Bachelor's Thesis (UAS)

Degree Program in Information Technology

Information Technology

2013

Yifei Ren

DATA PREPROCESSING FOR DATA MINING

TURUN AMMATTIKORKEAKOULU
TURKU UNIVERSITY OF APPLIED SCIENCES

BACHELOR'S THESIS │ ABSTRACT

TURKU UNIVERSITY OF APPLIED SCIENCES

Information Technology

2013 | 43

Supervisor: Patric Granholm

Yifei Ren

DATA PREPROCESSING FOR DATA MINING

People have increasing amounts data in the current prosperous information age. In order to improve competitive power and work efficiency, discovering knowledge from data is becoming more and more important. Data mining, as an emerging interdisciplinary applications field, plays a significant role in various trades' and industries' decision making. However, it is known that original data is always dirty and not suitable for further analysis which have become a major obstacle of finding knowledge.

This thesis aims to introduce this new field and data preprocessing as a critical step in a data mining project as well as its practical part using SPSS in order to show the effect of data preprocessing.

The thesis first introduces why people need data mining, what is data mining, what kind of data people mine, how to use data mining and what challenges in data mining i The second part compares data mining and relational technologies including data warehouse, OLAP, statistics and machine learning. Then a more detailed theoretical framework is suggested about the task and procedures in data mining especially measures of data preprocessing. At the end, a case study illustrates how to achieve data preprocessing by using SPSS. After preprocessing, the data is clean, integrated and reduced. As a conclusion of the experiment, SPSS can fulfill basically most of the data preprocessing tasks and give a better insight of the data.

KEYWORDS:

data mining, data preprocessing, DM, KDD, SPSS

TABLE OF CONTENTS

FIGURES

TABLES

1 Introduction to Data Mining

In recent years, the extensive application of database and computer networks coupled with the use of advanced automatic data generation and collection tools have made the amount of data increase rapidly. As the contradiction between the rapid increasement of data and hysteresis of data analysis methods more and more significant, people hope to use the analysis of large number of data to conduct scientific research and business decision or business management. Data mining precisely is developed to solve the lack of traditional analytical methods and deal with large-scale data . Data mining from large amounts of data to extract useful information hidden in the data is used in increasing number of areas, achieves good results, and provides a great help for decision-makers.

1.1 Background

It is well known that the human society has three critical elements: energy, materials, and information. From the industry age to the information age, knowledge from data has shown its extreme importance and dominance. The ability to obtain and store data easily with very low cost, and cheaper and more powerful s computers, make the amount of data increase exponentially. However, the traditional data analysis techniques are infeasible for raw data and as competition in human society is getting stronger, much of the data is never analyzed and buried in the data grave with undetectable information as shown in Figure 1. In this situation, extracting valuable wisdom from the vast expanse of the Information Ocean is a huge problem.

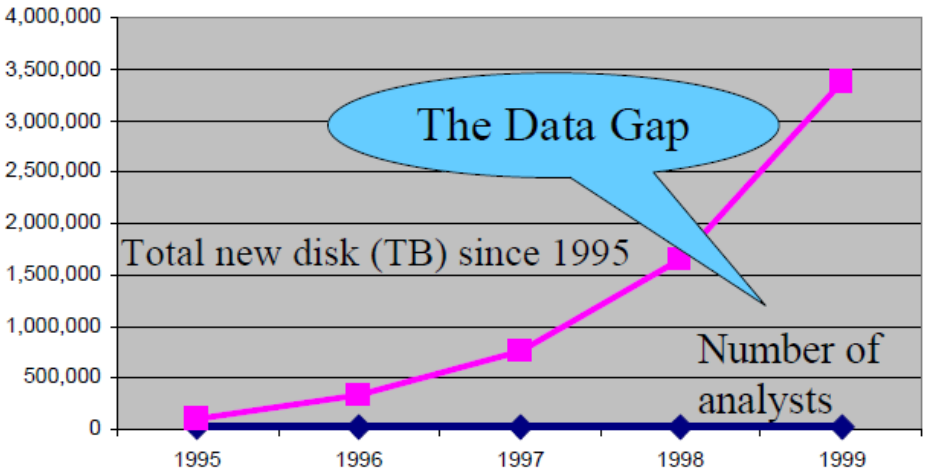


Figure 1. Big Data Gap (Grossman, 2001)

1.2 Definition

Many authors have suggested their own definitions of data miningsince the 1980s when data mining was first proposed as shown in Table 1. Briefly speaking, data mining refers to discovering meaningful patterns from large quantities of data by automatic or semi-automatic analyzing and exploring (Fayyad, Piatetsky-Shapiro, and Smyth 1996). It is also known as Knowledge Discovery in Databases (KDD) while the latter focus more on the whole process of knowledge discovering according to Fayyad's (1996) argument. As an interdisciplinary area, it incorporates analytical techniques including numerical analysis, neural networks and genetic algorithms, pattern matching and machine learning, and so on.

Table 1. Different definition of data mining (Xiaoli G, 2011)

Researchers	**Definitions**
SAS	The advanced method of data exploration and establishment of related models based on a large number of data.
Gartner	The process through careful analysis of large amounts of data to reveal.
Group	Meaningful new relationships, patterns and trends.
Aaron Zornes	The knowledge mining process from large databases to extract the operational information that we did not know before.
Fayyad	The important process to determine the effective, new and potentially information, and the model can be understood ultimately from the data.
Zekulin	Extracting previously unknown, understandable, actionable information from large databases.
Ferruzza	Used in the knowledge discovery process, some methods to identify unknown relationships and patterns existing in the data.
Jonn	Finding useful patterns while processing the data.
Parsay	A decision to support the process of studying the large data set for those unknown information models.
Bhavani	Finding meaningful new relationships, patterns and trends process in large amounts of data using pattern recognition technology, statistic and mathematical techniques.

1.3 Data Source

Different data sources lead to the need for different data mining tools. In fact, data mining projects often benefit from the use of several different types of data. In this way, the user can gain extra insight on accuracy and depth into project results. As suggested by SPSS Company (2005), now text mining and web mining are two major analysis directions. These two technologies make it easier to mine data known as "unstructured data" and log files from Web servers.

The sources used in data mining often have high capacity. For example, NASA's Earth observation satellites send data back at a speed up to 50GB per hour, and a supermarket deals with 20 million transactions every day. Most of data contain noise which means, they may be incorrect or incomplete in the real life. Besides, a source may have heterogeneous data which are a mixture of different data types, which is very common in data from Internet.

1.4 Application

The widespread use of barcode technology in the commercial accumulates large amounts of data every day, such as point of sale (POS) systems on the supermarket store tens of thousands of customer buying data everyday. Advanced modern scientific observation instruments result in a huge amount of data produced, for example, geostationary satellites send remote sensing image data of 50 giga (Gigabit) bytes per hour back to Earth. The rapid development of the Internet makes all kinds of resources on the network unusually rich which makes searching the information more difficult.

