



ESTARD

2004-2008 ©

ESTARD Data Miner 2.0

QuickStart Guide

CONTENTS

INTRODUCTION	3
GETTING STARTED	4
MINIMUM SYSTEM REQUIREMENTS	5
INSTALLING ESTARD DATA MINER	6
UNDERSTANDING HOW EDM WORKS	7
LOADING AND ANALYZING DATA	14
WORKING WITH STATISTICS	22
CREATING RULES	25
CREATING DECISION TREES	30
USING RULES	31
HOW THIS INFORMATION CAN BE USED?	34
FREQUENTLY ASKED QUESTIONS	36

ESTARD DATA MINER 2.0

QUICK START GUIDE

WELCOME to ESTARD DATA MINER!

ESTARD DATA MINER 2.0 is a newest data mining tool for organizations of all sizes. EDM flexibility allows using it in any field – from business up to science. With ESTARD DATA MINER 2.0 you can gain a power to find out what drives your business or to discover secrets of Universe within few clicks.

QUESTIONS?

View your help file

To open your help file:

1. Start ESTARD Data Miner
2. Select the Help menu
3. From the Help menu select “ESTARD Data Miner Help”
4. Help window will be displayed

SUPPORT

In case if you are experiencing problems with our product or if you need additional support of our experts for applying our product to your case, the best way to reach our support team is to email support@estard.com with a detailed description of your case. Our customer support team will create a case for you, contact you, and solve your problems.

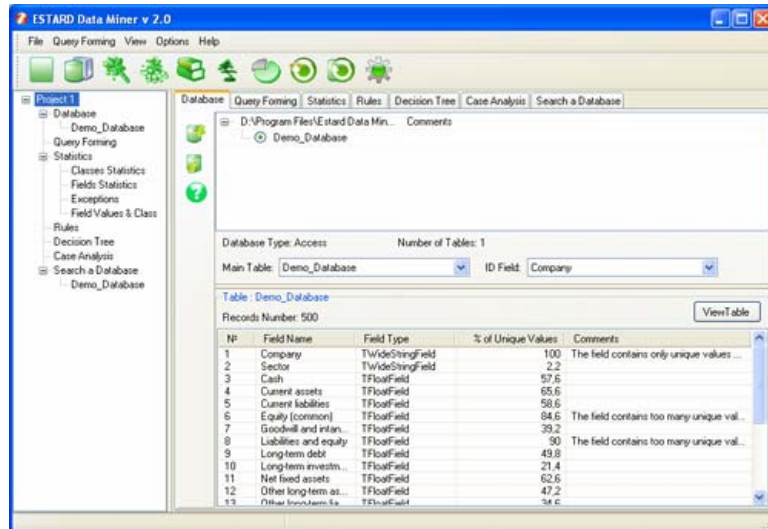
THANK YOU FOR CHOOSING ESTARD DATA MINER!

LETS GET STARTED

QUICK START GUIDE STRUCTURE

This Quick Start Guide provides all necessary information for starting your work with EDM. This document contains such sections:

- **Minimum system requirements**
- **Installing ESTARD Data Miner**
- **Understanding how EDM works**
- **Loading data**
- **Using settings**
- **Creating and viewing statistics**
- **Creating and viewing rules**
- **Creating and viewing decision trees**
- **Using rules and decision trees**
- **How this information can be used?**
- **Frequently asked questions**



MINIMUM SYSTEM REQUIREMENTS

- 1 Ghz Pentium/Celeron or compatible CPU
- 512 MB RAM
- A graphics card capable resolution 1024x768 resolution
- 200 MB of available disk space
- Windows Vista/ Windows 2000/ Windows XP/ Windows 98/ Windows ME
- Adobe Acrobat Reader 6.x or later

HINTS

- *Program performance will significantly increase with processor speed and memory increasing*
- *Only Intel and Intel compatible processors are supported*
- *Necessary disk space depends on program usage and varies depending on it*
- *Installation may require Administrator privileges*

INSTALLING ESTARD DATA MINER

The installation process will depend on how you obtained your software copy.

In case if you have downloaded the ESTARD Data Miner from ESTARD web site, the first step will be to locate the downloaded file. Installation file name is edataminer.exe. After locating the file follow these steps:

- Double click the edataminer.exe file. Installation will be started.
- Follow the on-screen instructions in the installer.

If you've purchased EDM and have received your copy on a CD, follow these steps:

- After inserting the CD into your CD or DVD driver the installation should be started automatically. If it does not, follow these steps:
 - Open directory to view CD contents.
 - Double click edataminer.exe file. Installation will be started.
 - Follow the on-screen instructions in the installer.

UNDERSTANDING HOW EDM WORKS

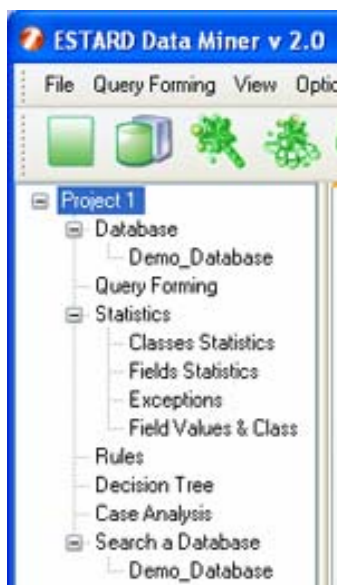
ESTARD Data Miner is a powerful business intelligence tool full of various features able to help you to make better decisions. Unlike most reporting and statistical business intelligence tools, EDM provides you with something more than raw statistics – it can detect hidden dependencies in your data, point out anomalies and help you to predict future events!

