

The count in the column under the question mark should be at least as high as any number in the matrix, to prevent the classifier from predicting “unknown.” If you predict a customer won’t churn, and you are correct, you neither win nor lose (represented by zero). If you predict a customer will not churn, thus fail to mail to them, and they do churn, you incur a loss of 10 (represented by positive 10—since the numbers represent loss). If you incorrectly predict a customer will churn, and they didn’t, you lose three, representing the cost of sending a mailing unnecessarily. If your mailing program works and you save yourself from losing a customer, you gain 10 (represented by negative 10).

Viewing a Return on Investment Curve

The Return on Investment curve lets you see what the cost is of making certain kinds of errors, and indicates to you the point at which it is no longer fruitful to continue taking action

1. Make sure *Backfit test set* and *Display Confusion Matrix*, and *Loss Matrix* are checked on in the Classifier options pane.
2. Ensure *Display ROI Curve* also is checked on.
3. Ensure *ROI/Lift label* is set to yes and click *OK*.
4. Click *Go!* in the Tool Manager window.

Three display windows appear, the Decision Tree and Confusion Matrix, and the ROI Curve. Notice that the Confusion Matrix shows the classifier is more conservative in make churn=no predictions, thus reducing false negatives. The Confusion Matrix now displays a different weighting, one of 4% which takes loss into account. The errors on one side have been increased, but those on the other have been decreased. Dismiss both the Confusion Matrix and the Decision Tree display using *File > Exit*, and examine the ROI curve window.

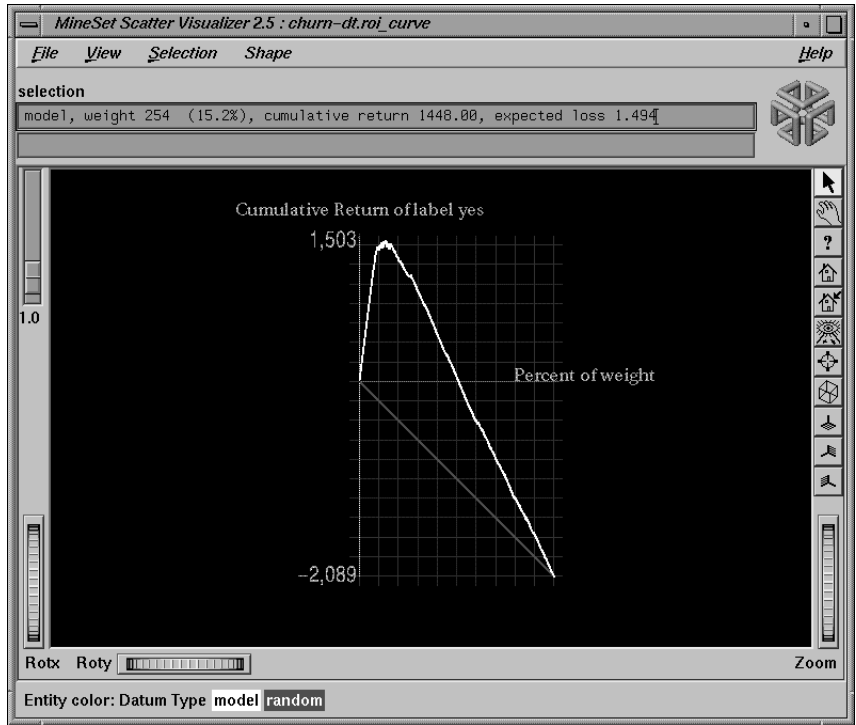


Figure 4-16 Return on Investment Curve

The ROI curve shown in Figure 4-16 bears a marked resemblance to a lift curve. The horizontal line across the middle represents zero profit and loss. The red line represents the expected performance if you were to take a random sample of the population and send them mail. You expect a loss if you mail to everyone, because of the cost of mailing. However, there is a point of optimum return on investment, represented by the knee of the curve, at 1448, or 15.2 percent of the population.

Further Exploration of MineSet

See the *MineSet User's Guide* for descriptions of these tools and what the analytical data mining algorithms can show. The manual is online and can be launched by selecting Help > MineSet User's Guide.

This tutorial has only been a brief introduction to the MineSet tool suite. Other aspects covered in the *MineSet User's Guide* include:

- Scatter Visualizer.
- Tree Visualizer for visualizing hierarchies.
- Option Tree Inducer and Classifier.
- Association Rules Generator and Visualizer.
- Regression, to allow you to predict a continuous value instead of discrete.
- Transformations, including binning, distribution, and indexing of arrays.
- Record weighting, which allows assigning different weights to different records, because some records are more important than others (for example, highly profitable customers).
- Learning Curve, which can help you determine whether sampling can be done on your dataset to speed up the knowledge discovery process, without losing much of the accuracy of the induced classifiers.
- Many tool options, including color manipulation, message boxes.
- Animation sliders for visual tools.
- Batch processing. The program `mineset_batch` can be used to execute operations non-interactively. This is useful if a job needs to run regularly (for example, once a night).
- Error estimation using advanced techniques such as cross-validation.

Also described in the *MineSet User's Guide* are the technical details of file and data manipulation.

Note: Data mining algorithms find correlations that may not be causal. A well-known discovery is the strong correlation between shoe size and reading ability: the larger one's shoe size, the better the reading ability. This correlation, while true, is not causal; both shoe size and reading ability improve with age (as children get older, their shoe size and ability to read both increase.) You are cautioned against attributing causality to discovered correlations. Wearing larger shoes is unlikely to increase your reading ability.

Tell Us About This Manual

As a user of Silicon Graphics products, you can help us to better understand your needs and to improve the quality of our documentation.

Any information that you provide will be useful. Here is a list of suggested topics:

- General impression of the document
- Omission of material that you expected to find
- Technical errors
- Relevance of the material to the job you had to do
- Quality of the printing and binding

Please send the title and part number of the document with your comments. The part number for this document is 007-3573-003.

Thank you!

Three Ways to Reach Us

- To send your comments by **electronic mail**, use either of these addresses:
 - On the Internet: techpubs@sgi.com
 - For UUCP mail (through any backbone site): *[your_site]!sgi!techpubs*
- To **fax** your comments (or annotated copies of manual pages), use this fax number: 650-932-0801
- To send your comments by **traditional mail**, use this address:

Technical Publications
Silicon Graphics, Inc.
2011 North Shoreline Boulevard, M/S 535
Mountain View, California 94043-1389