1.4.1 Scientific research

Data mining has very famous applications in astronomy: SKICAT (Sky Image Cataloging and Analysis Tool). SKICAT helps astronomers discover distant quasars. It is the first outstanding tool in data mining especially in application in astronomy and space science. By using SKICAT, astronomers have discovered 16 new extremely distant quasars (Weir et al., 1995).

The application of data mining in biology is mainly concentrated on molecular biologyespecially genetic engineering. As it has been suggested (Zhu and Davidson, 2007) that there are 10,000 different proteins, biological molecules series analysis method and gene database search technology have greatly contributed to many important discoveries.

1.4.2 Marketing

For marketing, it is very important to understand the patterns of the customer's shopping behavior. Now companies can improve competitiveness and promote sales via help of data analysis, specifically, database marketing and basket analysis. Database marketing selects potential customers by interactive query, data segmentation and model prediction. Through the analysis of existing customer data, users can be divided into different levels. The higher the level is, the greater the purchase likes. On the other hand, basket analysis identifies the customer's purchase behavior mode based on market sales data, POS database for instance. The working method is that if product A is bought, then the possibility of purchasing of product B is 95%. By finding out these relationships, the layout of the store shelves can be arranged better in order to promote products in a purposeful way (O'Brien and Marakas, 2006).

1.4.3 Fraud Detection

Fraudulence such as malignant overdrafts is always found in banks or commercial units , which has brought huge loss to the bank and commercial units. Fraud detection mainly focuses on finding the differences between normal behavior and fraud so that when a business complies with these characteristics the system can warn the policy makers. The FALCON and FAIS systems are very successful systems. The former one is developed by HNC and has been applied by a considerable number of retail banks for the detection of suspicious credit card transactions. The latter one is a system used to identify financial transactions related to money laundering which uses the government data form.

1.4.4 Internet applications

The rapid development of Internet and the global popularity of the Web make the information on the Web incomparably rich. Data from the Web is quite different from data from the database which has a specification structure. Information on the Web is mainly documents which are semi - structured or pure natural language text. The web mining tasks are Web information mining and Web user access pattern mining.

1.5 Challenges

As a young field, data mining has many challenges in researching and applying.

- Large databases: A common large database contains hundreds of tables and fields and thousands of records.
- Integration with other systems: Independent discovery in one system may not be very useful. Typical integration issues include integrating with database manager, integrating between spreadsheet and visualization tools and integrating with real-time reading sensors
- Determining the validity of statistics: There may be several possible models in a system.
- Over fit: When the algorithm finds the best parameters for a particular model with a limited data set, modeling can be applied not only for the general pattern data but also for proper noise, which results testing in a low-performance model.
- User interaction and prior knowledge: Some methods and tools are not truly doing interactive knowledge database discovery (KDD) and do not easily to absorb prior knowledge of the problem. The use of specialized knowledge in the whole KDD process is important.
- The complex relationship between fields: Better algorithms are required dealing with the relationship between attributes or attribute and value in hierarchical structure.
- The understood mode: It is important for many application systems to make it easier to be understood.
- Change of Data and knowledge: The rapidly changing data makes previously discovered patterns invalid. In addition, the measured variables in database can be modified, deleted, or conflict with new indicators.

- Missing data and noise: This problem is particularly serious in business databases. Important attributes may be lost if the designer of the database does not discover these problems.
- High dimension: The database has not only a large number of records but also a number of fields (attributes variables, etc...), so the dimension of the problem will be large.

2 Related Techniques VS Data Mining

Currently there are many so-called "data mining systems" on the market but many of them are merely tools based on statistical data analysis or machine learning. Data mining is an organic combination from a multi-discipline technologies including: database, mathematical statistics, high performance computing, neural networks, machine learning, space data analysis, pattern recognition, data visualization technology, image and signal processing, information retrieval and so on. By mining the data, meaningful knowledge, law, or a higher level of information can be found from the database. The found knowledge can support decisions, control processes, management of information and process queries. Therefore, data mining is considered as one of the most important frontline areas of research in the database system, and is also one of the most promising areas of database applications in the information industry (China BI, 2011).

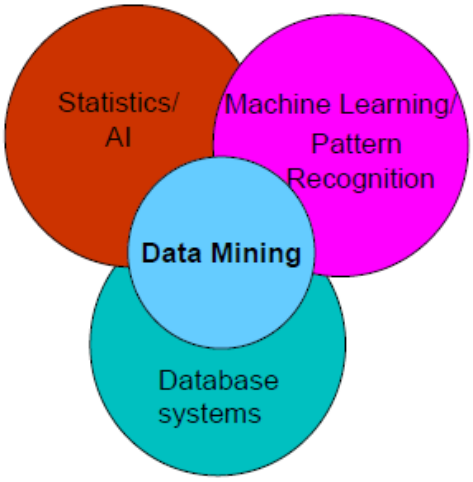


Figure 2. Data mining is an organic combination from a multi-disciplines technologies (KON, 2012)

2.1 Data warehouse

In fact, the history of development on computer and information technology is also the history of improvement and update of data and information processing means. In early

years, statistic analysis was carried out by artificial methods and summarized and reported by batch program. In the prevailing market circumstances, monthly and quarterly reports can meet information requirements for decision-making. However, with the amount of data grows and multiple data format being incompatible, it is necessary to integrate and store all the organization's data in uniform format, which becomes data warehouse. Different from the data database managing daily data, a data warehouse easily analyzes topic-oriented data which are integrated and time variant (Watson and Gray, 1997).

Figure 3. Data Warehouse Architecture (Watson and Gray, 1997)

2.2 Online analytical processing

After the emergence of the data warehouse, it is possible for more in-depth analysis of the data and the changes in the market have accelerated. Then emerges the online analytical processing (OLAP). OLAP is an online analysis system based on the connection of databases. It is a Business Intelligence (BI) solution including multidimensional consulting for large databases or tabulated data from the trading system. OLAP tools focus on understanding what already happened. They provide multidimensional data analysis which is better than the SQL calculated totals and straight through the multi-dimensional control. The main difference between OLAP and Data Mining is that the latter one is to generate hypothesis while the former one is to test and verify hypothesis. In other words, OLAP users have to explore data themselves and check whether they are right or wrong and data mining is the tool to help the user carrying out the exploration. Besides, OLAP tends to simplify and support interactive data analysis while the KDD tools' goal is to automate the process as much as possible.

From the perspective of data warehouse, data mining can be considered as an advanced stage of OLAP. However, the excellent data analysis capabilities let data mining far exceed the data warehouse OLAP. For example, according to reports and queries about the last month's total sales, OLAP shows sold products last month and data mining shows potential customers in the next month (SPSS Company, 2005).