To launch ESTARD Data Miner click on the Windows Start button, select ESTARD Data Miner and click ESTARD Data Miner icon. EDM will be launched.

Navigation

Depending on your goal you can use project tree view, page tabs or menu bar to navigate through program.

Project Navigation Tree



The project navigation tree is situated right below the toolbar. Clicking on this tree you can navigate to specific statistics page, rules and decision trees pages and other program pages. On the top of the tree the project name is displayed. After loading a database for analysis, database name and tables are also displayed in the tree. Clicking a table name you can view details of different tables: number of records, fields, etc.

The Toolbar

The toolbar gives you quick access to main features you will need during your work.



From left to right:

- New project button – creates a new blank project. A project is a collection of data you use during your work.
- Database – opens a program tab that lets you start the database loading wizard and view database details. Use this button if you want to open a database you'd like to use for statistics, rules, or decision trees creation.
- EDM Wizard – click this button to start a wizard that will help you to create statistics, rules or decision trees, guiding you step by step.
- Manual Query – use this button to switch to program tab for statistics, rules and decision trees creation.
- Rules – opens a tab that lets you view, edit, export and report created rules.
- Decision Tree – opens a tab that lets you view, export and report created decision trees.
- View Statistics – opens a tab that lets you view and print project statistics.
- Case analysis – opens a tab on which you can use obtained rules and decision trees to analyze how specific scenarios correlate with your data.
- Search a Database – use this button to open a tab on which you can use obtained rules and trees to find records in a database that meet criteria of selected rules or tree nodes.

- Options – opens a dialog for setting some general EDM options.

Program Pages (Tabs)

All project data you will use and obtain during your work with ESTARD Data Miner is placed on tabs. Below are the descriptions of program tabs:

- "Database" tab is used for loading and viewing currently loaded database used for data mining: obtaining statistics, decision rules and decision trees.
- "Query Forming" tab - use this tab instead of "Query Wizard" to create statistics, decision rules and decision trees.
- "Statistics" tab lets you view graphs after creating statistics. Use this page to view basic dependencies between fields and explored values.
- "Rules" tab - use this tab to view, export, edit and report rules.
- "Decision Tree" tab - this tab contains controls for viewing and working with decision trees.
- "Case Analysis" page - use this page for fast analysis of dependencies between your data and obtained rules and trees.
- "Search a Database" page - contains controls for searching for target records in a database using obtained rules and trees.

TIPS FOR UNDERSTANDING DATA MINING

Before creating your first project with ESTARD Data Miner it is important to know several rules of working with data mining application.

What Data Mining Can Do?

Data mining techniques implemented in ESTARD Data Miner can be powerful support in your decision making process. Decision Rules and Trees can help you to analyze and understand your business, create decision models that can be used for planning and predicting. For example, what if you could predict with high level of probability whether the new customer is potentially fraudulent, or detect what is most important in your service for increasing the number of loyal customers? ESTARD Data Miner can perform such analysis!

What Data Mining Can't Do?

It is important to remember that none of predictive techniques gives 100% results. The main aim of data mining is giving help in decision making, but the final decision is always after you.

Using Data Mining Results

ESTARD Data Miner gives you something more than a powerful tool for decision rules and decision trees creation - it also gives you a set of features for fast and easy implementation of results. As a result, you can use the program not only for planning and modeling situations, but also for fast analysis of a case or a database by using rules and trees.

TIPS FOR WORKING WITH DATA MINING

Use duplicates of databases

Although we've carefully developed an application that will analyze your database safely and most effectively, it is important to protect your information in every possible way. This is why for data mining we recommend to use copies of the data you want to analyze. For this purpose you can create a copy of your data by exporting it to Excel, or simply create a copy of your data warehouse. Besides, in such a copy you will be able to set experiments, by changing some data, or clean the data without any harm for the main data set. This is why we recommend not to use your main production databases.

Understand the results you will obtain

ESTARD Data Miner is a powerful tool that gives you technologies for understanding your business processes, for analyzing and predicting what to expect in the future. A BI application gives you an interpretation of data, but it is important to remember that all results you will obtain are an aid in decision making, and the final decision is always after you. And that there is no technology that is able to give 100% results.

Database preparing

Although ESTARD Data Miner is able to work with "unclean" data, duplicate records, empty or incomplete records and mixed data types in fields will most likely make the data mining results incomplete, or result in incorrect rules and decision trees. This is why it is recommended to use "clean" data. Besides, in case if you want to use several tables from one database, these tables will need an ID field - a field containing only unique values and identifying connection between records in two or more tables.

What to analyze?

Before starting the analysis from its goal, this will help with selecting fields to analyze and picking best rules and trees for prediction. Remember to select fields from the table that might have some correlations with your goal and don't use fields that contain only unique values, for example, don't use fields containing customers or companies names - this will most likely return "over-fitted" rules and trees, describing every single customer or company, but not giving information about trends or groups in your data. Forming a goal of analysis will help you to manage settings and results of analysis.

Rules and Decision Trees

Rules (also called if-then rules, or production rules) and decision trees are powerful data mining methods allowing analyzing hidden correlations in your data. In ESTARD Data Miner you have a possibility to use both these methods. With the help of these methods you can create models that will describe your data and will help in further decision making. For example, you could create a model of customers that are most likely to become fraudulent or bad debt. After that the decision model can be spread between employees responsible for working with customers. With the help of customers models they now could decide whether to proceed with new client, or to get more information before making the decision.

