Primary Approach for Fraud Detection of Health Care Using Data Mining Techniques

By
Maha Osman Mohamed Osman

B. Sc. In Computer Science Omdurmn Ahliya University 2000

A Dissertation
Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science and Information

Department of Computer Engineering
Faculty of Engineering and Technology
University of Gezira

Supervisor: Dr. Moawia Elfaki Yahia Aldaw

August 2005

Dedication

To my beloved family, who give me warmth and all their love.

……To my teacher Dr. Moawia who gives me help on knowledge and science.

……..To every one who assisted me to complete my project .

Primary Approach for Fraud Detection of Health Care Using Data Mining

Maha Osman Mohamed Osman

Master of Science in Computer Science and Information

August 2005

Department of Computer Engineering

Faculty of Engineering and Technology

Abstract

Data mining is one of the fastest growing fields in the computer industry; one of the greatest strengths of data mining is reflected in its wide range of methodologies and techniques that can be applied to a host of problem sets.

The major advantage that data mining tools have over the traditional analysis tools is that they use computer cycles to replace human cycles.

Patient health records represent comprehensive documents of the continuity of heath care and are a rich source of data for research. Data are generally accessible, accurate, and relatively inexpensive.

This research use data mining techniques to detect fraud from patients and providers .To detects those fraud we use combination of *PolyAnalyst* techniques.

We detect fraud from patients by performing analysis to the database and then create link chart, also we can detect fraud from providers by Transition Basket Analysis.

By performing those techniques we minimize the frauds from both patients and providers.

طريقة أولية لاكتشاف الحيل في البيانات الطبية باستخدام التنقيب عن البيانات

مها عثمان محمد عثمان

ماجستير العلوم في علوم الحاسوب والمعلومات ، أغسطس 2005

قسم هندسة الحاسوب

كلية الهندسة والتكنلوجيا

المستخلص

يعتبر تنقيب البينات من أحد المجالات سريعة التطور في مجال الحاسوب . تظهر قوة وفائدة تعدين البينات في المدى الواسع للطرق والتقنيات التي تطبق في حل المشاكل المتعلقة بالبيانات.

تأتي الفائدة الكبرى من استخدام اداوت تنقيب البيانات في أنها زيادة إلى طرق التحليل المعتادة تمكن الحاسوب من أن يحل محل دور الإنسان في التحليل.

تمثل ملفات المرضي مصدر غني بالبيانات للباحثين . هذه البيانات عموماً تكون صحيحة ، ودقيقة ، وغير مكلفة نسبياً.

في هذا البحث استعملنا أحد طرق التنقيب عن البيانات لاكتشاف بعض الحيل من قبل المرضي أو الجهات التي تقوم بدعم المرضي في مجال العلاج .ولاكتشاف هذه الحيل استعملنا مجموعة ادوات لبرنامج البولي انلست لاكتشاف الحيل .

استخدمنا ادوات برنامج البولي انلست لاكتشاف الحيل من قبل المرضي ، حيث قمنا بتحليل البيانات ثم قمنا بعمل لينك جارت. ولاكتشاف الحيل من قبل الجهات التي تقوم بدعم المرضي مادياً في العلاج قمنا باستخدام ترانس اكشن باسكت انليسس .

وباستخدام هذه التقنيات يمكننا أن نقلص الحيل المالية من قبل المرضي والممولين الذين يقومون بدعم المرضي.

Table of Contents

List of Tables

List of Figures

Chapter One

Introduction to Data Mining

1.1 Data Mining Overview

Data mining is the search for new, valuable, and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers.

Data mining is one of the fastest growing fields in the computer industry; one of the greatest strengths of data mining is reflected in its wide range of methodologies and techniques that can be applied to a host of problem sets.

Data mining, the extraction of hidden predictive information from database, is powerful new technology with great potential to help companies focus on the most important information.

There are two primary goals of Data Mining: prediction and description.

Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put data-mining activities into one of two categories:

- Predictive data mining, which produces the model of the system described by the given data set, or
- Descriptive data mining, with produces new, nontrivial information based on the available data set

On the predictive end of spectrum, the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation, or other similar tasks.

On the other descriptive, end of the spectrum the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets. The relative importance of prediction and description for particular data mining applications can vary considerable.

The goals of prediction and description are achieved by using data mining techniques [1].

1.2 Healthcare Applications

Applications of Data Mining in healthcare industry are widespread. One way Data Mining is helping healthcare providers cut costs and improve care by showing which treatments statistically have been most effective.

Patient health records represent comprehensive documents of the continuity of healthcare and are a rich source of data for research. The advantages of using healthcare records as a source of information are accurate and timely data, rich clinical detail and dates attached to data elements.

1.3 Techniques of Data Mining

Commonly used techniques in data mining are artificial neural networks, decision trees, genetic algorithms, method, and rule induction. Classification queries use decision variables or examples to partition data into subclasses. Characterization queries derive common features of a class regardless of the characteristics of other classes. An association query discovers associations among values grouped by selection phrase with a user specified minimum support requirement. Clustering queries partition data of a relational table with members of each cluster sharing a number of properties [2].

Five common types of information yielded by Data Mining are: association, sequences, classifications, clusters, and forecasting.

1.4 Objectives

In our project we will be searching for traces of two possible healthcare fraud mechanisms. By using a combination of PolyAnalyst data mining techniques we will obtain valuable results.

Also we can use combination of PolyAnalyst data mining techniques to analyze our data.

1.5 Plan of Study

- Introduction to Data Mining.
- Description of Data Mining concept.
- Description of PolyAnalyst software.
- Implementation using Data mining techniques (PolyAnalyst)

Chapter Two

Data Mining Concepts

2.1 Data Mining Concepts

Progress in digital data acquisition and storage technology has resulted in the growth of huge database. This has occurred in all areas of human endeavor from the mundane (such as supermarket transaction data credit card usage record) to more exotic (such as images of astronomical bodies and medical record)[6].

Data storage became easier as the availability of large amounts of computing power at low cost i.e. the cost of processing power and storage is falling, made data cheap. There was also the introduction of new machine learning methods for knowledge representation based on logic programming etc. in addition to traditional statistical analysis of data. The new methods tend to be computationally intensive hence a demand for more processing power [9].

Data mining is related to the sub area of statistics called *exploratory* data analysis, which has similar goals and relies on statistical measure, it is also closely related to the sub areas of artificial intelligence called knowledge discovery in data base (KDD) and machine learning.

The important distinguish characteristics of data mining is that the volume of data is very large. The large volume of data is used in order to guide decisions about future activities.

> Data mining consists of finding interesting trends or patterns in large datasets. The general exception that mining tools should be able to identify these patterns in the data with minimal user input. The pattern identified by such tools can give a data analyst useful and unexpected insights that can be more carefully investigated subsequently perhaps using other decision support tools [3] .

Data mining is an iterative process within which progress is defined by discovery, through either automatic or manual methods. Data mining

is the search for new, valuable, and nontrivial information in large volume of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the capabilities of computers.

The success of a Data Miningg depends largely on the amount of energy, knowledge, and creativity that the designer puts into it. Data Miningg is one of the growing fields in the computer industry [1].

The objective of data mining is to extract valuable information from your data to discover the hidden gold. This gold is the information.

Many traditional reporting and query tools and statistical analysis system use the term "data mining" in their product descriptions. Artificial Inelegance – based system are also being touted as new data mining tools [5].

The field that has come to be called data mining has grown from several antecedents. On the academic side are machine learning and statistics. Machine learning has contributed important algorithms for recognizing patterns in data. Machine-learning researchers are on the bleeding edge, conjuring ideas about how to make computers think. Statistics is another important area that provides background for data mining. Statisticians offer mathematical rigor; not only do they understand the algorithms; they understand the best practices in modeling and experimental design.The final thread is decision support.

Over the past few decades, people have been gathering data into databases to make better-informed decisions. Data mining is a natural extension of this effort [4].

Basically data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in set of data. It the computer, which is responsible for finding the patterns by identifying the underlying, rules and features in data.

Data mining analysis tends to work from the data up and the best techniques are those developed with an orientation towards large volumes

of data, making use of as much of the collected data as possible to arrive at reliable conclusions and decisions. The analysis process starts with a set of data, uses a methodology to develop an optimal representation of the structure of the data during which time knowledge is acquired.

Once knowledge has been acquired this can be extended to larger set of data working on assumption that the larger data set has a structure similar to the sample data [9].

2.2 The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database for example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- **Automated Prediction of Trends and Behaviors**. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

- **Automated Discovery of Previously Unknown Patterns**. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes.

Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

Databases Can be Larger in Both Depth and Breadth

- More Columns: Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.

- More Rows: Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

2.3 What is Data Mining

The term data mining has been stretched beyond its limits to apply to any form of data analysis. Some of the numerous definitions of Data Mining, or Knowledge Discovery in Databases are:

- Is the process of exploration and analysis, large quantities of data in order to discover meaningful patterns and rules [4].

- The entire process of applying a computer-based methodology, including new techniques, for discovering knowledge from data is called data mining [1].

- Is the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [6].

- Data Mining is a process of discovering various models, summaries, and derived values from a given collection of data [1].

- Data mining is the search for relationships and global patterns that exist in large databases but are `hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database [9].

2.4 What is Data Mining Good For

Data mining software allows users to analyze large databases to solve business decision problems. Data mining is, in some ways, an extension of statistics, with a few artificial intelligence and machine learning twists thrown in. Like statistics, data mining is not a business solution, it is just a technology. For example, consider a catalog retailer who needs to decide who should receive information about a new product. The information operated on by the data mining process is contained in a historical database of previous interactions with customers and the features associated with the customers, such as age, zip code, and their responses.

The data mining software would use this historical information to build a model of customer behavior that could be used to predict which customers would be likely to respond to the new product. By using this information a marketing manager can select only the customers who are most likely to respond. The operational business software can then feed the results of the decision to the appropriate touch point systems (call centers, direct mail, web servers, email systems, etc.) so that the right customers receive the right offers.

2.5 Data Mining Can Solve Most Difficult Problems

Every mature business has some seemingly insoluble problems-difficult judgments about customers, resource allocation, business strategy, or organization. The intractable residue that remains after years of work has solved the other 95 percent. These five-percent problems are the principal targets of data mining. Most of today's enterprise problems are no more difficult than those of fifty years ago.

Managers in the 1950s faced many of the same problems that decision-makers face today. But, there is one very important difference-scale. In terms of the number of customers, variety of products, array of marketing channels, speed of commerce, churn, fraud, etc., everything today is much bigger. So big, that unaided human decision-making processes are losing their ability to keep up. There may be nothing wrong with the processes themselves, other than their inability to scale up.

Further complicating the modern decision-maker's problem is a reduction in the time available for deliberation. Data mining techniques can be used to learn about the factors bearing on a decision and construct an application that uses those factors to help the enterprise make those decisions in an objective, consistent way.

These techniques and methods formalize important knowledge about how the enterprise operates. Good decision-makers are often successful because of the knowledge they possess. If intelligence is the engine, then knowledge is the fuel [4].

2.6 What Can Data Mining Do

We can use the term for a specific set of activities, all of which involve extracting meaningful new information from the data. The six activities are:

1. Classification.
2. Estimations.
3. Prediction.
4. Affinity grouping or association rules.
5. Clustering.
6. Description and Visualization.

The first three tasks—classification, estimation, and prediction—are all examples of directed data mining. In directed data mining, the goal is to use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data.

The next three tasks are examples of undirected data mining. In undirected data mining, no variable is singled out as the target; the goal is to establish some relationship among all the variables.

2.6.1 Classification

Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The objects to be classified are generally represented by records in a database. The act of classification consists of updating each record by filling in a field with a class code.

The classification task is characterized by a well-defined definition of the classes, and a training set consisting of preclassified examples. The task is to build a model that can be applied to unclassified data in order to classify it.

Examples of Classification Tasks

- Classifying credit applicants as low, medium, or high risk.

- Determining which home telephone lines is used for Internet access.

In all of these examples, there are a limited number of already-known classes and we expect to be able to assign any record into one or another of them.

2.6.2 Estimation

Classification deals with discrete outcomes: yes or no, debit card, mortgage, or car loan. Estimation deals with continuously valued outcomes .we use estimation to come up with a value for some unknown continuous variable such as income, height, or credit card balance.

In practice, estimation is often used to perform a classification task. Often classification and estimation are used together.