2.3 Statistics and Machine Learning

Over the years, the mathematical statistic and artificial intelligence have provided a solid and rich theoretical and technical basis for deep database analysis tools such as data mining.

2.3.1 Statistics

Statistics and data mining both aim to discover structure in data but data mining does not intend to replace the traditional statistical analysis techniques. Instead, it is the extension and expansion of the statistical analysis methodology. Most of the statistical analysis techniques are based on impeccable mathematical theory and superb skills. The prediction accuracy is satisfactory but very demanding for the users. With the constant enhancement of computing power, it is possible to relatively simply and fixedly use the computing power of the computer to perform the same function. However, statistics plays a significant role in data mining especially during developing and accessing models. Most of the learning algorithms use a statistical test to correct over-fitting model, constructing rules or trees. The statistical tests used to evaluate machine learning algorithms and to verify the machine learning models (Hastie et al., 2005).

2.3.2 Machine Learning

As suggested by Langley and Simon (1995), machine learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience. Its goal is the continuous improvement of thr automation level in the knowledge engineering process in order to release people from the time-consuming work. It can discover and use the training data with more accurate and

efficient automated technology. The Machine learning algorithms' core includes not only the data mining process but also the following important steps:

- Establishment and maintenance of the database
- Formatting and cleansing data
- Summarization and visualization of data
- Formulating the input for learning algorithm and evaluating the found empirical rule by using the knowledge of experts
- Determination of deploying the results

3 Working Theory of Data Mining

The usage of data mining can help to acquire knowledge for decision-making. In many cases, the users do not know what the valuable information and knowledge are so a data mining system should be able to find several modes in order to meet expectations and actual needs. In addition, data mining systems should also be able to mine the knowledge on different abstract levels. Besides, a data mining system should allow users guide the mining search and finally obtain valuable mode of knowledge.

3.1 Task

There are two types of methods in data mining according to the target outcome. Prediction methods are to predict unknown or future values of other variables (Han et al., 2006). By combining advanced analytic techniques with decision optimization, a user can apply analytical results to determine the best actions. On the other hand, description methods are to find human-interpretable patterns describing the data.

To be specific, there are mainly 4 techniques for tasks in data mining projects as summarized in Table 2.

1. Classification [Predictive]

   Classification is to conceptual identify a class. It represents the overall information of such data and applies the description to construct a model represented by the rules or decision tree method. Records are sets of attributes and each of them is a class. The goal is to assign previously unseen records to a class as accurately as possible. Classification can be used to rule description and prediction.

2. Clustering [Descriptive]

   Clustering is to group the data into several categories according to the similarity where the same class of data is similar to each other

and data points in separate clusters are less similar to one another. Clustering can create macro-concepts, find the distribution pattern of the data and relationship between attributes. Clustering is a common descriptive task in determining a limited set categories or clustering to describe the data.

3. Association Rule Discovery [Descriptive]

   Association means there is regularity between two or more values of variables. It is an important task to discover knowledge. The association contains simple association, temporal association and causal association. The purpose is to find the hidden relationship in the database. The association is generally measured by support level and confidence level and also introduced interestingness and correlation to make the rules more conformances to requirements.

4. Deviation Detection [Predictive]

   There are many abnormal conditions in a database and it is very important to inspect them. The basic method is to find the difference between the observations with reference. Deviation detection contains methods based on statistics, distance, deviation and density. The abnormality detection is an important aspect of data mining. It is always applied in:

   - Fraud in telecommunications and credit card
   - Loan approval
   - Drug research
   - Weather forecast
   - The financial sector
   - Customer classification
   - Network Intrusion Detection
   - Fault detection and diagnosis

Table 2. Data mining tasks and techniques (Jackson, 2002)

DATA ANALYSIS TECHNIQUES	Data Summari-zation	Segmen-tation	Classifi-cation	Prediction	Dependency Analysis
Descriptive & Visualization	■	■			■
Correlation Analysis					■
Cluster Analysis		■			
Disciminant Analysis			■		
Regression Analysis				■	■
Neural Networks		■	■	■	
Case-Based Reasoning					■
Decision Trees			■	■	
Association Rules					■

3.2 Process

There are seven processes in the whole Knowledge Discovery in Databases (KDD) project where data mining is only a step as shown in Figure 4.

1. Data cleaning: to clear data noise and apparently unrelated data
2. Data integration: to combine together data from multiple data sources
3. Data selection: to select data related to the project task.
4. Data transformation: to convert data forms which easier for data mining.
5. Data mining: to use intelligent method to mine patterns or knowledge. It is the fundamental step in knowledge mining.

6. Pattern evaluation: to filter meaningful patterns knowledge according to certain interesting measures from mining results

7. Knowledge presentation: to show users the knowledge excavated via visualization and knowledge representation techniques.

The first four steps are, in fact, different steps of data prepreparation. The data mining step may have interaction with user or knowledge base. Interesting patterns generated in step 6 can be presented to the user and stored in knowledge base as a new knowledge. In this structure, data mining means only one step which reveals the hidden patterns to be assessed (Han et al., 2006).

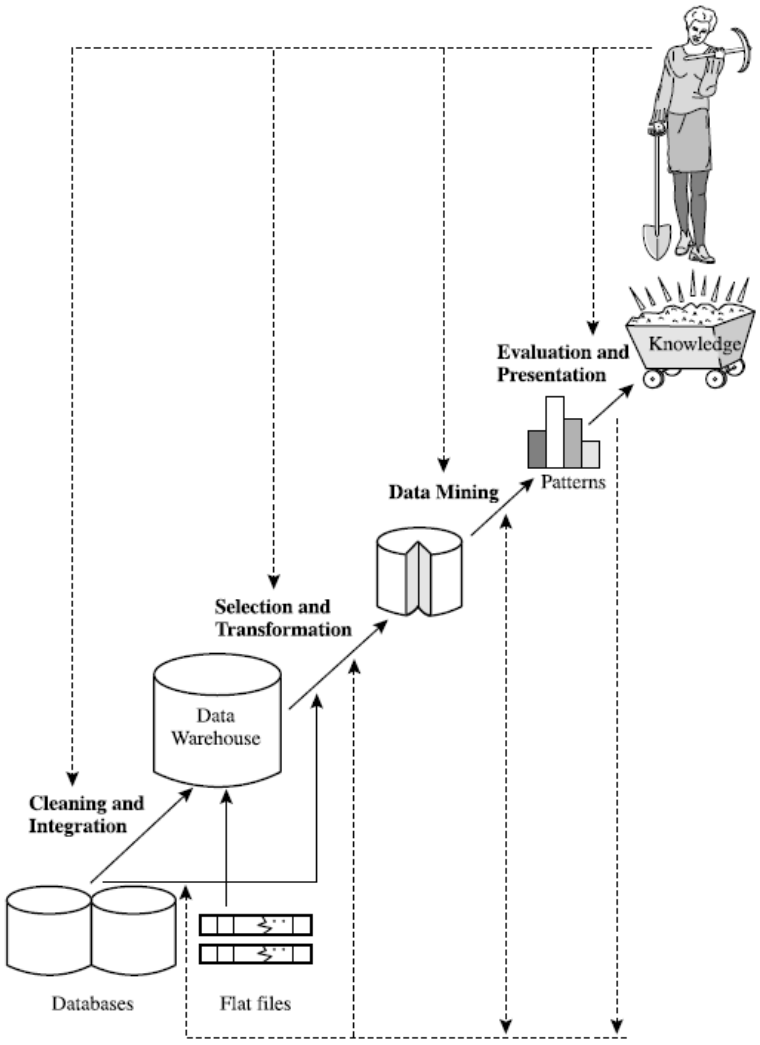


Figure 4. Data mining as a step in the process of knowledge discovery (Han et al., 2006)