Don't stop after one try

BI techniques implemented in ESTARD Data Miner allow to fast and easily repeat analysis as many times as you need, until you receive the best results. For each case you will need to try several times, changing rules or trees settings, selecting new fields, or reducing

their number, before you will achieve your goal. Don't stop after one analysis, change the settings and try again, and you will surely receive new results. It is hard to guess what settings will suite your case best, but it's better to start with higher values (for example, rule cases or rule probability) and then go on lowering these settings. In case if you start with low values in settings, you might have to wait for a long time before receiving rules, and their number will be very high, while they will describe very small groupings of data.

This is by no means the full list of rules for BI application, but they will help you with starting your work.

Using and Saving Projects

Use projects to save data obtained during your work with ESTARD Data Miner. A project can be shared with other EDM users.

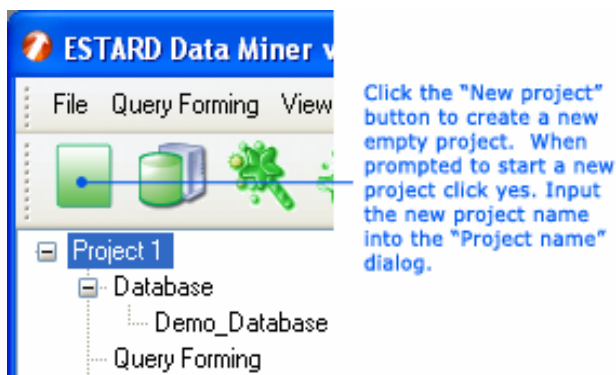
To save a project:

1. From the “File” menu select “Save Project”
2. Input the file name and click “Save” button
3. The project will be saves with a “.EPR” extension

LOADING AND ANALYZING DATA

First step to start data mining is loading the database you'd like to analyze. In Demo version of EDM a demo database is automatically loaded on startup. If you are working with full version of EDM or you would like to analyse your own database, your first step should be database loading.

The database used for creation of rules or decision trees is called the Learning database and can be loaded on the "Database" page or



from the main menu.

Loading a database is performed in a few steps with the help of a "Wizard".

To load your data follow these steps:

1. From the file menu select "Load Learning Database". This will open the database loading wizard.
2. Choose the database type from the "Database Type" drop down menu
3. If access to your database requires authorization, mark the "Authorization" checkbox and input username and password
4. Click Next
5. Click the "Open Database" button to start "Open" dialog
6. Find the database you want to open and double click to choose it

7. The database loading wizard will show you the path to the selected file and the list of tables in the file. If the loaded database contains no tables, you won't be able to load an empty database.
8. Click the "Finish" button
9. View the database on the "Database" page

The database and database tables are displayed in the project tree view

Click on a table on the Database page to view table fields, fields types and details

Click this button to view the selected table in a new window. Use it for exploring tables or search for values

Select the main table from this drop-down list, to change the ID field connecting tables

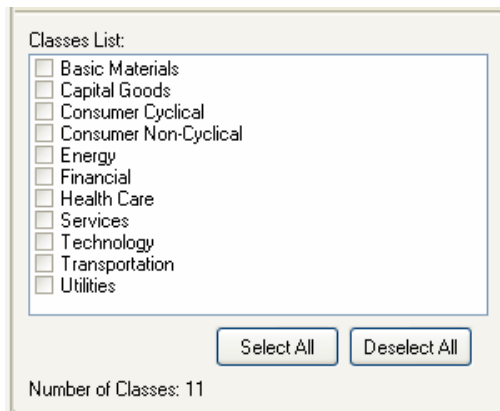
Use this column to explore the percent of unique values in a field. Fields with 100% can only be used as ID

Select an ID field from this list of table fields. Only fields with 100% unique values can be selected

After uploading a database EDM automatically performs the first analysis of fields and data density(% of unique values) .

Analyzing your data

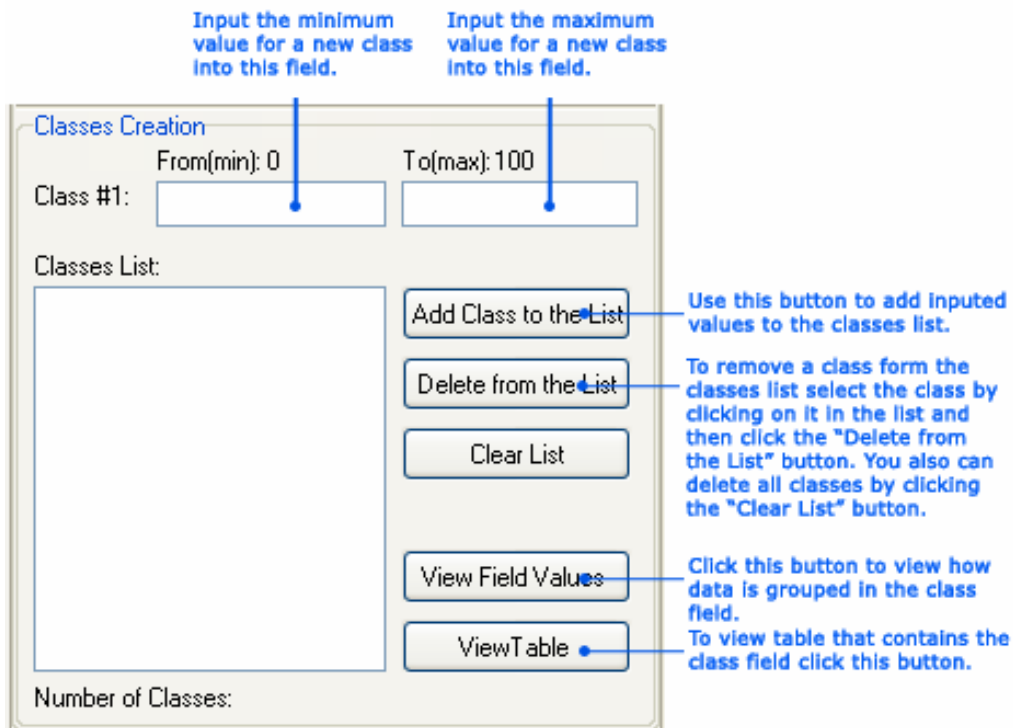
- Click the “Query Forming” tab
- To have a better view of your fields re-size EDM window, maximizing it. To re-size the “Class/Examined Field Selection” pane click-drag the divider below the pane.
- Click the check box next to field “Sector”. Note that the