Example of Estimation Tasks

- Estimating the number of children in a family.
- Estimating a family's total household income.

2.6.3 Prediction

There should not be a separate heading for prediction. Any prediction can be thought of as classification or estimation.

Predictive tasks feel different because the records are classified according to some predicted future behavior or estimated future value. With prediction, the only way to check the accuracy of the classification is to wait and see.

Example of Prediction Tasks

- Predicting which customers will leave within the next six months

Any of the techniques used for classification and estimation can be adapted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior.

2.6.4 Affinity Grouping or Association Rules

The task of affinity grouping is to determine which things go together. The prototypical example is determining what things go together in a shopping cart at the supermarket. Affinity grouping can also be used to identify cross-selling opportunities and to design attractive packages or groupings of products and services.

2.6.5 Clustering

Clustering is the task of segmenting a diverse group into a number of more similar subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes.

In clustering, there are no predefined classes and no examples. The records are grouped together on the basis of self-similarity. It is up to the miner to determine what meaning, if any, to attach to the resulting clusters. Clustering is often done as a prelude to some other form of data mining or modeling.

For example, clustering might be the first step in a market segmentation effort.

2.6.6 Description and Visualization

Sometimes the purpose of data mining is simply to describe what is going on in a complicated database in a way that increases our understanding of the people, products, or processes that produced the data in the first place.

A good enough description of a behavior will often suggest an explanation for it as well. At the very least, a good description suggests where to start looking for an explanation. Data visualization is one powerful form of descriptive data mining.

It is not always easy to come up with meaningful visualizations, but the right picture really can be worth a thousand association rules since human beings are extremely practiced at extracting meaning from visual scenes.

2.7 Benefits of Data Mining

Data mining offers three major advantages to the enterprise:

1- It provides information about business processing.
2- It takes advantages of data that may already be available in operational data collection, data mart & the data warehouse.
3- It provides patterns of behavior, reflect in data , that can drive the accumulation of business knowledge and the ability to foresee and shape future value [8].

2.8 Who Benefits from Data Mining

Organizations who are most likely to benefit from Data Mining:

- Have large volumes of data;
- Have communities of knowledge workers that need to understand data, but are not trained as statisticians;
- Have organizational data which is complex in nature; i.e., detailed and multifaceted, with complex data relationships; and
- Exist in competitive markets.

2.9 Variable Types of Data Mining

Variables come in a variety of types that can be distinguished by the amount of information that they encode. They are briefly reviewed here starting with the "simplest" (those that carry the least information) to those that carry the most information.

a- Nominal Variables: Essentially, these are no more than labels identifying unique entities. Personal names are nominal labels identifying unique individuals. So too are order numbers, serial numbers, tracking codes, and many other similar labels.

b- **Categorical variables**: These are group labels identifying groups of entities sharing some set of characteristics implied by the category. In addition to personal names, all readers of this book belong to the category of humans. (Apologies to any non - human readers for making unwarranted assumptions!)

c- **Ordinal Variables**: These are categories that can be rationally listed in some order. Examples of such categories might include small, medium, and large or hot, warm, tepid, cool, and cold. Notice that neither nominal nor categorical variables can be ordered; they are simply unordered labels for single entities and groups of entities, respectively.

d- **Interval Variables**: These are ordinal variables in which it is possible to determine a distance between the ordered categories. However, their intervals may well be arbitrary, as in a temperature scale. Additive distances between equidistant points are meaningful, but ratios aren't. For instance, there are 10 degrees between 20 and 30 degrees, and between 110 and 120 degrees. However, 50 degrees isn't twice as hot as 25 degrees on either Celsius or Fahrenheit scales.

e- **Ratio Variables**: These are interval variables in which ratios are valid, and which have a true zero point. An example is a bank account. The zero point is an empty account. The ratio between $10 and $20 is 1 to 2, and $20 is twice $10. The ratio between $100 and $200 is also 1 to 2, and $200 is twice $100. Same ratio—same relationship.

2.10 Data Mining Process

It is important to realize that the problem of discovering or estimating dependencies from data or discovering totally new data is only one part of the general experimental procedure used by scientists, engineers, and others who apply standard steps to draw conclusions from the data. The general experimental procedure adapted to data-mining problems involves the following steps:

2.10.1 State the Problem and Formulate the Hypothesis

Most database modeling studies are performed in a particular application domain. Hence, domain-specific knowledge and experience are usually necessary in order to come up with a meaningful problem statement.

Unfortunately, many application studies tend to focus on the data-mining technique at the expense of a clear problem statement. In this step, a modeler usually specifies a set of variables for the unknown dependency and, if possible, a general form of this dependency as an initial hypothesis. There may be several hypotheses formulated for a single problem at this stage.

The first step requires the combined expertise of an application domain and a data-mining model. In practice, it usually means a close interaction between the data-mining expert and the application expert.

2.10.2 Collect the Data

This step is concerned with how the data are generated and collected. In general, there are two distinct possibilities. The first is when the data-generation process is under the control of an expert (modeler): this approach is known as a designed experiment. The second possibility is when the expert cannot influence the data-generation process: this is known as the observational approach.

An observational setting, namely, random data generation, is assumed in most data-mining applications.

It is very important, however, to understand how data collection affects its theoretical distribution. Also, it is important to make sure that the data used for estimating a model and the data used later for testing and applying a model come from the same, unknown, sampling distribution. If this is not the case, the estimated model cannot be successfully used in a final application of the results.

2.10.3 Preprocessing the Data

In the observational setting, data are usually "collected" from the existing databases, data warehouses, and data marts.

1. Outlier detection (and removal) – Outliers are unusual data values that are not consistent with most observations. Commonly, outliers result from measurement errors, coding and recording errors, and, sometimes, are natural, abnormal values. Such non-representative samples can seriously affect the model produced later. There are two strategies for dealing with outliers:

One : Detect and eventually remove outliers as a part of the preprocessing phase, or

Two : Develop robust modeling methods that are insensitive to outliers.

2. Scaling, encoding, and selecting features – Data preprocessing includes several steps such as variable scaling and different types of encoding. For example, one feature with the range [0, 1] and the other with the range [−100, 1000] will not have the same weights in the applied technique; they will also influence the final data-mining results differently.

Therefore, it is recommended to scale them and bring both features to the same weight for further analysis. Also, application-specific encoding methods usually achieve dimensionality reduction by providing a smaller number of informative features for subsequent data modeling.

These two classes of preprocessing tasks are only illustrative examples of a large spectrum of preprocessing activities in a Data Mining process.

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

2.10.4 Estimate the Model

The selection and implementation of the appropriate data-mining technique is the main task in this phase. This process is not straightforward; usually, in practice, the implementation is based on several models, and selecting the best one is an additional task.

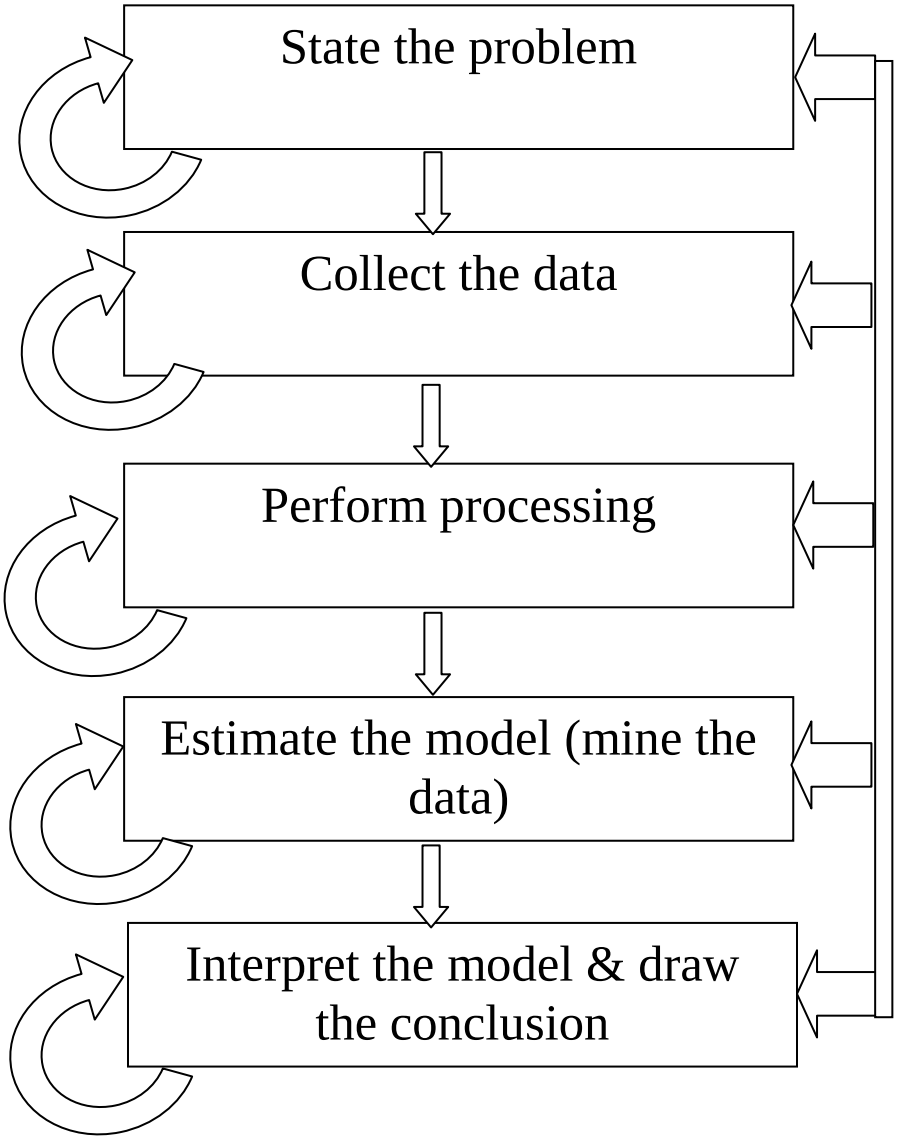
2.10.5 Interpret the Model and Draw Conclusions

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

Usually, simple models are more interpretable, but they are also less accurate. Modern data-mining methods are expected to yield highly accurate results using high dimensional models.

The problem of interpreting these models, also very important, is considered a separate task, with specific techniques to validate the results.

A user does not want hundreds of pages of numeric results. He does not understand them; he cannot summarize, interpret, and use them for successful decision-making.[1]



Figer (2.1) Data Mining Process

Data mining research has drawn on a number of other fields such as inductive learning, machine learning and statistics etc.

2.11 Data Mining & Inductive Learning

Induction is the inference of information from data and inductive learning is the model building process where the environment i.e. database is analyzed with a view to finding patterns. Similar objects are grouped in classes and rules formulated whereby it is possible to predict the class of unseen objects. This process of classification identifies classes such that each class has a unique pattern of values, which forms the class description. The nature of the environment is dynamic hence the model must be adaptive i.e. should be able learn.

Inductive learning where the system infers knowledge itself from observing its environment has two main strategies:

- Supervised learning - this is learning from examples where a teacher helps the system construct a model by defining classes and supplying examples of each class. The system has to find a description of each class i.e. the common properties in the examples. Once the description has been formulated the description and the class form a classification rule which can be used to predict the class of previously unseen objects. This is similar to discriminate analysis as in statistics.
- Unsupervised learning - this is learning from observation and discovery. The data mine system is supplied with objects but no classes are defined so it has to observe the examples and recognize patterns (i.e. class description) by it self. This system results in a set of class descriptions, one for each class discovered in the environment. Again this similar to cluster analysis as in statistics.

Induction is therefore the extraction of patterns. The quality of the model produced by inductive learning methods is such that the model could be used to predict the outcome of future situations [9].

2.12 Data Mining and Machine Learning

The machine-learning people come from the computer science and artificial intelligence worlds. They have focused their efforts on getting computers to display intelligence. In particular, the machine learning community is interested in writing computer programs that are capable of learning by example. The first kind

of learning manifests itself by a newfound ability to perform some task such as balancing a broom handle or recognizing written characters.

Machine learning is the automation of a learning process and learning is tantamount to the construction of rules based on observations of environmental states and transitions. This is a broad field, which includes not only learning from examples, but also reinforcement learning, learning with teacher, etc. A learning algorithm takes the data set and its accompanying information as input and returns a statement e.g. a concept representing the results of learning as output. Machine learning examines previous examples and their outcomes and learns how to reproduce these and make generalizations about new cases.