3.3 Data preprocessing

After the data files have been created, an indispensable key step in data analysis processes is to prepare data for further analysis. In fact, according to Fernandez's study (2003) data preprocessing occupies about 70% of time spent on knowledge discovery project. If data have impurities, for example missing or duplicate data, data mining tools may be misled and even give wrong results. Based on wrong results, companies may make fatal decisions. Besides, preparing data is an integral part of building a data warehouse so that it integrates data of uniform quality.

There are four main tasks in the data preprocessing as suggested in Figure 5.

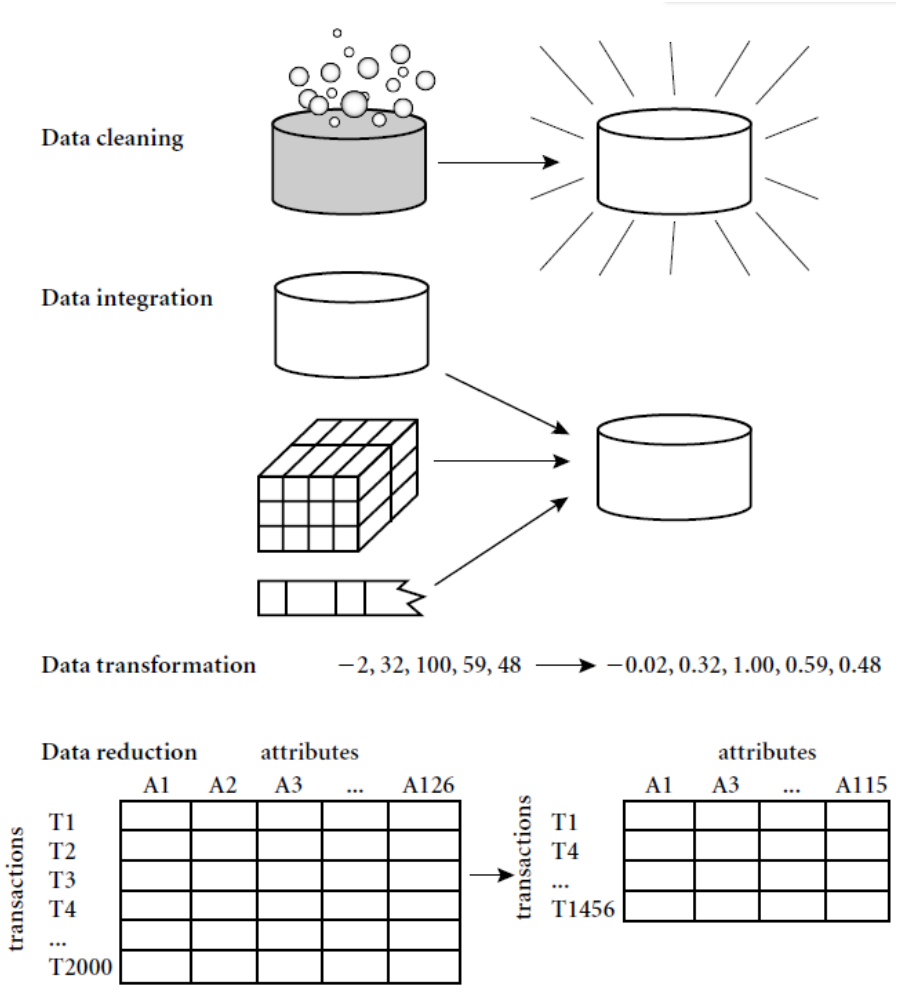


Figure 5. The four data preprocessing tasks (Han et al., 2006)

3.3.1 Data cleaning

In the real world, data are always accompanied by incomplete, noisy or inconsistent problems which would be handled in data cleaning process. Data may be missing because of missing collection, duplicate records or problematic equipment. Missing data can be ignored, filled in manually or automatically with a global constant. Noisy data refers to data with random error or variance in a measured variable. It may be caused during the data collection or transportation step or because of different locution or variant spellings. To solve this problem, the first method is binning which means sorting data and parting them in equal frequently bins and then can smooth bins by means, median or smooth boundaries. The other methods are regression, clustering and combination of computer and manual inspection.

3.3.2 Data integration and transformation

Data integration is combining data from multi sources to a consistent warehouse. Different data can be combined by matching the same records, for example, matching the primary key and foreign key or identifying based on common sense. In this process, there are always redundancies because a same object can have different values and should be removed by correlation analysis.

Numerical correlation analysis uses a correlation coefficient (also called Pearson's product moment coefficient)

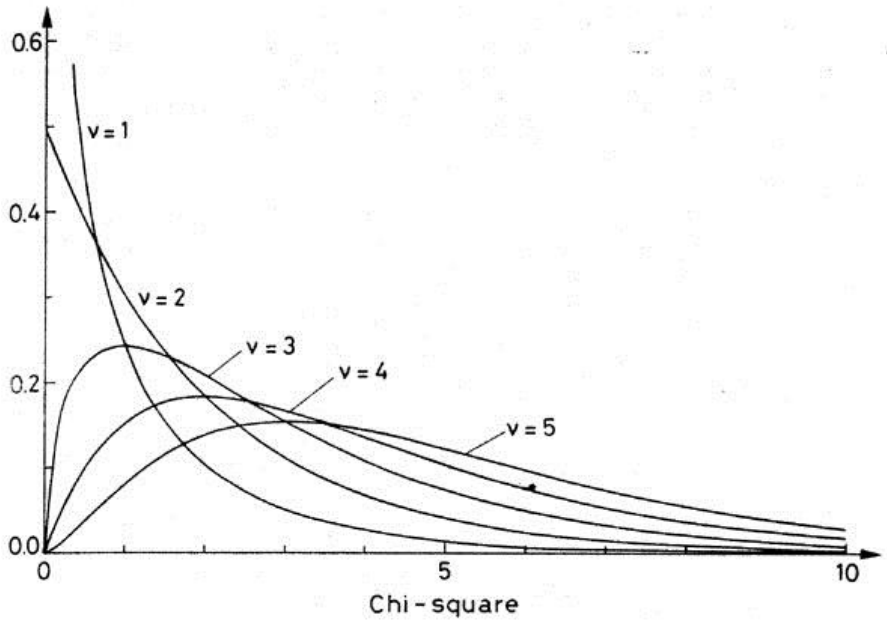
$$r_{A,B} = \frac{\sum\left(A-\overline{A}\right)\left(B-\overline{B}\right)}{\left(n-1\right)\sigma_A\sigma_B} = \frac{\sum\left(AB\right)-n\overline{A}\overline{B}}{\left(n-1\right)\sigma_A\sigma_B}$$

Where n is the number of tuples, $\overline{A}$ and $\overline{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, then Σ (AB) is the sum of the AB cross-product. If result is positive, A and B are positively correlated which means A

and B increase together. If the result is negative, A and B are negatively correlated. If result is 0, A and B are uncorrelated.