“Classes List” pane displayed all entries(unique values) detected in the field. Text fields always provide a pre-populated list of values based on data from selected

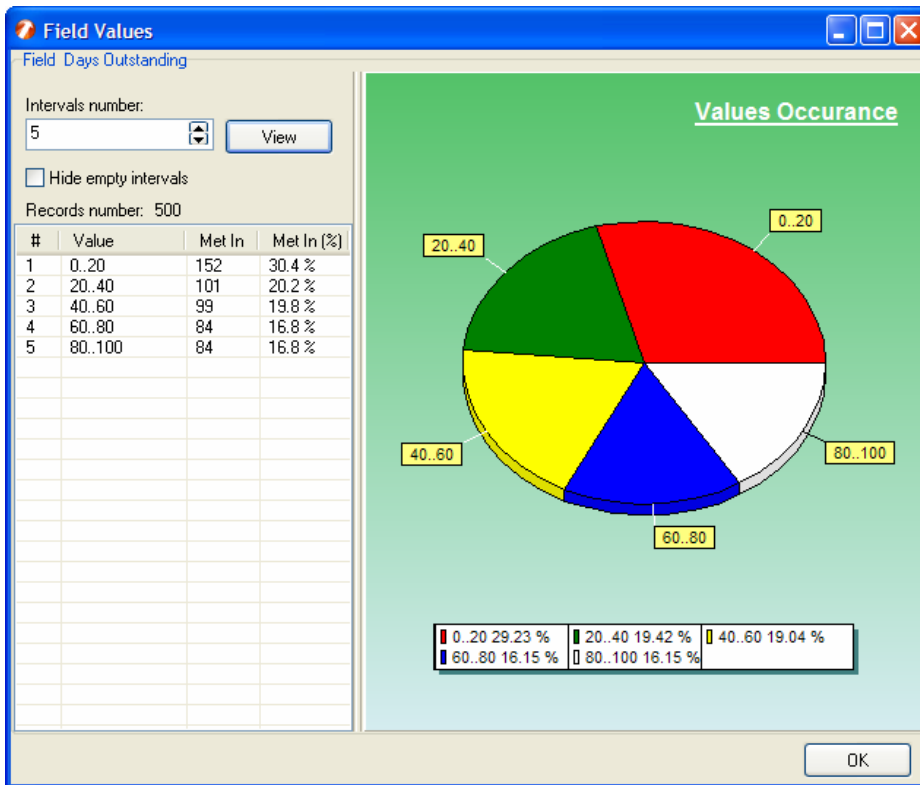
database table.

- Note that the class field is marked by red color in the “Fields Selection” list.
- Click the check box net to “Days Outstanding” field. Note that the “Classes List” pane changed and now displays a form for numeric classes creation.



- In our exercise we will analyze this field, but for better understanding how to analyze your data we recommend to try working with different fields and viewing the dependencies.
- Note that in the class creation pane near the “Class #1” label there are two fields for inputting values. Above the fields there are labels containing minimum and maximum values (“0” and “100”) found in the “class” field. This is displayed to give you a hint where you can begin and end your class ranges.
- Creating ranges for class field can be a hard task for the first time. To further assist with the analysis you can ask EDM to show you quick analysis of your field by clicking the “View Field Values” button. You might need to re-size the window, maximizing it for further work.

- In the opened window you will see that EDM has created two range values – class 1 from 0 to 50 and class 2 from 51 to 100. You can edit the number of ranges by changing the “Intervals Number” value and clicking the “View” button.
- Select five in the “Intervals Number”, and click the “View” button. EDM will create five equal ranges. These ranges are provided as a guide for you. You may use them or create your own. On the graph you can view how many records belong to each of created ranges. If you create more ranges, you might see that some of them do not have any corresponding records at all. This gives you information about what ranges will give better results for your analysis. Try using your own ranges and proposed ranges and see how results will vary.



- In the “From” and “To” fields input the values 0 and 10 and click “Add Class to the List” button. Note that in the “Classes List” the value “0..10” is now displayed. This is the first “class”.
- Repeat the above step with values 11 to 20, 21 to 40, 41 to 60, 61 to 80, 81 to 100.

HINT

As you can see, the created classes do not overlap. For example, the first class starts at 0 and ends at 10 and the next class starts at 11, instead of 10. When creating classes it’s important to remember that they shouldn’t overlap. EDM will notify you if you will try to create overlapping classes and will point you where is the mistake.

- After you create all classes as described above “Classes List” pane should look like in the example below.