Generally a machine learning system does not use single observations of its environment but an entire finite set called the training set at once. This set contains examples i.e. observations coded in some machine-readable form. The training set is finite hence not all concepts can be learned exactly.

The term, data mining was first used by people who took the methods of the machine learning and began to apply them to fields outside of computer science and Artificial Intelligence (AI)—fields such as industrial process control and direct marketing.

The choice of the term data mining for the new, business-oriented applications of AI research shows how little overlap there was between this group and the statisticians, actuaries, and economists who had long been doing predictive modeling. For the latter group, the term "data mining" meant searching for data to support a particular point of view rather than letting the facts speak for themselves. The data miners were smart people getting good results, but they were not mathematicians.

2.13 Data Mining and Statistics

Statistical techniques alone may not be sufficient to address some of the more challenging issues in Data mining, especially those arising from massive data set. Nonetheless, statistics plays a very important role in Data mining it is necessary component in any Data mining enterprise.

With large data sets (and particularly with very large data sets) we may simply not know even straightforward fact about the data. Simple eye-balling of the data is not an option. This means that sophisticated search and examination methods my be required to illuminate features which would be readily apparent in small data set.

The most fundamental difference between classical statistical applications and Data mining is the size of data set. To a conventional statistician, a 'large' data set may contain a few hundred or a thousand data points.

To someone concerned with Data mining, however, many millions or even billions of data points are not unexpected – gigabyte and even terabyte database are by no means uncommon.

Massive data sets can be tackled by sampling or by adaptive methods, or by summarizing the record in terms of sufficient statistics.

Further difficulties arise when there are many variables. One that is important in some context is the curse of dimensionality, the exponential rate of growth of the number of unit cells in a space as the number of variables increases.

The curse of dimensionality manifestation itself in the difficulty of finding accurate estimations of probability densities in high dimensional space without astronomically large databases.

Various problem arise from the difficulties of accessing very large data sets [6].

Statistics has another important thread that has supported data mining. For centuries, people have used statistical techniques to understand the natural world. These have included predictive algorithms (which statisticians call regression), sampling methodologies, and experimental design. Now, they are applying these techniques to the business world [4].

There are verities of statistical methods used in data mining and it is not data mining tools. Statistical tools are widely used in science and industry and provide excellent features for describing and visualizing large data.

Statistical analysis is often a good in understanding data (it could be the first step to understand the data).

2.14 Data Mining and Decision Support

Decision support is a broad term for the entire information technology infrastructure that companies and other organizations use to make informed decisions. The term covers both relational and dimensional databases used for decision support. Decision support systems contrast with online transaction processing systems (OLTP). OLTP databases are designed to process large numbers of transactions very fast. In database terminology, a transaction is a complete action that must either finish successfully or appear not to have happened

Decision support databases have very different requirements. In decision support, it is rarely useful to look at individual records. Decision support databases are designed to support complex queries such as "Which customers spent more than $100 at a restaurant more than 100 miles from home in two of the last three months?" To answer this question, you need to be able to translate each customer's address into geographic coordinates and do the same for the restaurants before beginning to aggregate the charges.

2.15 Data Mining Applications

Data mining has many and varied fields of application some of which are listed below.

2.15.1 Retail/Marketing

- Identify buying patterns from customers
- Find associations among customer demographic characteristics
- Predict response to mailing campaigns
- Market basket analysis

2.15.2 Banking

- Detect patterns of fraudulent credit card use

- Identify `loyal' customers
- Predict customers likely to change their credit card affiliation
- Determine credit card spending by customer groups
- Find hidden correlations between different financial indicators
- Identify stock trading rules from historical market data

2.15.3 Insurance and Health Care

- Claims analysis - i.e which medical procedures are claimed together
- Predict which customers will buy new policies
- Identify behavior patterns of risky customers
- Identify fraudulent behavior

2.15.4 Transportation

- Determine the distribution schedules among outlets
- Analyse loading patterns

2.15.5 Medicines

- Characterise patient behavior to predict office visits
- Identify successful medical therapies for different illnesses [9].

2.16 Data Mining Problems & Issues

Data mining systems rely on databases to supply the raw data for input and these raises problems in that databases tend be dynamic, incomplete, noisy, and large. Other problems arise as a result of the adequacy and relevance of the information stored.

2.16.1 Limited Information

A database is often designed for purposes different from data mining and sometimes the properties or attributes that would simplify the learning task are not present nor can they be requested from the real world. Inconclusive data causes problems because if some attributes essential to knowledge about the application domain are not present in the data it may be impossible to discover significant knowledge about a given domain. For example cannot diagnose malaria from a patient database if that database does not contain the patients red blood cell count.

2.16.2 Noise and Missing Values

Databases are usually contaminated by errors so it cannot be assumed that the data they contain is entirely correct. Attributes, which rely on subjective or measurement judgements, can give rise to errors such that some examples may even be mis-classified. Error in either the values of attributes or class information are known as noise. Obviously

where possible it is desirable to eliminate noise from the classification information as this affects the overall accuracy of the generated rules.

Missing data can be treated by discovery systems in a number of ways such as;

- simply disregard missing values
- omit the corresponding records
- infer missing values from known values
- treat missing data as a special value to be included additionally in the attribute domain
- or average over the missing values using Bayesian techniques.

Noisy data in the sense of being imprecise is characteristic of all data collection and typically fit a regular statistical distribution such as Gaussian while wrong values are data entry errors. Statistical methods can treat problems of noisy data, and separate different types of noise.

2.16.3 Uncertainty

Uncertainty refers to the severity of the error and the degree of noise in the data. Data precision is an important consideration in a discovery system.

2.16.4 Size, Updates, and Irrelevant Fields

Databases tend to be large and dynamic in that their contents are ever-changing as information is added, modified or removed. The problem with this from the data mining perspective is how to ensure that the rules are up-to-date and consistent with the most current information. Also the learning system has to be time-sensitive as some data values vary over time and the discovery system is affected by the `timeliness' of the data.

Another issue is the relevance or irrelevance of the fields in the database to the current focus of discovery for example post codes are fundamental to any studies trying to establish a geographical connection to an item of interest such as the sales of a product.

Chapter Three

Data Mining Techniques

3.1 The Foundations of Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery.

Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.

3.2 Data Mining Techniques

We have broken the discussion into two sections, each with a specific theme:

- Classical Techniques: Statistics, Neighborhoods and Clustering
- Next Generation Techniques: Trees, Networks and Rules

3.2.1 Classical Techniques: Statistics, Neighborhoods and Clustering

The Classics

These two sections have been broken up based on when the data mining technique was developed and when it became technically mature enough to be used for business, especially for aiding in the optimization of customer relationship management systems. Thus this section contains descriptions of techniques that have classically been used for decades the next section represents techniques that have only been widely used since the early 1980s.

3.2.1.1 Statistics

By strict definition "statistics" or statistical techniques are not data mining. They were being used long before the term data mining was coined to apply to business applications. However, statistical techniques are driven by the data and are used to discover patterns and build predictive models. And from the users perspective you will be faced with a conscious choice when solving a "data mining" problem as to whether you wish to attack it with statistical methods or other data mining techniques.

For this reason it is important to have some idea of how statistical techniques work and how they can be applied.

- What is Statistics

Statistics is a branch of mathematics concerning the collection and the description of data. Usually statistics is considered to be one of those scary topics in college right up there with chemistry and physics. However, statistics is probably a much friendlier branch of mathematics because it really can be used every day. Statistics was in fact born from very humble beginnings of real world problems from business, biology, and gambling!

Knowing statistics will help the average business person make better decisions by allowing them to figure out risk and uncertainty when all the facts either aren't known or can't be collected. Even with all the data stored in the largest of data warehouses business decisions still just become more informed guesses. The more and better the data and the better the understanding of statistics the better the decision that can be made.

- Linear Regression

In statistics prediction is usually synonymous with regression of some form. There are a variety of different types of regression in statistics but the basic idea is that a model is created that maps values from predictors in such a way that the lowest error occurs in making a prediction. The simplest form of regression is simple linear regression that just contains one predictor and a prediction.

The relationship between the two can be mapped on a two dimensional space and the records plotted for the prediction values along the Y-axis and the predictor values along the X-axis. The simple linear regression model then could be viewed as the line that minimized the error rate between the actual prediction value and the point on the line (the prediction from the model).

3.2.1.2 Nearest Neighbor

Clustering and the Nearest Neighbor prediction technique are among the oldest techniques used in data mining. Most people have an intuition that they understand what clustering is - namely that like records are grouped or clustered together.

Nearest neighbor is a prediction technique that is quite similar to clustering - its essence is that in order to predict what a prediction value is in one record look for records with similar predictor values in the historical database and use the prediction value from the record that it "nearest" to the unclassified record.

3.2.1.3 Clustering

Clustering for Clarity

Clustering is the method by which like records are grouped together. Usually this is done to give the end user a high level view of what is going on in the database. Clustering is sometimes used to mean segmentation - which most marketing people will tell you is useful for coming up with a birds eye view of the business

A Simple Example of Clustering

A simple example of clustering would be the clustering that most people perform when they do the laundry - grouping the permanent press, dry cleaning, whites and brightly colored clothes is important because they have similar characteristics. And it turns out they have important attributes in common about the way they behave (and can be ruined) in the wash. To "cluster" your laundry most of your decisions are relatively straightforward. There are of course difficult decisions to be made about which cluster your white shirt with red stripes goes into (since it is mostly white but has some color and is permanent press).

When clustering is used in business the clusters are often much more dynamic - even changing weekly to monthly and many more of the decisions concerning which cluster a record falls into can be difficult.

What is the Difference Between Clustering and Nearest Neighbor Prediction

The main distinction between clustering and the nearest neighbor technique is that clustering is what is called an unsupervised learning technique and nearest neighbor is generally used for prediction or a supervised learning technique.

Unsupervised learning techniques are unsupervised in the sense that when they are run there is not particular reason for the creation of the models the way there is for supervised learning techniques that are trying to perform prediction. In prediction, the patterns that are found in the database and presented in the model are always the most important patterns in the database for performing some particular prediction.

In clustering there is no particular sense of why certain records are near to each other or why they all fall into the same cluster. Some of the differences between clustering and nearest neighbor prediction can be summarized in Table below.

Table (3.1) Diffrences Between Clastering and Nerest Neighbor

Nearest Neighbor	Clustering

Used for prediction as well as consolidation.	Used mostly for consolidating data into a high-level view and general grouping of records into like behaviors.
Space is defined by the problem to be solved (supervised learning).	Space is defined as default n-dimensional space, or is defined by the user, or is a predefined space driven by past experience (unsupervised learning).
Generally only uses distance metrics to determine nearness.	Can use other metrics besides distance to determine nearness of two records - for example linking two points together.

There is no particular rule that would tell you when to choose a particular technique over another one. Sometimes those decisions are made relatively arbitrarily based on the availability of data mining analysts who are most experienced in one technique over another. And even choosing classical techniques over some of the newer techniques is more dependent on the availability of good tools and good analysts.

3.2.2 Next Generation Techniques: Trees, Networks and Rules

The Next Generation

The data mining techniques in this section represent the most often used techniques that have been developed over the last two decades of research. They also represent the vast majority of the techniques that are being spoken about when data mining is mentioned in the popular press. These techniques can be used for either discovering new information within large databases or for building predictive models. Though the older decision tree techniques such as CHAID are currently highly used the new techniques such as CART are gaining wider acceptance.

3.2.2.1 Decision Trees

- What is a Decision Tree

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification

- Where Can Decision Trees be Used

Decision trees are data mining technology that has been around in a form very similar to the technology of today for almost twenty years now and early versions of the algorithms date back in the 1960s. Often times these techniques were originally developed for statisticians to automate the process of determining which fields in their database were actually useful or correlated with the particular problem that they were trying to understand.

Partially because of this history, decision tree algorithms tend to automate the entire process of hypothesis generation and then validation much more completely and in a much more integrated way than any other data mining techniques. They are also particularly adept at handling raw data with little or no pre-processing. Perhaps also because they were originally developed to mimic the way an analyst interactively performs data mining they provide a simple to understand predictive model based on rules (such as "90% of the time credit card customers of less than 3 months who max out their credit limit are going to default on their credit card loan.").