Categorical correlation analysis is using the Chi-square test where the larger the X2 value, the more likely the variables are related. The Chi-square test is based on Chi-square distribution as shown in Table 3.

Table 3. Chi Square Distribution (Leo, 1994)



Data transformation includes smoothing, aggregation, generalization, attribute or feature construction and normalization.

There are three kinds of normalization methods.

Min-max normalization: to [new_min_A, new_max_A]

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

Z-score normalization: where μ is mean and σ is standard deviation

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Normalization by decimal scaling: where j is the smallest integer such that Max (|v'|) < 1

$$v' = \frac{v}{10}$$

3.3.3 Data reduction

As a database or data warehouse may occupy terabytes of space and it will take quite a long time doing complex data mining, it is important to reduce the size of data set and obtainthe same or approximate analytical results. The methods of data reduction contain: (Pyle, Dorian, 1999)

- Data cube aggregation: reduces data cube and aggregates data for individual entity of interest
- Dimensionality reduction: e.g., removes unimportant attributes
- Data Compression:
  - String compression: usually lossless, has plenty theories and mature algorithms
  - Audio/video compression: usually loss, sometimes reconstructing small fragments instead of reconstructing the whole
  - Time sequence: usually short and changes slowly with time
- Numerosity reduction: e.g., fits data into models
- Discretization and concept hierarchy generation
  - Discretization: divides the continuous attribute within the range into intervals and assigns the interval labels to actual data values

so that it reduces the data size and easily applies some classification.

- Concept hierarchy formation: reduces the data via recursive method by collecting and replacing low level concepts by higher level concepts
- Typical methods:
  - Binning (covered above)
  - Top-down split, unsupervised,
  - Histogram analysis (covered above)
  - Top-down split, unsupervised
  - Clustering analysis (covered above)
  - Either top-down split or bottom-up merge, unsupervised
  - Entropy-based discretization: supervised, top-down split
  - Interval merging by χ^2 Analysis: supervised, bottom-up merge

4 SPSS Applications in Data Preprocessing

Data preprocessing is an important step in the KDD project not only because it takes a long period of time but also the four tasks may be executed several or zero times in the real application and there is no sequence between methods. From this point of view, data preprocessing is easy to be carried out by software like SPSS but need experience and data mining knowledge even some other professional knowledge if we want to achieve well prepared data.

4.1 Data mining software comparison

Nowadays there are many kinds of data mining tools in the market. The following are some famous data mining software packages with general purpose analysis (Abbott et al., 1998).

- IBM Intelligent Miner

  Pros: A fair degree of customization

  Scalability

  Index fast

  Advanced linguistic analysis, aggregation and filtering capabilities

  Powerful API library

  Ability of handling the huge amount of data

  Support the parallel processing

  Cons: GUI not friendly

  Need to be very familiar with UNIXS

  Batch operations is difficult

Metadata is not open

Lack of error code detailed explanation in documentation

No detailed description of algorithm.

- Oracle Darwin

  Pros: A high degree of scalability: enabled algorithms parallel implementation

  The model can be easily exported and integrated with other applications

  Windows-style client is easy to be used

  Cons: Lack of visualization exploration before data mining

  The workflow cannot be visual editing

- SAS Enterprise Miner

  Pros: A graphical interface

  Visualization operation

  Guidance for users without much experience to follow the principles of the SEMMA and mine successfully

  Cons: Temporary files spend much space

  Printing decision tree is difficult

- SPSS Clementine

  Pros: Friendly interface

  No require on programming

  Powerful statistics

  Cons: Has two environments "SPSS Data Editor" and “SPSS Output Navigator". The storing output and storing data are confusing.

- Silicon Graphics MineSet

  Pros: Data visualization

  Beautiful GUI

  Cons: No industry-specific custom applications

  RAM is very sensitive, needs at least 1G of memory to run in multi-threaded mode

4.2 Case study using SPSS

SPSS can select part of the sample in Data Editor Window according to the specified need of a user and analyze only the selected data until the user cancels selection. It can be used for data cleaning, data mining and statistical analysis.

Functions:

- Missing Value Analysis estimates the missing value by analysis the internal relations and models of large data sets.
- SPSS analysis generates database Scores.
- SPSS can create a graphical depiction of the model and exported to PowerPoint.
- SPSS Base contains several mining products: Answer Tree, Clementine and Goldminer. Specific technical: Kohonen neural network, regression, factor analysis, decision trees, aggregation, association rules, rule induction, monotonic regression, OLAP environment.
- SPSS for Windows migrates data from multiple data sources into a common data set for analysis including Basic Frequency Distributions to Correlations, regression and the more advanced Econometric Modeling.
- Clementine can find a model and convert into C language code.

In this case study, we used an SPSS Statistics Data Document '1991_U.S._General_Social_Survey.sav' as the original database and carried out some data prepreparation using SPSS. The first step, sorting, is a part of data cleaning which can give a better view of data and check missing data. The second step normalization is a part of data integration in order to find relationship between two variables. The third and fourth step visualization and calculation are parts of data reduction. They suggests histogram analysis and buils new variables and bin them into five groups, then in the following data mining processing, a user can deal with only one variable instead of three.

4.2.1 Data Sorting

Data sorting is to rearrange the data in the edit window according to one or more of the specified variable values in ascending or descending order. Ordering based on one sort variable is called single-valued sort. Ordering based on plural sort variables is called multiple sorts. By data sorting, a user is able to find missing data or outlines and initial analysis of data discrete degree according to the maximum and minimum values of the data.

Operation:

1. Choose 'Data' 'Sort Cases…'
2. Choose one variable for example 'Number of Brothers and Sisters' as the sort variable and order by ascending.
3. Check the 'Viewer' window, it records 'SORT CASES BY sibs (A)' where 'sibs' is the column name and A for ascending.