The screenshot shows a dialog box titled "Classes Creation". At the top, there are two input fields: "From(min): 0" and "To(max): 100". Below these is a "Class #7" label and two empty input boxes. The main area is labeled "Classes List:" and contains a list of six classes: "0...10", "11...20", "21...40", "41...60", "61...80", and "81...100". To the right of the list are five buttons: "Add Class to the List", "Delete from the List", "Clear List", "View Field Values", and "ViewTable". At the bottom left, it says "Number of Classes: 6".

- Next step is selecting a list of fields that ESTARD Data Miner should analyze. Click the “Recommended” button below the “Fields Selection” list. Note that all recommended fields, marked by bold black color were automatically selected.

HINT

ESTARD Data Miner automatically performs analysis of table fields and determines which are best suited for analysis. Fields containing too many unique values, like names, surnames, etc, often don't give good results for data mining. Such fields are considered to be unsuitable and are marked by grey color in the “Fields Selection” list. “Unsuitable” fields are not selected when clicking the “Recommended” button.

- In our exercise we will also use several fields that belong to not recommended. Select fields “Total operating expenses” and “Sales”. To select them check the boxes next to field names in the “Fields Selection” list.

When selecting fields for analysis it is important to answer a question “can this field have influence on the “class” field?”. If you know that some field has direct influence on the field selected as “class”, or if you don't know and just want to find out if such impact exists, select the field. For example, if you are analyzing your customer database for payment information, the company size information might be valuable, while name of the company will not have any impact on results.

- To start the analysis, click the “Process Statistics Query” button.

- **Once EDM performs the analysis it will display a dialog notifying that now you can view statistics or create rules and decision trees. Click “OK” to dismiss the dialog.**

WORKING WITH STATISTICS

After performing first analysis of selected data, the next step is to view obtained by ESTARD Data Miner statistics. Viewing statistics is recommended for understanding which fields will give better results if selected for rules or decision trees creation. After detecting such fields you can go back to query forming page, adjust settings, change analyzed classes, and repeat statistics query as many times as you need. Using statistics will help you to concentrate on fields that you feel will give best results.

Viewing Statistics

To view project statistics click on “Statistics” tab.

- The first tab that will be displayed is called “Classes Statistics”. On this page you can see how the examined class values were grouped and how many times each value was met. On a graph you can see a graphical representation of classes data. This tab is helpful to understand whether the classes you created are providing appropriate groupings, or whether you might need to go back and make some changes in classes or class field.
- After viewing class statistics click “Fields Statistics” sub-tab to view statistics for other fields you’ve selected for analysis. Use “ + ” and “ – ” buttons to collapse and expand fields and field values tree view. Fields that were not used for analysis are marked with grey color. Click on a box next to field names to view graphical map of distribution of values for the field.
- To view anomalies in your data, click on “Exceptions” sub-tab. In case if this tab is empty, no exceptions were

detected. In other case you can view a list of exceptions and use these anomalies to avoid creating rules describing such anomalies – incorrect rules.

- Next click on “Field Values and Class” sub-tab. This tab is very useful to view dependencies of fields you’ve analyzed to classes. View how your data is distributed by classes and how field values can be helpful during rules and trees creation. Use “ + ” and “ – ” buttons to collapse and expand fields and field values tree view.
- Click on a box next to “Cash” field and note the graph displays distribution of the “Cash” field values for class “0..10”. Select another class from the “Classes” drop-down list or other fields from tree view. Note that some field values can have obvious correlations to field values. For example, for “0..10” class value “High End” in the “Product” field is met 40 times. This means that our first correlation can be established.

HINTS

Statistics are useful for two reasons – data analysis and planning for rules and decision trees creation. As you can see from this exercise, EDM provides you fast analysis on data distribution for each selected field. Besides, using statistics you can easily understand how the fields you’ve selected impact on classes you created. The more distributed the field data is, the more rules can be created with this field, but from the other point of view, such rules will describe smaller data patterns.

For example, go back to statistics query processing and click the “Select All” button below the fields list. Repeat the statistics query and go back to “Field Values and Class” sub-tab. Make sure the

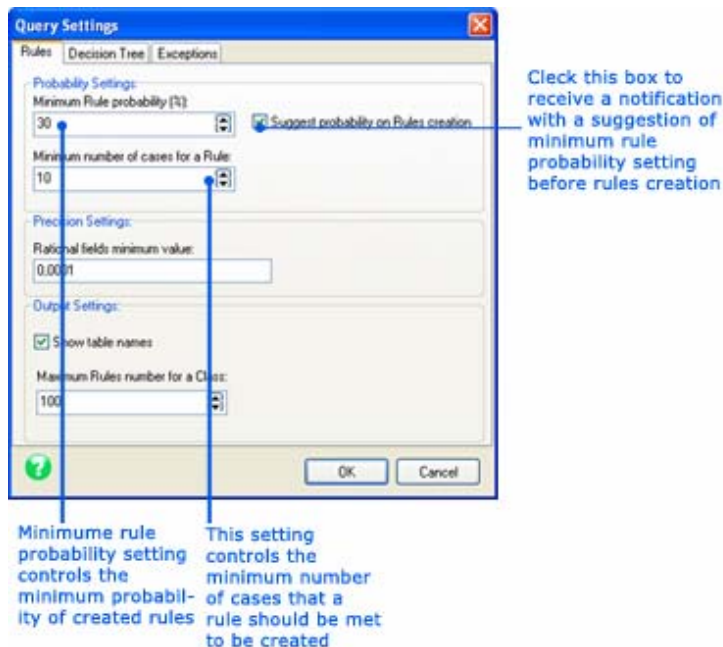
“Company” field is checked in the fields list. As you can see, the distribution of values in this field is very wide, and each value describes only one record in the table. It is not recommended to use such fields for rules creation, because they can not return good results, but will have influence on program performance.