- Decision Tress for Prediction

Although some forms of decision trees were initially developed as exploratory tools to refine and preprocess data for more standard statistical techniques like logistic regression. They have also been used and more increasingly often being used for prediction. This is interesting because many statisticians will still use decision trees for exploratory analysis effectively building a predictive model as a by product but then ignore the predictive model in favor of techniques that they are most comfortable with.

Sometimes veteran analysts will do this even excluding the predictive model when it is superior to that produced by other techniques. With a host of new products and skilled users now appearing this tendency to use decision trees only for exploration now seems to be changing

- ID3 and an Enhancement C4.5

In the late 1970s J. Ross Quinlan introduced a decision tree algorithm named ID3. It was one of the first decision tree algorithms yet at the same time built solidly on work that had been done on inference systems and concept learning systems from that decade as well as the preceding decade.

Initially ID3 was used for tasks such as learning good game playing strategies for chess end games. Since then ID3 has been applied to a wide variety of problems in both academia and industry and has been modified, improved and borrowed from many times over.

ID3 picks predictors and their splitting values based on the gain in information that the split or splits provide

ID3 was later enhanced in the version called C4.5. C4.5 improves on ID3 in several important areas:

- predictors with missing values can still be used
- predictors with continuous values can be used.
- pruning is introduced.
- rule derivation .

3.2.2.2 Neural Networks

- What is a Neural Network

When data mining algorithms are talked about these days most of the time people are talking about either decision trees or neural networks. Of the two neural networks have probably been of greater interest through the formative stages of data mining technology. Neural networks do have disadvantages that can be limiting in their ease of use and ease of deployment, but they do also have some significant advantages. Foremost among these advantages is their highly accurate predictive models that can be applied across a large number of different types of problems.

- Where to Use Neural Networks

Neural networks are used in a wide variety of applications. They have been used in all facets of business from detecting the fraudulent use of credit cards and credit risk prediction to increasing the hit rate of targeted mailings. They also have a long history of application in other areas such as the military for the automated driving of an unmanned vehicle at 30 miles per hour on paved roads to biological simulations such as learning the correct pronunciation of English words from written text.

What Does a Neural Net Look Like

A neural network is loosely based on how some people believe that the human brain is organized and how it learns. Given that there are two main structures of consequence in the neural network:

The node - which loosely corresponds to the neuron in the human brain.

The link - which loosely corresponds to the connections between neurons (axons, dendrites and synapses) in the human brain.

- How does a Neural Networks Make a Prediction

In order to make a prediction the neural network accepts the values for the predictors on what are called the input nodes. These become the values for those nodes those values are then multiplied by values that are stored in the links (sometimes called links and in some ways similar to the weights that were applied to predictors in the nearest neighbor method).

These values are then added together at the node at the far right (the output node) a special thresholding function is applied and the resulting number is the prediction. In this case if the resulting number is 0 the record is considered to be a good credit risk (no default) if the number is 1 the record is considered to be a bad credit risk (likely default).

3.2.2.3 Rule Induction

Rule induction is one of the major forms of data mining and is perhaps the most common form of knowledge discovery in unsupervised learning systems. It is also perhaps the form of data mining that most closely resembles the process that most

people think about when they think about data mining, namely “mining” for gold through a vast database

Rule induction on a data base can be a massive undertaking where all possible patterns are systematically pulled out of the data and then an accuracy and significance are added to them that tell the user how strong the pattern is and how likely it is to occur again. In general these rules are relatively simple such as for a market basket database of items scanned in a consumer market basket you might find interesting correlations in your database such as:

- If bagels are purchased then cream cheese is purchased 90% of the time and this pattern occurs in 3% of all shopping baskets.
- If live plants are purchased from a hardware store then plant fertilizer is purchased 60% of the time and these two items are bought together in 6% of the shopping baskets [10].

3.3 Overview on Health Care

Healthcare and information technology has made huge strides over the years. Information Technology includes data mining in which healthcare workers storing patient health information and sharing it with each other as well as for insurance and billing purposes.

Patients can now look up their own diseases and symptoms to help educate themselves before going to the doctor or after for a broader understanding. There are support groups, newsletters, medical libraries and wealth of other information available at the click of a button. These benefits do not come without the risk of private patient information possibly being mishandled and falling into the wrong hands.

3.4 Data Mining In Health Care

Data mining has often been defined as a "process of extracting previously unknown, valid and actionable information from large databases and then using the information to make critical business decisions".

This definition is based on the premise that an organization can link its myriad sources of data into a data warehouse (to potentially include data marts). Further, these data sources can evolve to a higher degree of analyses to include exploration using on-line analytical processing (OLAP), statistical analyses, and querying.

In general, the rationale for mining data is a function of organizational needs and type of firm's role in the industry. Moreover, researchers have established that data-mining applications, when

implemented effectively, can result in strategic planning and competitive advantage. In an effort to minimize the collection and storage of useless and vast amounts of data, mining can identify and monitor which data are most critical to an organization – thereby providing efficient collection, storage, and exploration of data [7].

In particular, data mining offers the health care industry the capabilities to tackle imminent challenges germane to its domain.

With the amount of information and issues in the healthcare industry, not to mention the pharmaceutical industry and biomedical research, opportunities for data-mining applications are extremely widespread, and benefits from the results are enormous. Storing patients' records in electronic format and the development in medical-information systems cause a large amount of clinical data to be available online.

Regularities, trends, and surprising events extracted from these data by data-mining methods are important in assisting clinicians to make informed decisions, thereby improving health services.

The healthcare industry faces contradictory pressures of lowering cost and increasing quality of service, both of which require efficient decision-making. These business challenges of healthcare delivery require greater operational efficiencies and the tools necessary to provide real-time access to information. Healthcare facilities have at their disposal vast amounts of data from administrative and clinical databases. Capabilities for data storage have created databases of immense size that can be tapped to generate knowledge. However, the challenge is to extract relevant information from this data and act upon it in a timely manner. Efficient decision-making is a by-product of thorough analysis of available data on a given problem.

Healthcare providers are confronted daily with constantly changing information needs to manage care. In every two of three patient encounters, the average clinician has unmet information needs, even though managed care and other healthcare system innovations mandate that providers be knowledgeable about the details of patient care. As practice data are computerized, the ability to capture, store, retrieve, organize, and analyze the information of clinical practice can provide information for decision support, enhancement of documentation, and identification of care trends and costs, with the ultimate goal of improved patient care.

3.5 Healthcare Information

The cost of information is usually not stated or, indeed, even known, but information represents a large percentage of the healthcare cost structure.

The cost of information technology is definitely high, but the cost of manual information handling is also expensive .25% of hospital cost is spent on information handling, primarily as a means of communication. It is now much easier to access more information at a substantially lower cost. This ability should facilitate more informed choices and better decisions, but with all the increases in "cheaper information," little progress has been made in deriving knowledge from this information.

There are indications that the most frequent problem with healthcare information is a lack of availability. Too much information, information in the wrong place, incomplete, inaccurate, inconsistent, illegible or difficult to understand information is also noted.

3.6 Healthcare Records

Patient health records represent comprehensive documents of the continuity of healthcare and are a rich source of data for research. Such data are generally accessible, accurate and relatively inexpensive. The advantages of using healthcare records as a source of information are accurate and timely data, rich clinical detail and dates attached to data elements. Traditionally, the paper record is documented in a "diary" style, and includes documentation that produces a defensive legal record. Documentation in health records is assumed to be legally and medically accurate and reliable. Historically, the nursing documentation in the patient's "chart" has been seen as a "transaction log rather than an evolving repository of practice based on nursing knowledge". As nursing diagnoses, interventions and patient outcomes are captured; the nursing record becomes a document that records actual nursing practice. This kind of nursing documentation offers the opportunity for study of actual nursing practice and its effects on patient outcomes.

There are disadvantages to using healthcare records as a data source. Concerns related to such data are that data are collected as a by-product

of some other processes; data are probably collected and entered by many people without any quality check; data may have different structure even within the same database; and missing data may be common. Other disadvantages are related to the non-research purpose of the record, the presence of selective information, the need for interpretation of certain information in the record, and the difficulty with data verification. Although concern has been expressed about the reliability and validity of health record data, most investigators operate on the premise that healthcare records provide fairly accurate information [2].

Chapter Four

Introduction to PolyAnalyst

Befor we began describe our case study we must introduce some informations about the software we used .

4.1 What is Data Mining and Knowledge Discovery

PolyAnalyst is a next-generation data mining system. Data mining represents a new and promising branch of Artificial Intelligence (AI) that embraces different technologies aimed at the automated extraction of knowledge (meaningful patterns and rules) from databases (large amounts of raw data).

For a long time, knowledge acquisition has been the bottleneck in the process of turning the raw data into informed and successful business decisions. That is why data mining is the hottest AI application today – gradually, more and more people understand the necessity and advantages of using machine-learning methods for intelligent data analysis. The goal of data mining for a company is frequently to improve profitability through a better understanding of its customers, sales, products, or operations. However, data mining is not restricted to marketing and business – finding patterns and rules in raw data can be important to science, medicine, academia, and engineering as well.

The growing amount of information available from computerized storage and the increasing complexity of this information make it impossible for a human to come up with an effective solution. This makes data mining the technology of the future for every profession.

Tasks well suited to data mining include prediction (determining the value of one variable based on patterns found in others), classification (dividing data into categories based on its attributes), clustering (the undirected finding of categories that a given dataset falls into naturally), and description (putting a given pattern or relationship into explicit form.)

PolyAnalyst is a powerful multi-strategy data mining system that implements a broad variety of mutually complementing methods for the automatic data analysis. One of the most advanced exploration engines, Find Laws, utilizes our unique Symbolic Knowledge Acquisition Technology™ (SKAT) – a next-generation data mining technique. PolyAnalyst automatically finds dependencies and laws hidden in data, presenting them explicitly in the form of rules and algorithms. The system builds empirical models of an investigated object or phenomenon based on the raw data. The user does not have to provide the system with any assumptions about the form of the dependencies –PolyAnalyst discovers the hidden laws automatically, regardless of how complex they are.

PolyAnalyst works with data extracted from flat files or relational databases, and can work with:

1. Numerical (floating-point).
2. Integer.
3. Yes/no (binary).
4. Date.
5. Discrete (categorical or string)
6. Variables.

Relationships in the data can be discovered, predictions made, and the data classified and organized using PolyAnalyst's suite of analytical algorithms. Data Mining can perform tasks beyond the scope of statistical analysis software.

4.2 What Makes PolyAnalyst Unique in This Field

PolyAnalyst is unlike any other data mining software package for a variety of reasons. First and foremost, PolyAnalyst is a complete suite of data mining algorithms, allowing data mining to be performed in whatever way is appropriate for the data being analyzed. PolyAnalyst includes eleven different approaches to data mining in one package. This is useful both because different tools are appropriate for different data and because sometimes the use of one tool will produce data that can be better analyzed using another tool. PolyAnalyst combines all your data mining requirements in a single package.

In addition, PolyAnalyst's exploration engines are superior to the engines provided by other vendors for the same purposes. While many data mining packages offer clustering, PolyAnalyst's clustering uses the LA (Localization of Anomalies) algorithm that overcomes many of the drawbacks of conventional clustering, such as the dependence on functional transformations and the inability to place new points into existing clusters. Linear Regression is a ubiquitous data analysis technique, but PolyAnalyst's Stepwise Linear Regression automatically determines which attributes are most important to the analysis and which can be ignored, as well as correctly incorporating categorical and yes/no attributes into the regression.

Finally, PolyAnalyst's Find Laws engine is unique – no other product offers the ability to generate symbolic high-order rational expressions from a dataset. While neural network tools offer good predictive ability, their output cannot be placed in a simple, human-readable format or incorporated into a cell in a spreadsheet. Find Laws provides this capability, generating rules with the predictive power of a neural network but which show the analyst the exact form of the dependence. Sometimes, it can be as important to know why the relationship exists and exactly how it works, as it is to know the relationship itself. Unlike other data mining tools, Find Laws provides this explanatory power.

4.3 About the Exploration Engines

Most of PolyAnalyst's power comes from its rich suite of exploration engines – the machine learning algorithms that make PolyAnalyst a sophisticated data-mining tool instead of just a statistical analysis and charting package. The exploration engines are what allows PolyAnalyst to extract useful knowledge from raw data. As mentioned before, one of PolyAnalyst's greatest advantages is that it contains eleven different exploration engines, each with a different purpose and a different place in the data mining process. The following exploration engines are a part of PolyAnalyst 4.6.