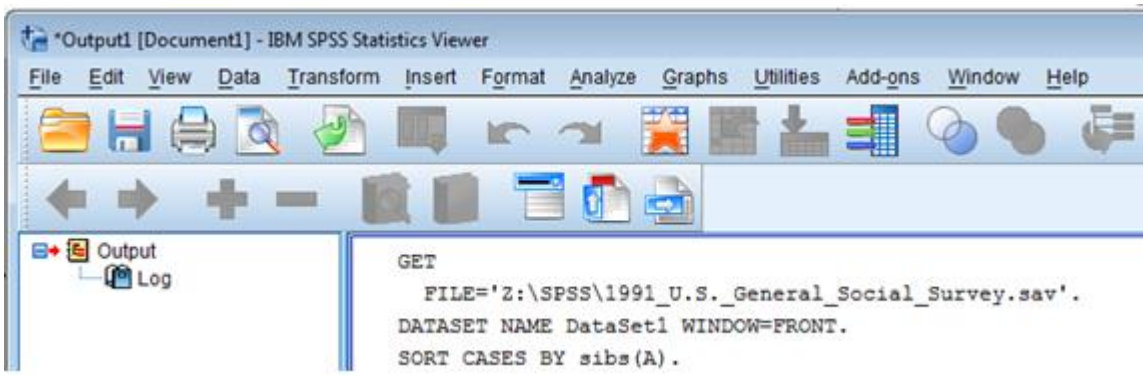
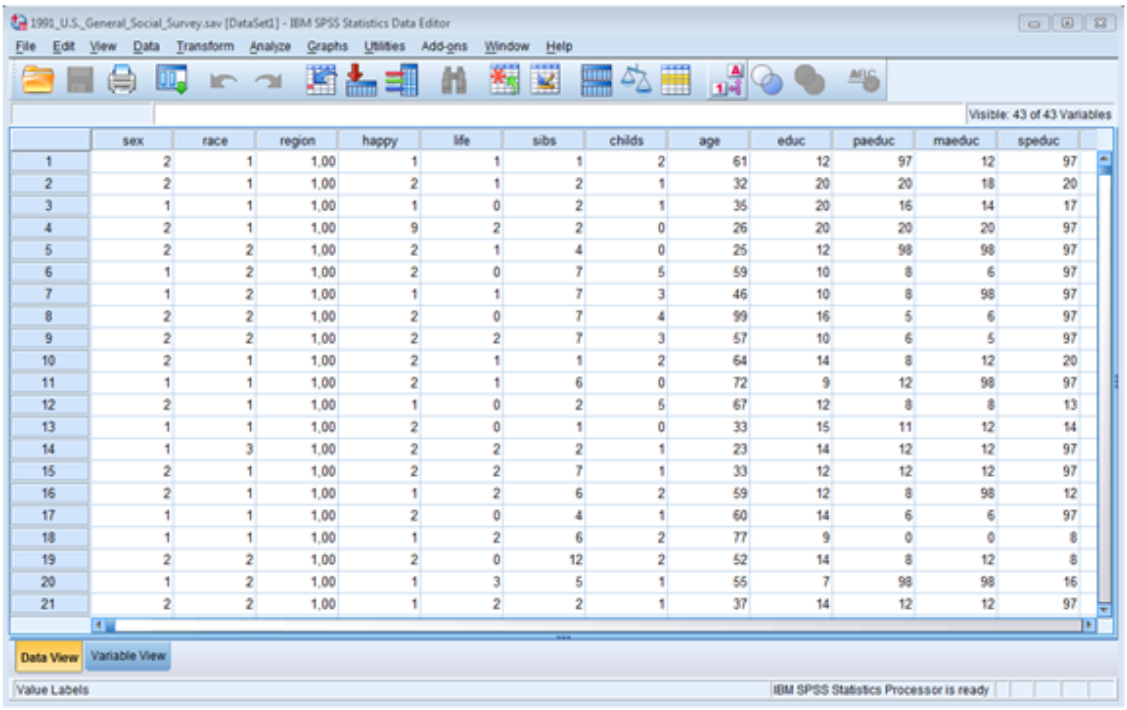