CREATING RULES

Ability to create rules, also called production rules, or IF-THEN rules is one of most powerful features of ESTARD Data Miner. Used properly, this feature will allow you to detect dependencies in your data, and receive best results from data mining. We recommend you not to stop after obtaining your first rules, but keeping playing with settings until you find your optimum results.

To create rules follow these steps

- Before creating the rules make sure you've created Statistics.
- To create decision rules check the current rules query settings ("Query Forming" tab, settings button).



- It is better to start with higher settings values and then repeat rules creation with lower settings. First time you might get no rules at all, or get a small number of rules, but if you are working with a large database with many thousands of

records, such action sequence will highly increase the performance and as a result you won't get "over-fitted" rules. For example, you want to analyze a database with 40 000 of records. As the class field you've selected a field that contains such values: True/False. In this case if you set rules cases equal 5, you would probably get thousands of rules that will describe very small data patterns.

- If it is hard to decide what value to set for "Minimum Number of Cases for a Rule" - check "Classes Statistics", select the smallest value in the "Met In" column and set it for the "Minimum Number of Cases for a Rule". The program will not create rules that are met in less cases than set in this setting.
- "Minimum rule probability" setting also has direct influence on the number of rules, and, as a result on time necessary for their creation and output. For example, if you set this value to 10% then only rules with probability higher than 10% will be created. It is also recommended to start with higher values for this setting. For your first rules creation you can use maximum value in the "Met In %" column on "Classes Statistics" page.
- "Maximum rules number for a class" setting controls the maximum number of rules that will be created for each selected class. If you want to obtain as many rules as possible, set this value to number larger than number of records in the table, in other case select a smaller value. This setting allows to stop rules creation for a class when their number is very large and program performance becomes too low.
- Set "Minimum rule probability" to 20 % and "Minimum Number of Cases for a Rule" to 5. Set "Maximum rules number for a class" to 5000.

HINT

Before obtaining your final rules you might need to come back to these settings for several times until you find the settings that yield the rules you find most useful. Depending on data you are analyzing, these settings might need to be set very low in order to obtain rules.

After you've checked the rules settings, start their creation by pressing "Create Rules" button. In case if you've selected rules creation settings that can correspond to rules number larger than "Maximum rules number for a class" EDM will ask you whether to continue rules creation for the class. Click "No" if you don't want to obtain all rules. If you want to obtain ALL rules, click yes. In this case rules creation can take several minutes.

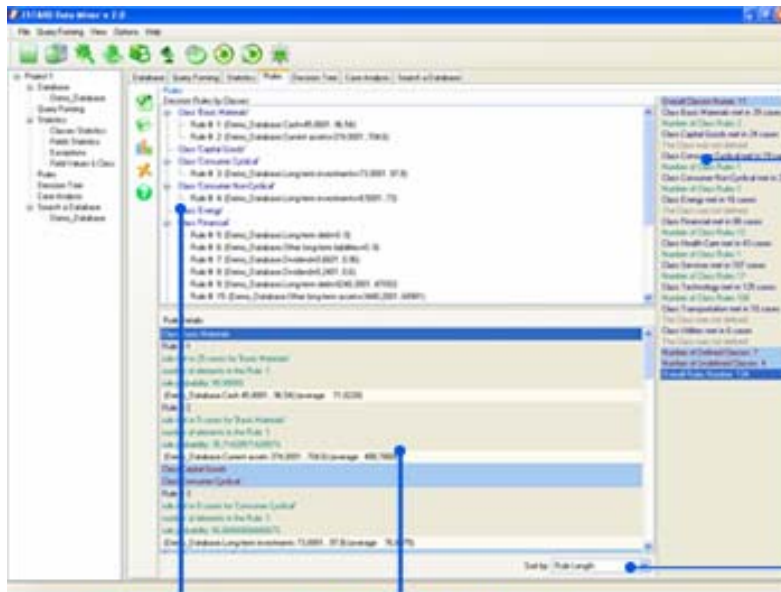
The program will automatically switch to the "Rules" page after creating the rules.

For exercise use the mentioned settings for the first time and then try lowering down "Minimum rule probability" setting for 20% and 19%. You will see the growth of number of created rules.

NOTE

In ESTARD Data Miner demo version the number of outputted rules is limited to 15 rules.

Viewing rules



The Class pane gives you information on rules statistics, including rules number for each class

Use this drop-down for sorting rules in "Rule Details" pane by rules length, probability or number of rules cases

The "Decision Rules by Classes" pane contains all detected rules grouped by classes

The "Rule Details" pane displays detailed information for each rule. Click on a class to view details of all rules that belong to it, or click on a selected rule to view its details.

EDM RULES TAB

All rules are grouped by classes and sorted by length. In the "Rule Details" form you can view details for each rule, or sort rules by length, probability and cases. After sorting, rules will remain grouped by classes. On this page you also can view the number of rules for each class, undefined classes and overall rules number. You also can create a report or export rules to .TXT or .XLS files.

HINT FOR ADVANCED USING

You also can use this page for rules editing. Click the "Edit Rules" button. The "Edit Rules" window will appear. To edit a certain rule first select the class it belongs to from the drop-down. Then select the rule from the "Selected Rule" drop-down. Now you can delete the whole rule, or delete a part of this rule, or change rule values. After you've finished editing click the "Close" button. Remember that

rules are edited immediately after you delete them or change their values.