4.3.1 Find Laws

The powerful Find Laws algorithm, unique to PolyAnalyst, generates complex symbolic rules that represent nonlinear dependencies in your data. Using Symbolic Knowledge Acquisition Technology (SKAT), Find Laws is one of PolyAnalyst's most powerful and useful data mining tools. It searches for functional dependencies hidden in data and expresses the discovered knowledge explicitly in the symbolic form as mathematical formulae, including rational polynomials, relational operators and conditional blocks. The ability of Find Laws to automatically build a wide variety of mathematical constructions, including complex nonlinear algebraic expressions and functions, makes it an unmatched knowledge discovery tool. However, Find Laws is resource-intensive and takes a great deal of time to run. As a result, Find Laws is often used in the final stages of data mining to present a human-readable rule explaining the analysis.

4.3.2 Nearest Neighbor (Memory-Based Reasoning)

The Nearest Neighbor exploration engine, based on the PAY algorithm, was introduced in PolyAnalyst version 4.0. It uses a memory-based classification system (k-nearest neighbor), assigning values to data points based on their "proximity" to other data points. The rules it produces cannot be viewed symbolically like Find Laws rules, but can be used by PolyAnalyst to classify other data points.

4.3.3 PolyNet Predictor

PolyNet Predictor is PolyAnalyst's neural network tool. It generates a network of nodes, each of which contains a mathematical expression, which can be used to predict the value of one attribute based on the values of several others. If the resulting neural network is sufficiently simple, it will be displayed as a symbolic rule, but in real-world cases with large amounts of data, a neural network usually operates as a "black box". That is, you can pass data through the network to get valid predictions, but can't "look inside" to see how the network works or what criteria are being used to categorize the data.

Thus, PolyNet Predictor should be used when the primary goal is to learn to predict values of a target attribute, not to obtain explicit knowledge about the form of its dependence on other attributes. The key advantage of PolyNet Predictor is that it can work with databases containing a large number of records – much greater than can be efficiently handled by Find Laws or Nearest Neighbor. However, PolyNet Predictor should not be used to explore datasets with a large number of attributes.

4.3.4 Find Dependencies

The Find Dependencies algorithm is used to find relationships and dependencies in the data quickly without finding the specific form of those dependencies. Unlike Find Laws, PolyNet Predictor, Linear Regression, Nearest Neighbor, and others, Find Dependencies does not produce a predictive rule.

Rather, it tells you how strong the connections between attributes are, so that you can further explore those attributes with other engines. In this way, it is a statistical data preprocessing module that is aimed primarily at creating the best operating conditions for Find Laws or another predictive algorithm. In addition, Find Dependencies can be used to identify points that do not obey the dependencies found in most of the data, thus allowing identification of "stray" data points and errors in the data.

This makes it very useful for preprocessing and determining which attributes are most important to subsequent analysis.

As a result, Find Dependencies is often the first engine used in a data mining exploration.

4.3.5 Stepwise Linear Regression

Linear Regression is one of the oldest and most well known methods of statistical prediction – it is the process of creating a line through a space such that the sum of the squares of the distance between the line and each point is minimized. It provides an excellent model of phenomena that are actually linear, but produces models of nonlinear functions that are accurate at part of the function and quite poor at predicting other parts.

Nevertheless, Linear Regression is a valuable tool as it is very fast and produces easily readable and interpretable results. PolyAnalyst's stepwise linear regression can work with any number of attributes, and automatically determines which attributes give the best linear prediction rule. It is based on a very fast algorithm that performs the multiparametric linear regression search on a variable number of parameters, with automated selection of the most influencing independent attributes and rigorous statistical estimation of the significance of the obtained result.

4.3.6 Market Basket Analysis

Market Basket Analysis is an algorithm that examines a long list of transactions in order to determine which items are most frequently purchased together. It takes its name from the idea of a person in a supermarket throwing all of their items into a shopping cart (a "market basket"). The results can be useful to any company that sells

products, whether it is in a store, a catalog, or directly to the customer. In addition, Market Basket Analysis can be used for non-marketing purposes whenever correlations between a large number of different items are of interest.

4.3.7 Transaction Basket Analysis

Transaction Basket Analysis performs the same product association analysis as Market Basket Analysis except it is designed for data provided in a transactional format. Transaction Basket Analysis is performed on data where each individual product purchase is considered a transaction. Therefore, if a customer purchased three products at the same time, three separate records/transactions are recorded. This is by far the most common format for storing seller data.

4.3.8 Cluster

The Cluster exploration engine examines a dataset for areas of similarity. The datasets records are compared by all attributes, and similarities and differences are found. The use of all attributes makes the Cluster algorithm very useful for beginning a data mining project, especially since it is an undirected method and does not require the selection of a target attribute. Using the Cluster algorithm at the start of the project will help determine which attributes and datasets are most important to the rest of analysis. It outputs both datasets representing each of the clusters that can then be further explored and a rule that can be used to quickly sort new data into the existing clusters.

4.3.9 Classify

The Classify algorithm is used to solve a very common problem in data mining – splitting a dataset into two groups. The two groups could be buyers and non-buyers, loyal bank customers and those likely to leave, fraudulent transactions and legitimate ones, or good products and defective ones. Classify produces both a scoring rule and a threshold, finding not only a way to score records but also the point at which most accurate classification is achieved.

Classify is one of PolyAnalyst's derived exploration engines – the development of the classification rule can be done by Find Laws, PolyNet Predictor, or Linear Regression. Classify employs the other engine to develop the rule, then uses its own fuzzy logic algorithms to determine the classification threshold that maximizes the number of correct classifications.

4.3.10 Discriminate

Discriminate, PolyAnalyst's other derived exploration engine, is used to compare a dataset with the World dataset. It is a form of undirected data mining in the sense that it does not have a target attribute, but must be used later in the project

since it requires that you have multiple datasets. Discriminate will find a rule that can be used to predict if a given data point will fall into the selected dataset, or would be placed elsewhere in the World dataset. It develops its rule and threshold the same way as the Classify algorithm.

4.3.11 Summary Statistics

Summary Statistics is not a machine-learning algorithm, but it is nevertheless a vital part of the analysis procedure. The Summary Statistics exploration engine provides basic statistics about your data, including means, standard deviations, and frequencies. In addition, the Summary Statistics report includes frequency analysis charts for each category, string, and yes/no variable. Summary Statistics allows you to perform analyses that would normally be performed using other statistical software without having to leave PolyAnalyst.

4.3.12 Decision Tree

The Decision Tree exploration engine helps solve the task of classifying cases into multiple categories. Decision Tree is PolyAnalyst's fastest algorithm when dealing with large amounts of records. Decision Tree report provides an easily interpreted decision tree diagram and a predicted versus real table.

4.3.13 Decision Forest

The Decision Forest algorithm solves multi-category analysis by growing a set of competing classification trees, one for each of the selected classes. When attempting to categorize an attribute into several categories, traditional decision tree analysis often lacks the efficiency and accuracy needed for precise solutions. The problem originates from the fact that a single decision tree does not provide a robust mechanism for assigning records to multiple classes. The Decision Forest results are displayed as a tree diagram along with a predicted versus real table.

4.3.14 Text Analysis

Text Analysis analyzes data in textual format by extracting key concepts from text stored in a database format, categorizing individual database records, and deriving from text quantitative knowledge. This knowledge is delivered in a format that can be exploited by other machine learning engines of this data mining system. Every organization has a significant portion of business-related knowledge residing in the form of natural language documents. These documents can be in the form of memorandums, contracts, regulations, manuals, e-mail messages, or natural language fields in a database. Text Analysis enables the user to turn this unformatted information into quantifiable knowledge.

4.3.15 Text Categorization

Text Categorization engine automates categorization of dataset records on the basis of their textual field contents. This knowledge is delivered in an easily navigated categorization tree. The categorization tree allows for the creation of rules, based on the nodes, to be applied to your datasets allowing textual categorization.

4.3.16 Link Analysis

The Link Analysis exploration engine reveals and visually represents complex patterns of correlations between individual values of all categorical and Boolean attributes. Results of the analysis are displayed as a graph of linked objects supporting various object manipulation and drill-down operations. The visual output of LA facilitates better understanding of the hidden structure of investigated data, and helps quickly isolated interested patterns for further investigation.

4.3.17 Link Terms

Link Terms exploration engine creates an intuitive display of most characteristics terms for textual fields of explored datasets with the purpose of deep understanding of dataset's structure. Easy navigation and selection of specific subsets of records on the basis of their semantics allows analyst to get a true understanding of the textual fields.

4.3.18 Taxonomies

Taxonomies give you the opportunity to classify document arrays and provide a quicker navigation mechanism over the created classification tree. Moreover, Taxonomies allow you to trail document distribution statistics (and their dynamic) over different rubrics and store the document subsets being browsed into new datasets.

4.3.19 Text OLAP

PolyAnalyst OLAP (Online Analytical Processing) gives you an opportunity to perform quick navigation through the project's data. During this navigation, you are given the ability to store the subsets being browsed into new datasets for further investigation or exportation. OLAP is performed through two PolyAnalyst objects: the dimension matrix and the OLAP report.

4.4 The Exploration Engine Used in this Project

In this project we will used the following Exploration Engine:

4. 4.1 - Summary Statistics.

4.4.2 -Link Terms.

4.4.3 - Transaction Basket Analysis.

Bellow we a breif descreption about Summary Statistics.

4.4.1 Summary Statistics

Summary Statistics is not a machine-learning algorithm, but it is nevertheless a vital part of the analysis procedure. The Summary Statistics exploration engine provides basic statistics about your data, including means, standard deviations, and frequencies.

In addition, the Summary Statistics report includes frequencies charts for each category, string, and yes/no variable.

A- What Problems Can be Solve by Summary Statistics

Summary Statistics is not an algorithm generally used for problem solving, but rather to find the characteristics of the dataset and identify any anomalies. It can, however, be used to compare the frequencies of categorical attributes between two datasets.

In addition, Summary Statistics allows you to perform analyses that would normally be performed using statistical software without having to leave PolyAnalyst.

For example, when reviewing the basic statistic parameters you can detect some unusual values in your data, find constant attributes, see categorical attributes that take only two values and hence should be transformed to yes/no attributes, and solve other simple preprocessing problems. There for be e by summary statistics we can do the following :

- Getting a quick overview of a dataset
- Comparing the characteristics and distributions of two or more datasets
- Comparing frequencies of categorical attributes between two or more datasets
- Performing simple statistical analysis without having to use a separate statistics software package

B- When to Use This Algorithm

Summary Statistics is frequently used throughout analysis to get an idea of the shape of a dataset. It is often used on the World dataset at the beginning of analysis to see if any of the attributes contain unexpected data, such as some records having an invalid value in a category or string attribute.

In addition, once World has been split into many other datasets using the Split function or other exploration engines, the frequency analysis capabilities of Summary Statistics can be used to look for differences between new datasets.

Summary Statistics can be used any time a conventional statistical package would normally be used.

C- Optimal Number of Records

- Minimum of 2 records.
- Maximum of 3,000,000 records.

D- The Data Used

Summary Statistics will work with any type of data, but it provides different results for each data type. Numerical and integer attributes have their mean, standard deviation, minimum, maximum, range, and median calculated.

In addition, integer attributes include the mode and number of different values. Category and yes/no attributes, on the other hand, have the number of values, mode, and frequency of each value calculated.

When you choose **Explore | Summary Statistics**, the dialog displayed prompts you to enter a report name and choose which attributes to include. Summary Statistics requires no target attributes, or specific options to be set, as it determines summary statistics for the entire dataset. The dialog box contains a check box, "Remove constant fields" that that cleanses the selected dataset by removing the column(s) containing a constant field for each record.

If a constant field(s) are found, a new cleansed dataset will be created and appear in the tree diagram.

E- Output Format

The Summary Statistics report begins by listing the number of attributes and records that were selected from the dataset. Next, a table is displayed containing each statistic that was calculated for the numerical and integer attributes. This lists the mean, standard deviation, minimum, maximum, range, and median for each attribute. In addition, integer attributes include the mode and number of different values. These are all standard statistical measures of the characteristics of the data.

The next table lists category, string, and yes/no attributes. The table begins by listing the first categorical attribute, the number of records containing a value for that attribute, the mode (the value that occurs most frequently for that attribute), and the number of different values that attribute can take on. Beneath this line is a list of each of the possible values of that attribute, followed by the frequency (number of times it occurs) of the value. This table repeats for each of the category, string, or yes/no attributes.