Figure 6. Single-valued sort Viewer

4. In the ‘Data Editor’ window, now the column ‘sibs’ is ordered as shown comparing to the former in Figure 7.

Before



After

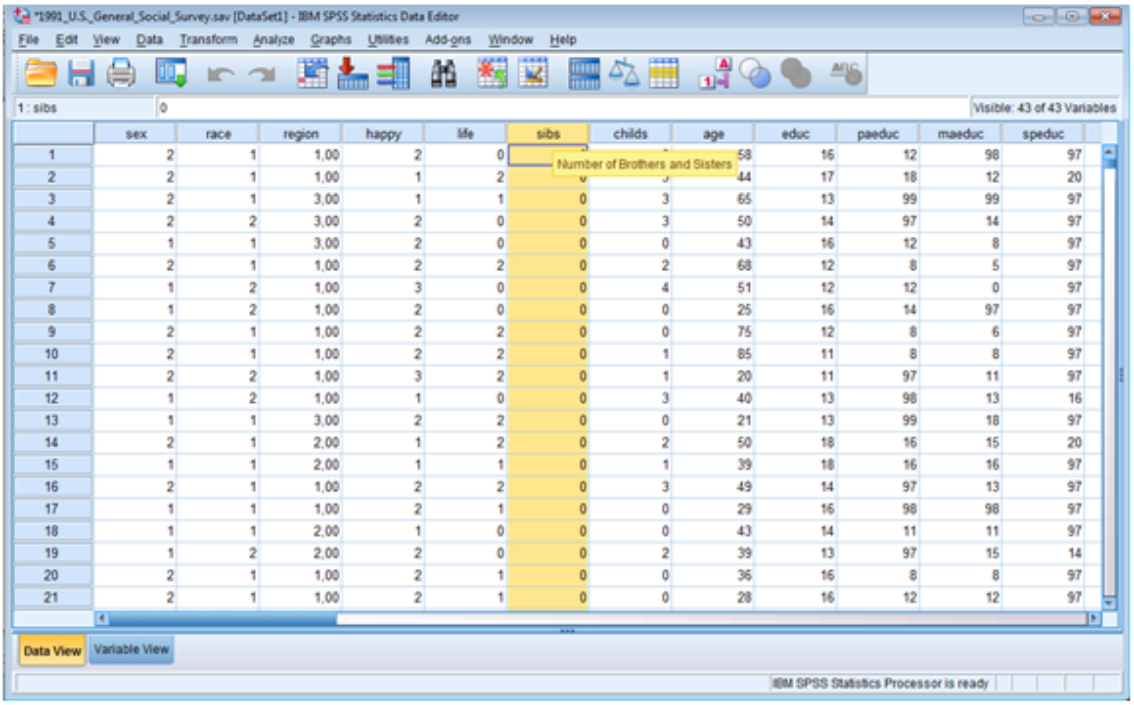
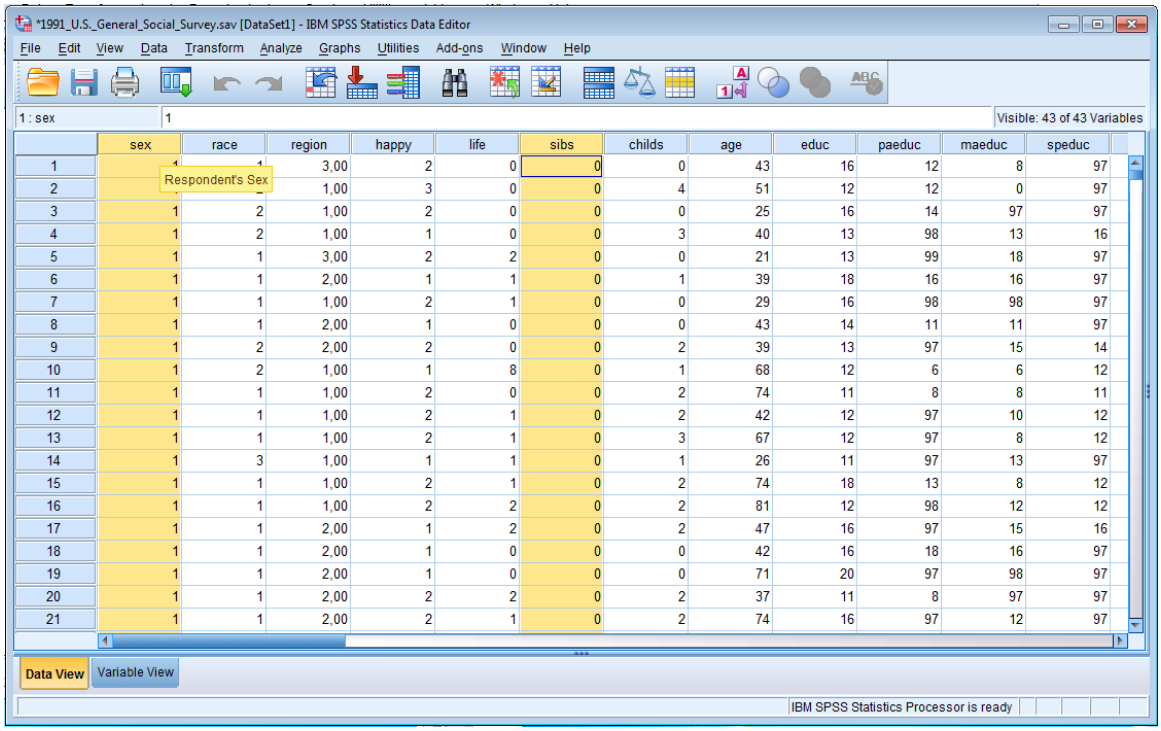


Figure 7. Single-valued sort Data View

5. For multiple sorts also choose ‘Data’ ‘Sort Cases…’ and then choose, for example, two variables ‘Respondent’s Sex’ and ‘Number of Brothers and Sisters’ by ascending order.

6. And here is the data view after multiple sort.



	sex	race	region	happy	life	sibs	childs	age	educ	paeduc	maeduc	speduc
1	1	1	3,00	2	0	0	0	43	16	12	8	97
2	1	2	1,00	3	0	0	4	51	12	12	0	97
3	1	2	1,00	2	0	0	0	25	16	14	97	97
4	1	2	1,00	1	0	0	3	40	13	98	13	16
5	1	1	3,00	2	2	0	0	21	13	99	18	97
6	1	1	2,00	1	1	0	1	39	18	16	16	97
7	1	1	1,00	2	1	0	0	29	16	98	98	97
8	1	1	2,00	1	0	0	0	43	14	11	11	97
9	1	2	2,00	2	0	0	2	39	13	97	15	14
10	1	2	1,00	1	8	0	1	68	12	6	6	12
11	1	1	1,00	2	0	0	2	74	11	8	8	11
12	1	1	1,00	2	1	0	2	42	12	97	10	12
13	1	1	1,00	2	1	0	3	67	12	97	8	12
14	1	3	1,00	1	1	0	1	26	11	97	13	97
15	1	1	1,00	2	1	0	2	74	18	13	8	12
16	1	1	1,00	2	2	0	2	81	12	98	12	12
17	1	1	2,00	1	2	0	2	47	16	97	15	16
18	1	1	2,00	1	0	0	0	42	16	18	16	97
19	1	1	2,00	1	0	0	0	71	20	97	98	97
20	1	1	2,00	2	2	0	2	37	11	8	97	97
21	1	1	2,00	2	1	0	2	74	16	97	12	97

Figure 8. Multiple sort Data View

4.2.2 Data normalization

Cross tabulation is to find out two certain variables’ relationship. In this part we used the Chi-square test to determine the relationship between the variables ‘Respondent's Sex’ and ‘General Happiness’.

Operation:

1. Choose ‘Analyze’, ‘Descriptive Statistics’, ‘Crosstabs…’
2. Choose the variable ‘Respondent's Sex’ as Row and ‘General Happiness’ as Column. Click the ‘Statistics’ box and choose ‘Chi-square’.

3. Check the ‘Viewer’ window, it shows the result of Chi-square test. From the table we can see that the value of χ^2 is 7.739 and degree of freedom is 2. The probability P ($\chi^2 \geq 7.739$) is 0.021which is less than 0.05. This shows that the variables ‘Respondent's Sex’ and ‘General Happiness’ are independent.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7,739[a]	2	,021
Likelihood Ratio	7,936	2	,019
Linear-by-Linear Association	4,812	1	,028
N of Valid Cases	1504		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 69,44.

Figure 9. Result of Chi-square Test

4.2.3 Data visualization

SPSS has an intuitive user interface and powerful tabular and graphical functions. The analysis result can be presented clearly via sixteen kinds of table format. In addition, it has a top graph analysis functions which can automatically generate graphics of statistical results as part of the analysis and draw graphics and graphical analysis independent from statistical process.