CREATING DECISION TREES

About Decision Trees

Decision trees are another data mining method available in ESTARD Data Miner. Algorithm of this method differs from rules (if-then rules) algorithm and thus differ their results. Combining both results allows looking at your data from different points of view. For first time users decision trees might seem easier to view and understand. Decision trees however are not as powerful as rules.

Decision trees can be helpful for several reasons:

- For identifying direct relations in data that normally wouldn't be apparent
- For data groups identifying
- For obtaining a simple model of how one data pattern depends on other patterns

To create decision trees follow these steps:

- Make sure you've created statistics as described above.
- To create your first decision tree just click the "Create Decision Tree" button on the "Query Forming" tab.
- When the process of decision tree creation is complete you will see a notification dialog. After closing the dialog the program will automatically switch to the "Decision tree" tab.
- Scroll the decision tree to the top and analyze it. Navigating through decision tree labels (nodes) you will see how decision can be made starting with one field value and going down to the "class" field value.

USING RULES

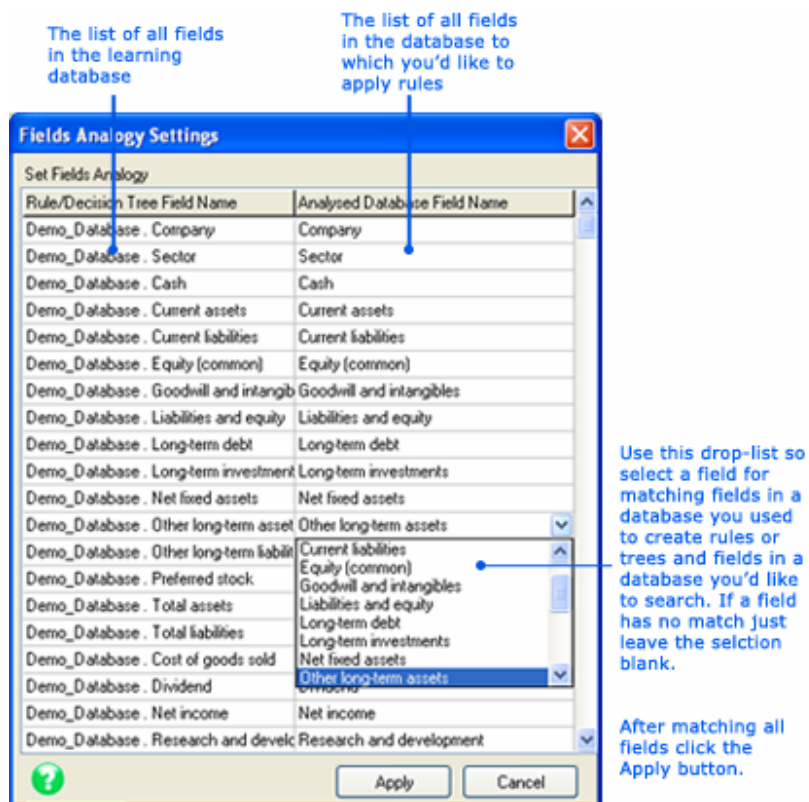
One of most useful and powerful functions of ESTARD Data Miner is ability to apply obtained rules and decision trees to databases that contain similar or same data that was used for generating rules or trees. Other powerful program function is ability to apply rules for analysis of a certain case.

Performing Searching a Database

To search a database with the help of rules or decision tree follow these steps:

- Make sure you've created statistics and rules as described in examples above, using such rule generation settings "Minimum rule probability" equal 20% and "Minimum number of cases for a rule" equal 5.
- Go to the "Rules" tab and view rules. Now sort rule details by probability and view top rules for class "Technology" – we will use the top two rules - #5 and #6.
- Go to the "Search a Database" tab.
- In program Demo version the demo database is loaded automatically on program startup. You can load the analyzed database by clicking on the "Load Database" button. The analyzed database is loaded just like the learning database.
- Click on the "Field Analogy" button. You can see here a list of fields that correspond to each other in the learning and analyzed databases. This list is filled automatically, if fields name and type coincide. In our case we are using the same database and all fields were set up automatically. If using a

database with other field names or types, use drop-down fields on the right side to find a match for a field.



- After matching all fields click the “Apply” button.
- From the toolbar on the left click the “Search” button.
- The “Setting Simple Select Query” dialog will be opened. As you can see, the dialog contains all rules generated by program. Select rules #5 and # 6 by marking them.
- Make sure the “Join Rules with OR” option is selected. Selecting this option will result in returning database records that correspond to at least one of the selected rules. Option “Join Rules with AND” will return you records that are described by all of the selected rules. If you select several rules and mark this option, most probably you won’t get any records as a result.
- Click the “Process Query” button.

- After processing the search you will notice that EDM has detected a number of records corresponding to one of the selected rules.

Performing Analysis of a Case

With the help of this feature you can easily check a new customer's profile, analyzing available customer data.

To perform analysis of a certain data case with the help of rules or decision tree follow these steps:

- Make sure you've created statistics and rules as described in examples above, using such rule generation settings "Minimum rule probability" equal 30% and "Minimum number of cases for a rule" equal 10.
- Go to "Case Analysis" tab.
- Input "50" into the "Value" field in front of "Cash".
- Click the "Start" button.
- In the "Analysis Result" list you will see a list of rules grouped by classes they belong to. Now you can see that if "Cash" equals "50" it can be referred to "Basic Materials" class.

HOW THIS INFORMATION CAN BE USED?