The summary of the output can be :

- Table of statistics for integer and numerical attributes.
- Frequencies table for yes/no attributes.
- Frequencies table and mode for category and string attributes.
- Frequencies graph for yes/no, category, and string attributes [11].

Chapter Five

Case Study

5.1 Case Description

Fraud Detection is an area of vital importance to numerous businesses: credit companies, banks, insurance companies, telecom and airline companies alike have to be able to discern fraudulent transactions from the main stream of legitimate business transactions. Inability to catch fraud can become an extremely costly, painful and damaging problem to a business and thus the need for efficient Fraud Detection solutions resides high on the "to do" list of every company.

5.2 Project Objectives

Having to deal with millions of insured customers and dozens of thousands of providers of medical services, healthcare insurance companies have to routinely address the task of verifying the authenticity and legitimacy of all processed transactions. In fact, fraudulent transactions might be originating from all involved parties.

For example:

1. Fraudulent "providers" can be sharing lists of valid patient IDs and trying to bill the insurance company for the services that were never rendered in reality.
2. Fraudulent "patients" might try to bill insurance company for a large number of ghost procedures.

As fraud schemes get more sophisticated and the volume of transactions grows, it becomes increasingly more difficult to discern fraud from legitimate transactions. Investigators have to utilize advanced data analysis tools capable of processing large volumes of data and detecting unusual events deviating from the normal operation patterns. Fraud schemes are rapidly changing and analysts need to be able to discern new fraud patterns without an explicit prior knowledge of these patterns.

In our project we will be searching for traces of two possible healthcare fraud mechanisms listed above. Using a combination of PolyAnalyst data mining techniques we will obtain valuable results.

Also we can use combination of PolyAnalyst data mining techniques to analyze our data.

5.3 Descriptions of the Data

A large health insurance company provided the following dataset containing about 15,000 records representing individual procedure transactions. This is a subset of the year 2000 patient records of a health insurance company. We are going to look at a small subset of the resulting data, containing data on only 100 first patients. The data is located in the 'Patients.csv' dataset. The fields and their definitions are listed below:

1. PatientID: Individual patients in this healthcare system are designated by unique PatientIDs.
2. PatientName: Disguised patient's name.
3. ProviderID: Individual healthcare providers are also designated by unique provider IDs.
4. Provider Name: Disguised provider's name.
5. Service Date: A date when the service was provided.
6. Procedure Code: Individual procedures have their unique codes.
7. Net Payments: The net Sudanese Dinar amount that the insurance company paid to a provider for the service.

Algorithm 5.1 Detect the Fraud from Patients and Providers

1. Load the data from patient.pvc file.
2. Splitting the dataset.
3. Create summary statistics.
4. Create Link Chart (for possible patient's frauds).
5. If there is a patient have more than one procedure with the same date and same procedure but with different payment , select this patient name and make new dataset for more analysis to determine if there a fraudulent behavior or not.
6. Create Transaction Basket Analysis (for possible provider frauds).
7. If there are many providers sharing the same patients we determine farther analysis if there a fraudulent behavior or not.

5.4 Performed Analyses

5.4.1 Load the Data to the Project

At first we load 'Patients.csv' dataset in PolyAnalyst and open the World dataset to take a look at the structure the data.

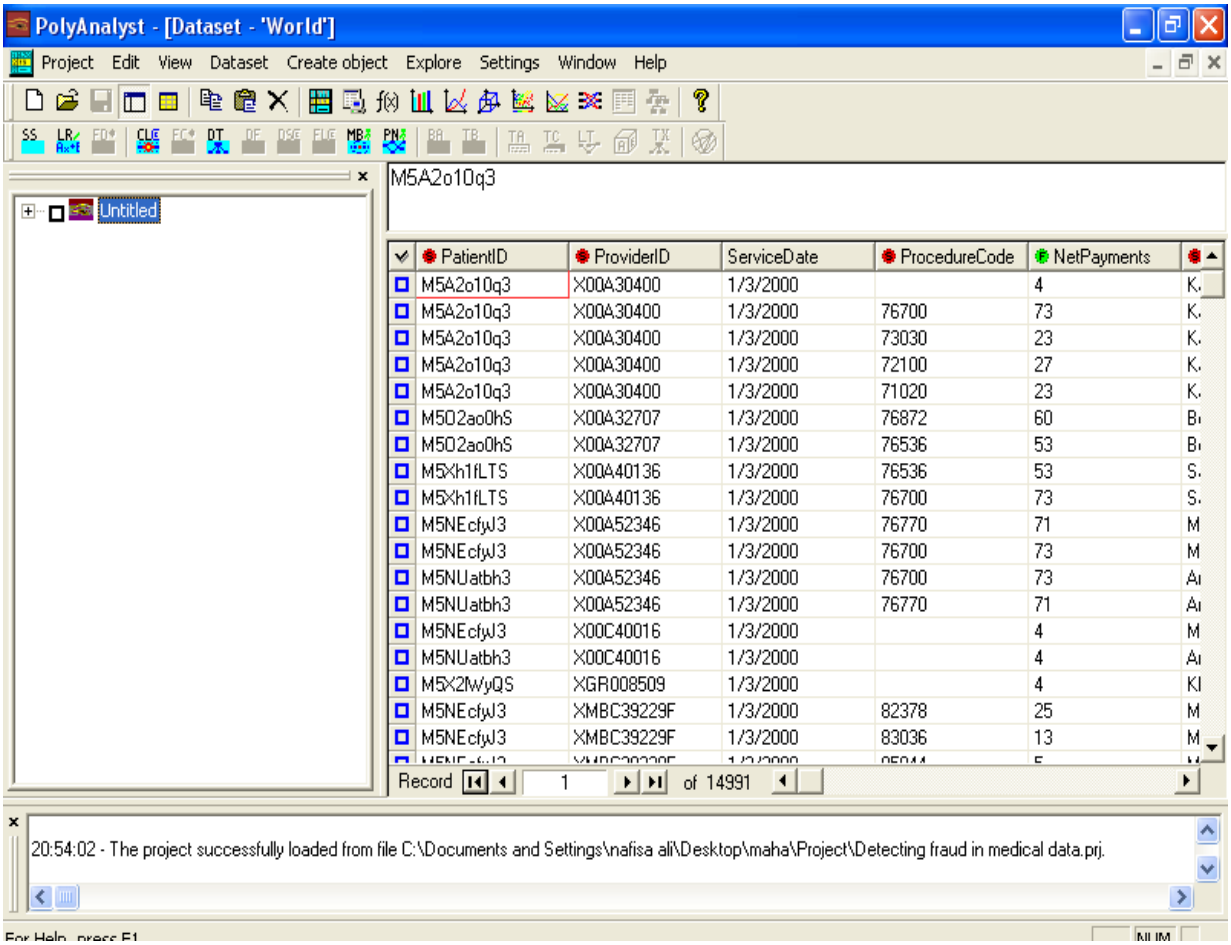


Fig 5.1 Patients World Data Set

5.4.2 Splitting Datasets

In our project there are different categories of patients, we see that we can divide up what we are studying because their medical characteristics may overlap.

There for according to the patient world data set we can spilt it to "40" data set.

PolyAnalyst provides the functionality to segment a data set by using the Split option. The World data set contains records of about 273 Provider Name.

We right click on the World data set. Select Split | To equal intervals from the context menu.

5.4.3 Create a Summary Statistics for the Data

1. Create a new dataset *Explored*, excluding PatientID and ProviderID attributes from this dataset.

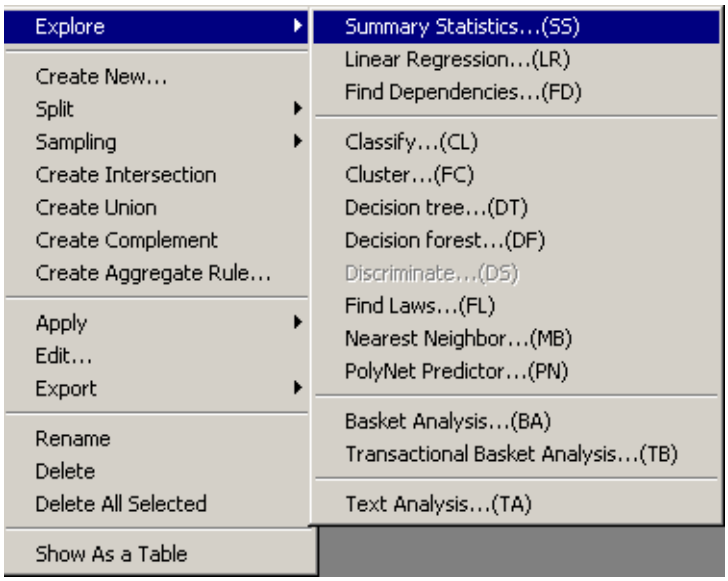


Fig 5.2 The Steps of Summary Statistics

2. Launch Summary Statistics exploration engine on the Explored dataset in order to gain an overall picture of the data. The important observations are that there are 39 unique patients present in the dataset, 273 individual providers, and 270 different procedures. Finally, one can observe that some patients had a large number of procedures done during the period of a year (some of these procedures can be the same).

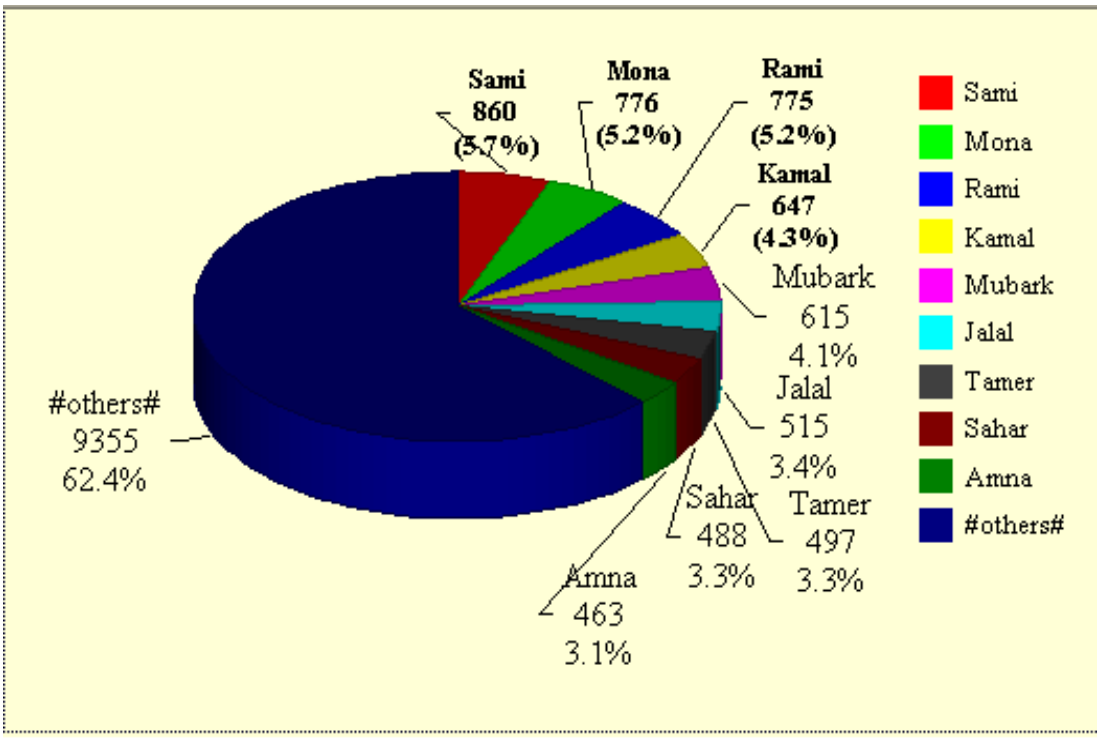


Fig 5.3 Graph Result of Summary Statistics

From the above graph we observe that patients Sami, Mona, and Rami had more than 700 procedures during the year 2000. These patients might be considered as primary suspects for fraud. For this reason we can run a simple Link Analysis exploration in order to figure out whether this is indeed the case.

PolyAnalyst generates a Summary Statistics report named Summary of medial data in the Results panel. Below is the statistical information of the attributes.

5.4.4 Creating a Histogram to Analyze Data

PolyAnalyst provides us by the ability to view insights into the data through graphical output. For example, it may be of interest to examine the Provider Names with Service Date.

- From the main file menu, select Create Object | Create Histogram….

The Create new Histogram window appears where the data sets are listed on the left and the attributes are listed on the right.

- Select all the data sets expect the World data set by right clicking on the data set names.

5.5 Possible Patient Frauds

Create a Link Chart for the Explored dataset selecting PatientName as antecedent attribute and ProcedureCode as consequent attribute.

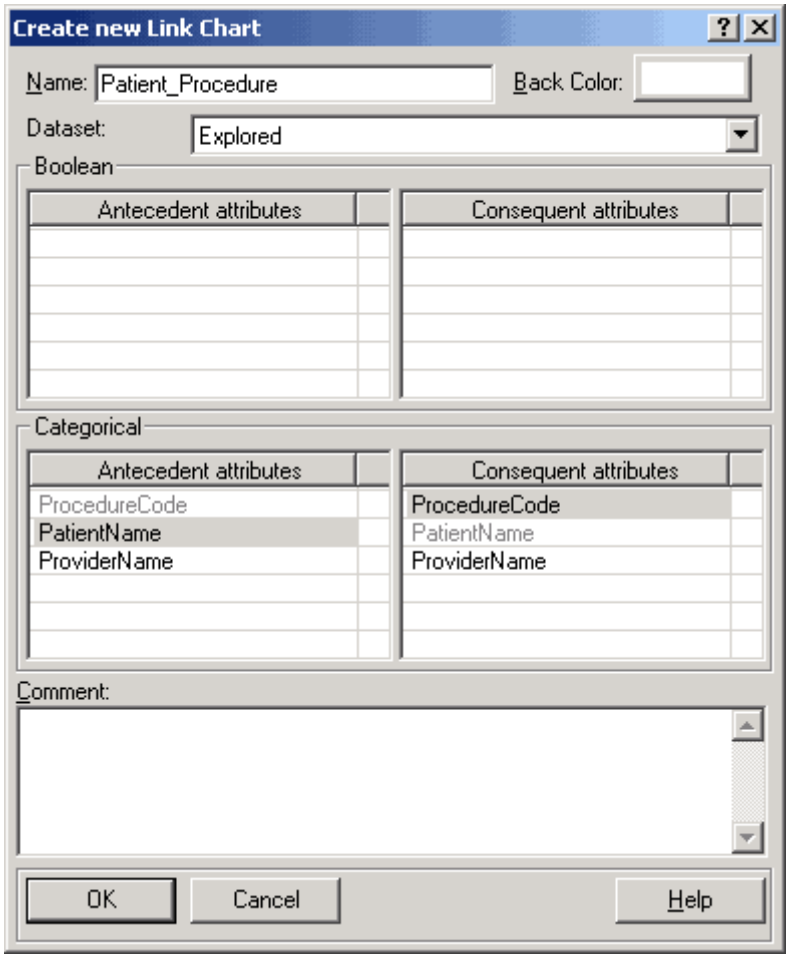


Fig 5.4 Create new Link chart

Select the radio button to show only positive links. Set the minimum Correlation to 7.75 by moving the slider in the top left corner of the link chart control bar. Then we locate those patients, who have the largest number of links to different procedures and evaluate some of these links in more detail.

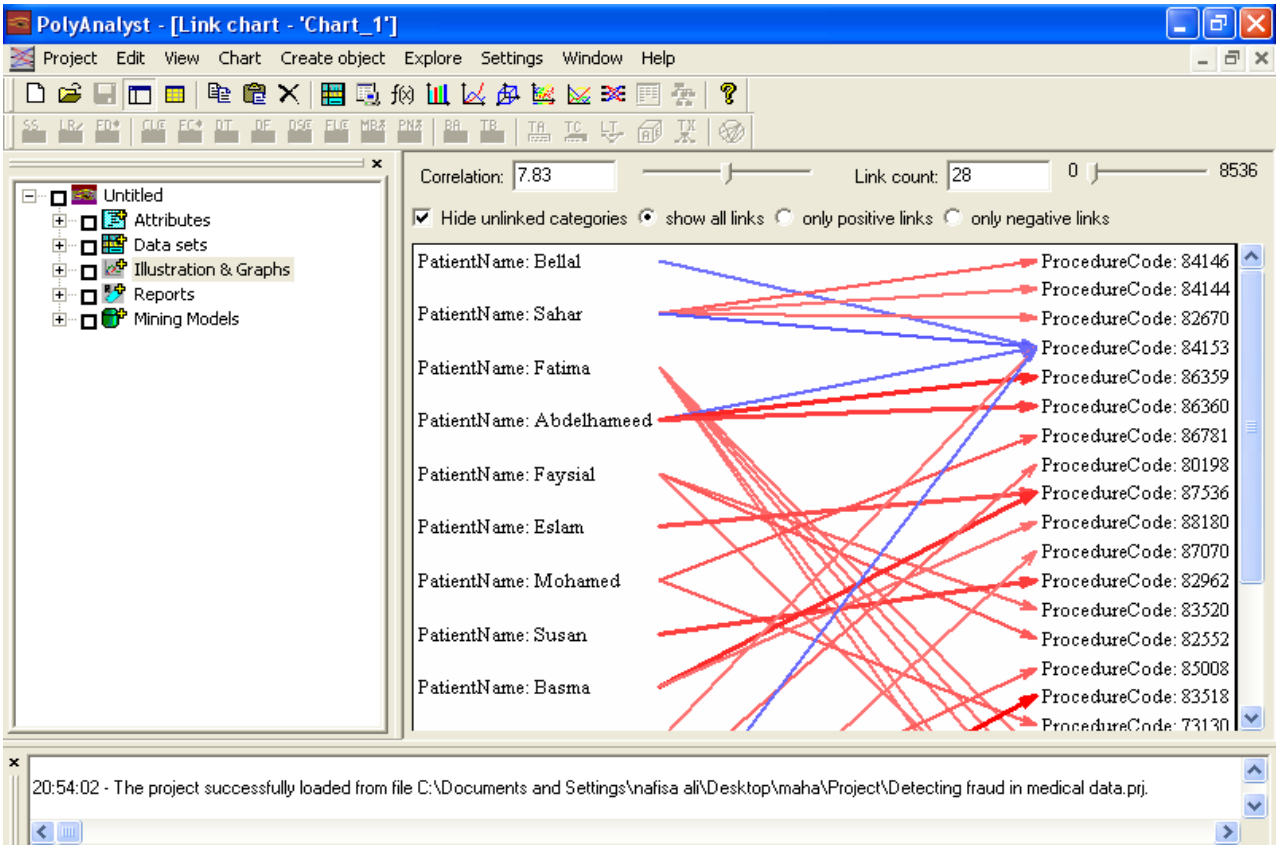


Fig 5.5 Result of Link Chart

For example, consider the patient that has the largest number of links to different procedures, the patient *Fatima* has 5 links to procedures associated with her, and we can evaluate these links in more detail. By Pressing and hold the mouse button on a selected link. We will see a prompt displaying some statistical characteristics of the link.

The selected link becomes highlighted with a different color. From the pop-up prompt we learn that patient *Fatima* had 218 individual procedures – this is one of the lowest numbers of procedures for the considered list of patients.

By Right click anywhere on the link chart and select an option for creating a dataset representing the data related to this link. We will see the corresponding dataset appearing in the project navigation tree: this dataset is named after the name of the original dataset and values of attributes associated with both names of the link. We repeat the same procedure for four other links originating from the same patient name.

One of the most frequent manifestations of fraud of the patients might be the presence of the same services obtained by a patient on the same date from the same or different providers. Sequentially open all newly created datasets. It is interesting to note that in all these cases *Fatima* had the same procedure performed twice on the same date and by the same provider (but being charged different amounts for the same procedure). This sequence of transactions looks suspicious.

Now we become really suspicious about the relationship between *Fatima* and Provider #177 and might want to carry out much more in-depth analysis to determine whether we indeed identified some fraudulent behavior.

Surprisingly, we discovered the first example of suspicious behavior not for those patients who had the largest number of procedures performed during the year 2000 (of order 700), but for one of the patients who had about the smallest number of procedures compared to other patients in the selected dataset.

5.6 Possible Provider Frauds

What is the best way to identify fraudulent "providers" sharing lists of valid patient IDs and billing the insurance company for the services that were never rendered?

The most direct method offered by PolyAnalyst is to utilize Transactional Basket Analysis (TB)**.**

Launch the (TB) engine on the Explored dataset and select ProviderName as Basket, and PatientName – as Product. This way, one can reveal groups of patients shared by sets of providers and the corresponding sets of providers themselves.

In this exploration we are interested not only in seeing the composition of the groups of patients shared between providers, but rather in identifying sets of providers sharing their customer lists. Correspondingly, we want to isolate datasets representing transactions related to each found cluster of shared patients.

To achieve this, check the Create datasets checkbox in the window for launching the (TB) algorithm. Note that the Basket transactions only checkbox will become checked automatically by default too: we are

interested in seeing only those transactions, which involve patients and providers that belong to the considered cluster.

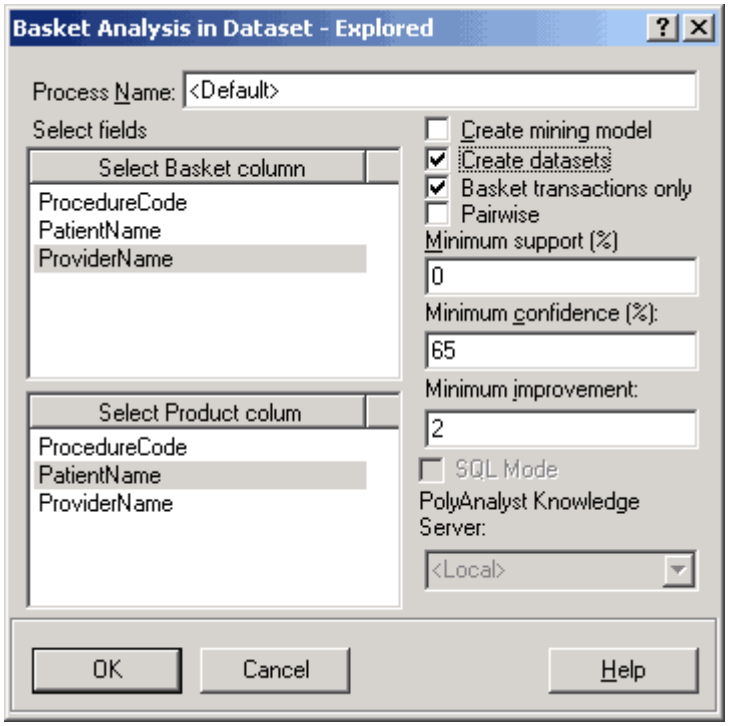


Fig 5.6 Creation of Basket Analysis

By the Transaction Basket Analysis (TB) we found five groups of patients shared between different sets of providers varying in number.

The first groups of patients represents one very interesting result, where 5 patients (out of 39 totals – each eighth patient) are shared by 10 different providers (the found Support for this group of patients is 10).