In the example we have completed we wanted to use our customer database to create rules that will identify customers likely to pay their invoices between 0 and 10 days. Then we wanted to use this information to identify potential prospects from our sales database who also are “responsible”, we also can identify customers who could end up in category of “irresponsible” customers. To create such customers profiles and receive the list of such customers we went through such steps:

- Loading and analysing the customers database to select fields that are best suited for analysis and rules and trees creation
- Using the statistical views for better understanding customers data through quick reporting
- Creation of rules using fields and settings that would allow us to identify the probability of failing our customers to different payment categories
- We used the most appropriate rules to search a database for deals that are likely to be closed in 10 days and deals that can end up in “late” category.

Obviously the example we’ve just completed can be a powerful way of using known data (in our case it was customers database) for identifying areas of business that are connected with high risk and could cause harm for our business. Now the key question is “How to use this information?”.

The answer to this question differs depending on your business situation and goals. It’s important to remember that there is no method that can give 100% probability for decision making, for

example we can create customers profiles that may become bad debt customers with higher probability than other customers. The best way of using such information is creating a business process that will help to minimize risks and maximize profit. For example, in the described case the CFO might control how many current deals are of high risk for the company. To minimize risks CFO could give instructions for the sales team to control the high risk deals and to prioritize other deals over these.

Regardless of your company field and strategy, predictive analysis is able to help to identify risks and give you the power to arrange business processes that will minimize your risks saving your time and money.

FREQUENTLY ASKED QUESTIONS

1. What is a Class?

A class is a unique value from the analyzed column. For example, if some column is of logical type and it contains values "True" and "False" this means it contains two classes.

2. What field should be used as the "Class field"?

The best way to receive valuable data from data mining is to analyze the field that contains key information in record. It can be of any type, but not the ID field. Besides, setting Text fields with high level of unique values will probably not return good results. For example, field "Customer name", containing names of the customers of a company, will probably contain lots of unique values. If you set this field as a Class field, you will receive Rules equal to records, which are the best descriptions of every customer. This doesn't apply to numeral fields, because there Classes are created manually, and you can create any number of classes.

3. How do I create numeral classes?

Creating a numeral class means creating intervals you want to analyze. To decide what intervals you want to analyze use the "View Field Values" button to analyze how many values are met in equal intervals. Use "View Table" button to view values met in the table selected for analysis. You also can see minimum and maximum values met in the Class field above the input fields. The intervals you create should not intersect. For example, intervals "1..10" and "11..20" do not intersect, while "1..10" and "6..20"

intersect. If you create intersecting intervals, you will be asked to change the inputted value.

4. Why do I get too many classes?

If you have too many classes, then probably you've used a text field as the Class field, because for numeric fields you will be asked to create classes by yourself and you can create as many classes as you wish. If you have too many classes in a text field, then you can split them into groups and analyze step by step. In case if this column contains numeric data, and is detected as a text column, this means that the column contains some incorrect records, containing text instead of numbers.

5. Why all classes are analyzed when I select only one class for the Statistics Query?

All classes have to be analyzed during the Statistics query, because EDM has to detect differences between them. After performing the statistics query you can select values you want to analyze further.

6. Why EDM detects empty classes or values in statistics?

For better results, it is good to use "clean" databases. Empty records are automatically detected in database, and though they are not filled, they are also been analyzed. To avoid this problem you can deselect empty values, if such are detected in analyzed column, before creating rules and trees.

7. What number of classes is recommended for use for rules and decision trees creation?

The more classes are detected in the analyzed column, the more precise results will be obtained during the statistics query. But their number shouldn't be too big, for example, several hundreds of classes used for rules creation will be analyzed much longer, than 20 classes. So the number of classes should be reasonable from the point of view of statistics and performance.

8. What is the difference between the Learning and Analyzed databases?

The Learning database is used for creating statistics, if-then rules and decision trees. The Analyzed database is not used for these operations. It is used for selecting records from the dataset with the use of rules or trees. This additional feature allows you to easily apply the results of data mining to some other database.

9. Why ID field is necessary for analyzing two or more databases together?

The ID field is not necessary if you want to analyze only one table. But it is vital for estimating which record in one table corresponds to some other record in another table. Without an ID field such relation cannot be set.

10. How values in statistics are divided into groups?

The values are automatically grouped in such way, that they detect differences between classes in the best way.

11. What is the difference between Rules and Decision Trees?

These two data mining methods allow looking at the same problem from different points of view. Rules can be represented in the form of tree, a decision tree - in a form of rules, but these two methods differ not in the way of representation. They differ in the way of obtaining. Used together, they create a good combination of models for understanding relations in data.

12. How to use settings for Rules and Decision Trees creation?

It is better to start from high values for "Probability" and "Rule cases" settings. If the number of obtained rules is low or not created at all, low down these values and create rules once again.

13. Why no Rules are created after I press "Create Rules" button?

This might happen if the "Probability" or "Rule cases" settings ("Query options" dialog) values are too high for the dataset you are working on. For example, the number of records for some class might be lower than "Rule cases", or the analyzed data might contain too many unique values, so the probability of rules might be lower than the one in the settings.

14. What algorithm is used for rules creation?

The ESTARD Data Miner algorithm used for creation of rules was specially created by our analysts. It allows obtaining ALL if-then rules, of ANY length. So the length of rules depends only on the number of fields you select for rules creation.



ESTARD

© 2004-2008 ESTARD

All rights reserved.

Reproduction or redistribution
without the expressed written
consent of ESTARD is
prohibited.

<http://www.estard.com>