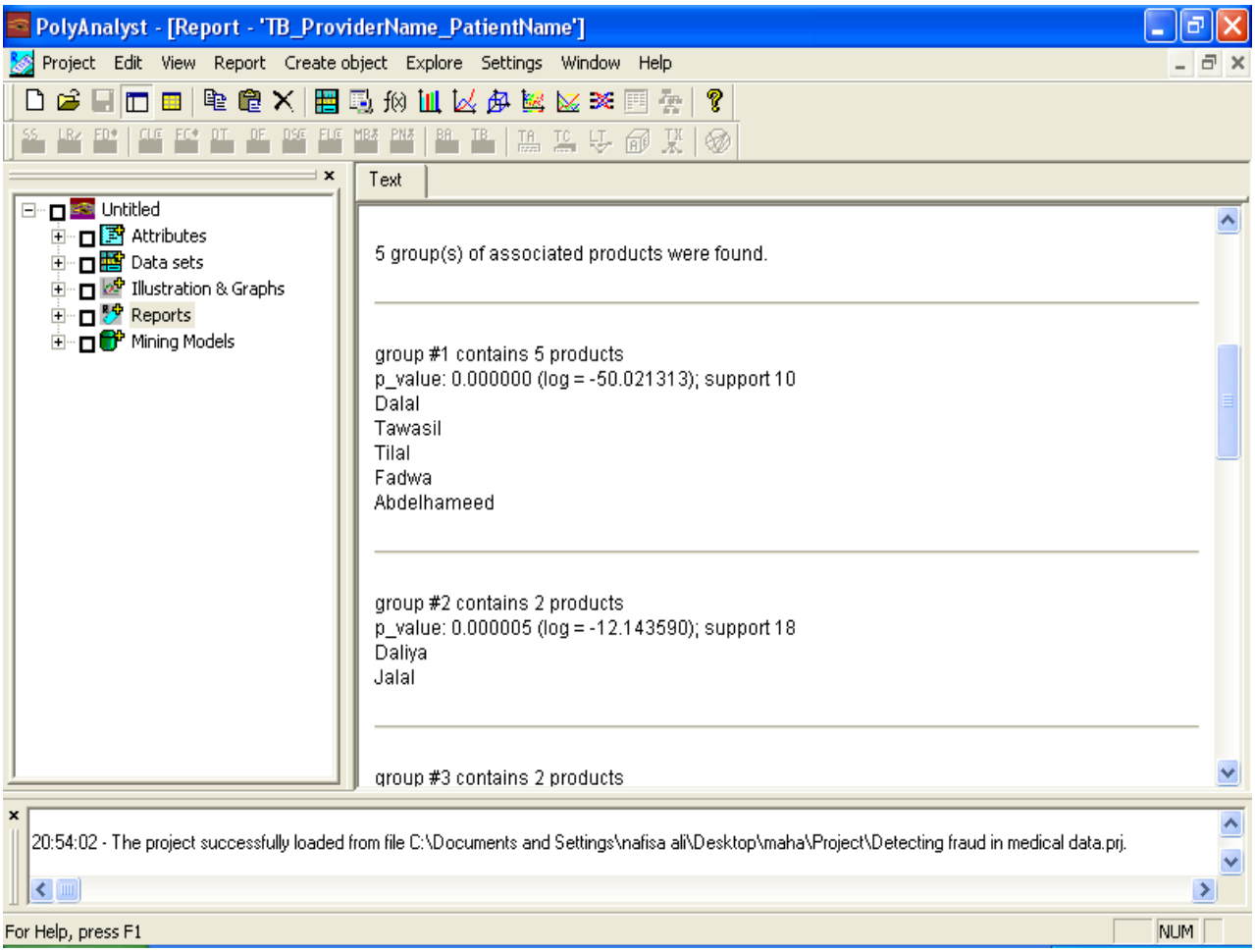


Fig 5.7 TB ProviderName_PatientName

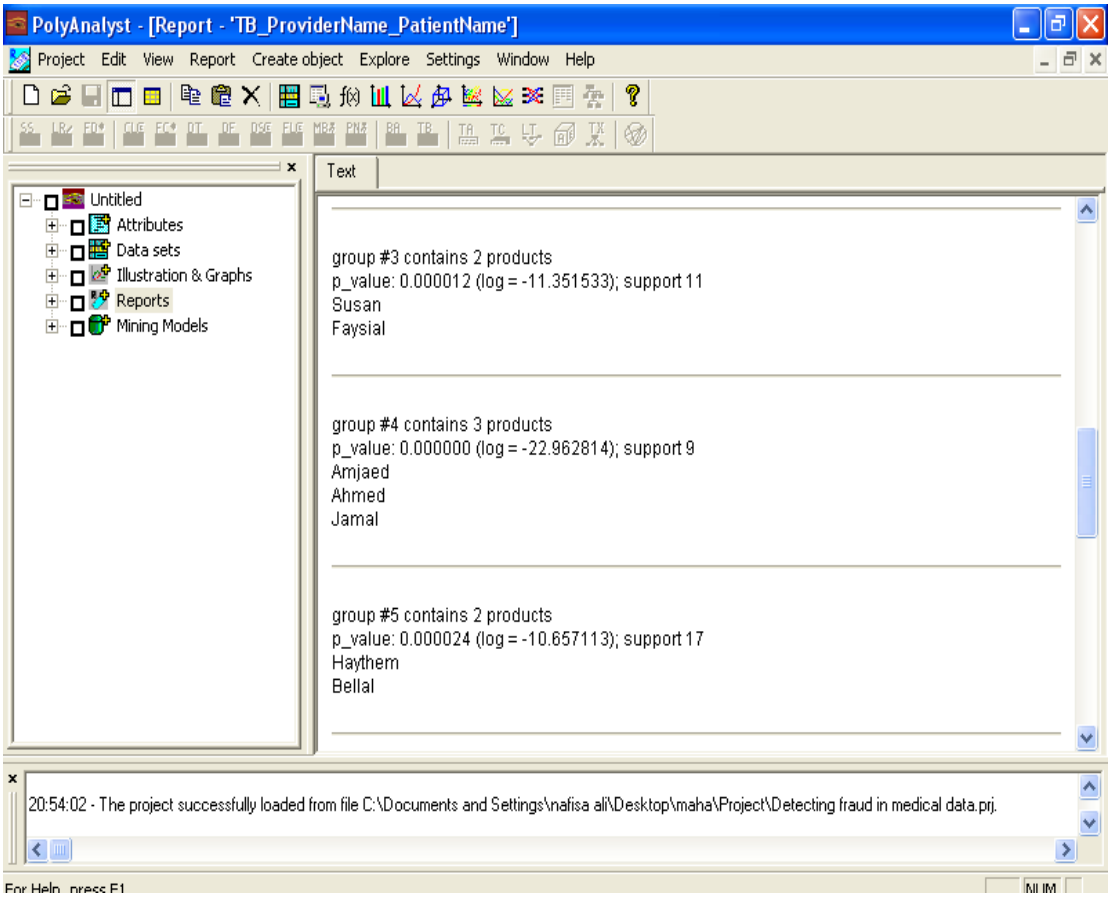


Fig 5.8 TB ProviderName_PatientName

The Transaction Basket Analysis (TB) engine will automatically created five datasets that contain all transactions (procedures) corresponding to five found groups of products (patients). We will also find these datasets appearing in the project navigation tree.

If we comparing the results of Summary Statistics run on each of these datasets, we can identify providers serving the same groups of patients, possibly even across five groups of patients found by the Transaction Basket Analysis **(**TB) engine. The table below represents the Summary Statistics results for each found dataset.

Table (5.1) Transaction Basket Analysis for the ProviderName _PatientName_1

TB for the ProviderName _PatientName_1
5 patients, 10 providers
Provider #225 179 (17.02%)
Provider #234 155 (14.73%)
Provider #229 147 (13.97%
Provider #223 135 (12.83%)
Provider #224 131 (12.45%
Provider #235 101 (9.601%)
Provider #268 101 (9.601%)
Provider #232 76 (7.224%)
Provider #16 20 (1.901%)
Provider #67 7 (0.6654%)

Table (5.2) Transaction Basket Analysis for the ProviderName _PatientName_2

TB for the ProviderName_ *PatientName*_2
(2 patients, 18 providers)
Provider #210 187 (29.08%)
Provider #224 81 (12.6%)
Provider #223 67 (10.42%)
Provider #234 52 (8.087%)
Provider #235 52 (8.087%)
Provider #246 46 (7.154%)
Provider #248 35 (5.443%)
Provider #237 35 (5.443%)
Provider #236 32 (4.977%)
Provider #232 28 (4.355%)
Provider #106 7 (1.089%)
Provider #124 6 (0.9331%)
Provider #158 3 (0.4666%)
Provider #6 3 (0.4666%)
Provider #89 3 (0.4666%)
Provider #118 2 (0.311%)
Provider #257 2 (0.311%)
Provider #39 2 (0.311%)

Table (5.3) Transaction Basket Analysis for the ProviderName _PatientName_3

TB for the ProviderName_ *PatientName*_3
(2 patients, 11 providers)
Provider #234 116 (33.33%)
Provider #268 48 (13.79%)
Provider #224 39 (11.21%)
Provider #223 37 (10.63%)
Provider #235 36 (10.34%)
Provider #236 30 (8.621%)
Provider #232 29 (8.333%)
Provider #247 4 (1.149%)
Provider #63 4 (1.149%)
Provider #228 3 (0.8621%)
Provider #105 2 (0.5747%

Table (5.4) Transaction Basket Analysis for the ProviderName _PatientName_4

TB for the ProviderName_ *PatientName*_4
(3 patients, 9 providers)
Provider #270 102 (25.19%)
Provider #224 81 (20%)
Provider #225 69 (17.04%)
Provider #227 57 (14.07%)
Provider #232 44 (10.86%)
Provider #235 30 (7.407%)
Provider #16 13 (3.21%)
Provider #198 5 (1.235%)
Provider #247 4 (0.9877%)

Table (5.5) Transaction Basket Analysis for the ProviderName _PatientName_5

TB for the ProviderName_ *PatientName*_5
(2 patients, 17 providers)
Provider #229 58 (14.36%)
Provider #270 48 (11.88%)
Provider #224 47 (11.63%)
Provider #261 42 (10.4%)
Provider #235 41 (10.15%)
Provider #223 40 (9.901%)
Provider #268 40 (9.901%)
Provider #222 34 (8.416%)
Provider #232 20 (4.95%)
Provider #246 10 (2.475%)
Provider #16 6 (1.485%)
Provider #44 4 (0.9901%)
Provider #31 4 (0.9901%)
Provider #257 3 (0.7426%)
Provider #129 3 (0.7426%)
Provider #193 2 (0.495%)
Provider #163 2 (0.495%)

From the Previous sets we can see that:

1- Three providers are found in all five tables: included in all five groups of patients. These providers, *Provider #224*, *Provider #232*, and *Provider #235* are sharing 14 patients out of the total selected group of 39 patients. This represents almost 40% of all considered patients! From this fact we can imply that either these providers have tightly related services or one or more of these providers might be submitting fraudulent claims. In such a case, the insurance company should launch an additional in-depth investigation of the activities of the identified group of providers.

2- We can note that three more providers with a high number of overlapping patients with the three providers listed above. These are: *Provider #223* sharing 11 patients with other providers; and *Provider #234* sharing 9 patients, and *Provider #225* sharing 8 patients. These providers may also be good candidates for more in-depth investigation of their activities.

Chapter Six

Conclusion and Recommendations

6.1 Conclusion

There is a growing gap between more powerful storage and retrieval systems and the users' ability to effectively analyze and act on the information they contain. A new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining tools can make this leap. Data mining techniques have tremendous capabilities for navigating this massive data.

It may seem clearly that it is very hard to some one deciding to implement a Data mining system to determine which techniques should be used and also when this technique is use.

It is important to determine the criteria of the technique that it may used to solve the problem, this criteria may be depend on the type and size of the database.

6.2 Recommendations

Data mining tools have many powerful analysis tools that may use computer cycles to replace human cycles.

Many data mining tools still require a significant level of expertise from users. So user must design better interfaces if they hope to gain wider acceptance of their products

So I recommend to using data mining tools for its useful and powerful result that can help in marketing good result and in improving the performance depending on there stored database.

References

[1] Mehmed Kantardzic , Data Mining: Concepts, Models, Methods, and Algorithms . John Wiley & Sons 2003.

[2] Gerald Grant (ed) , ERP & Data Warehousing in Organizations: Issues and Challenges. Idea group publishing © 2003.

[3] Raghu Ramakrishnan / Johannes Gehrke , Database Management systems .second edition 1999.

[4] Michael J.A. Berry and Gordon Linoff , Mastering Data Mining. John Wiley & Sons 2000.

[5] Rhonda Delmater and Monte Hancock ,Data Mining Explained : A Manager's Guide to Customer-Centric Business Intelligence. Digital press © 2001.

[6] David Hand, Heikki Mannila and Padhraic Smyth.s, Principles of data mining. The MIT press ©2001 .

[7] John Wang (ed) ,Data Mining: Opportunities and Challenges. Idea group publishing © 2003.

[8] Barry de Ville , Microsoft Data mining: Integrated Business Intelligence for e-commerce & knowledge Management .Digital press © 2001.

[9] **http://www.qub.ac.uk/**Data mining . Accessed on January 2005.

[10] Alex Berson, Stephen Smith, and Kurt Thearling , Building Data mining Aplications for CRRM .Computing Mc Graw – Hill © 2000.

[11] http://www.megaputer.com Accessed on May 2005**.**