Operation:

1. Choose ‘Graphs’, ‘Legacy Dialogs’, ‘Histogram’
2. Choose variables ‘Number of Children’
3. Check the ‘Viewer’ window. It shows the histogram with the variable name, mean, standard division and number of cases.

```
GRAPH
  /HISTOGRAM=childs.
```

Graph

```
[DataSet1] Z:\SPSS\1991_U.S._General_Social_Survey.sav
```

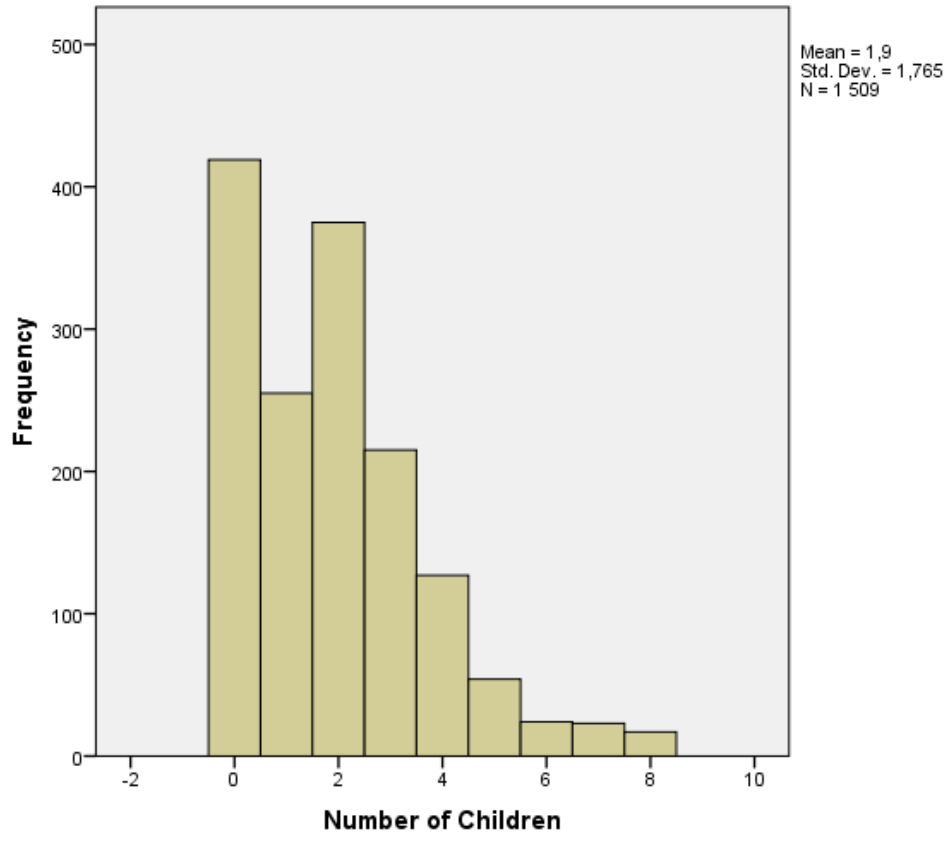


Figure 10. Histogram Viewer

4.2.4 Data calculation

Variable calculation is one of the most important and widely applied processes in the data analysis. It can generate new variables according SPSS arithmetic expressions and functions given by the user on the basis of the original qualified cases. As variable calculation is for all qualified cases, each case has its own result and the result should be saved to a specified variable. The variable data type should be consistent with the data type of the result calculated.

Variable calculation can handle mainly two tasks. The first one is data conversion. On the basis of the original data, calculations are made to produce new data with richer information and more intuitive. For example, according to the employees' basic wage, unemployment insurance and bonuses data, we can calculate the actual monthly income. Secondly, the calculation process can convert the original distribution state of the data. Some models in data analysis and modeling have certain requirements on the distribution of the data. Therefore, variable calculation can change the original distribution.

There are three kinds of variable calculation:

- Numeric expression is a formula consisting of constants, variables, arithmetic operators, parentheses, and function. In numeric expression, character constants should be enclosed in quotation marks. Variable refers to the original variable that already exists in the data. The arithmetic operators are +, -, *, /, ** (exponentiation) operation where the objects' data type is numeric. The constants and variables, data types in the same arithmetic expressions should be consistent; otherwise, they cannot be calculated.
- Conditional expression is a formula to judge conditions. The result is true or false depending on the determination of condition statement. Conditional expressions include simple conditional expressions and complex conditional expressions.
- The SPSS function is a pre-programmed and stored program in the SPSS software in order to achieve certain computational tasks. The specific form of the function is FUNCTION (PARAMETER) where the function name has already been provided in SPSS and a parameter can be a constant, variable or an arithmetic expression. Parameter may be one, or possibly more than one. The parameters are separated by commas.

Operation:

1. Choose ‘Transform’ ‘Compute Variable...’

2. In the ‘Target’ box, enter the name of the variable to store calculation results. If the specified variable is a new variable, SPSS will automatically create it. In case the specified variable already exists, SPSS will ask whether to overwrite the original values. The new variable default is numerical value and can be modified by clicking ‘Type & Label’ button.

3. Numeric expression is needed in the ‘Numeric Expression’ box. It can be entered manually or press the function drop-down menu. Here we add a variable named ‘Toeduc’ for ‘Total education year’ and plus the variables ‘educ’, ‘paeduc’, ‘maeduc’ and ‘speduc’ together.

4. In Data Editor, variable ‘Toeduc’ is calculated and shown in a new column.

	work4	work5	work6	work7	work8	work9	prob1	prob2	prob3	prob4	Toeduc
1	2	2	2	2	2	2	.	.	.	.	36,00
2	2	2	2	2	2	2	.	.	.	.	24,00
3	2	2	2	2	2	2	.	.	.	.	30,00
4	2	2	2	2	2	2	.	.	.	.	42,00
5	2	2	2	2	2	2	1	.	.	.	31,00
6	2	2	2	2	2	2	4	.	.	.	50,00
7	2	2	2	2	2	2	5	7	.	.	16,00
8	2	2	2	2	2	2	.	.	.	.	36,00
9	2	2	2	2	2	2	.	.	.	.	42,00
10	0	0	0	0	0	0	.	.	.	.	36,00
11	2	2	2	2	2	2	.	.	.	.	38,00
12	0	0	0	0	0	0	.	.	.	.	34,00
13	2	2	2	2	2	2	1	.	.	.	32,00
14	1	2	2	2	2	2	1	.	.	.	24,00
15	2	2	2	2	2	2	1	.	.	.	51,00
16	0	0	0	0	0	0	.	.	.	.	36,00
17	2	2	2	2	1	2	.	.	.	.	47,00
18	2	2	2	2	2	2	.	.	.	.	50,00
19	2	2	2	2	2	2	.	.	.	.	20,00
20	2	2	2	2	2	2	.	.	.	.	19,00
21	0	0	0	0	0	0	.	.	.	.	28,00

Figure 11. Data calculation SUM in Data View

5. Next we want to deal with the variable only meet certain conditions. For example, we divided the ‘Toeduc’ into five groups: Toeduc<26→1, Toeduc≥26 & Toeduc<36→2, Toeduc≥36 & Toeduc<46→3, Toeduc≥46 & Toeduc<56→4, Toeduc≥56→5.

First, create variable ‘group’ with label ‘Group for Toeduc’ and type is Numeric. Second, click the ‘If’ button and select ‘Include if case satisfies condition’, then enter the conditional expression Toeduc < 26 and set value equals 1 as shown in Figure 12. The not qualified cases will have no value of the variable if it had no value yet which means it will become a system-missing value or keep the old value. Third, add four other conditions in the same way.

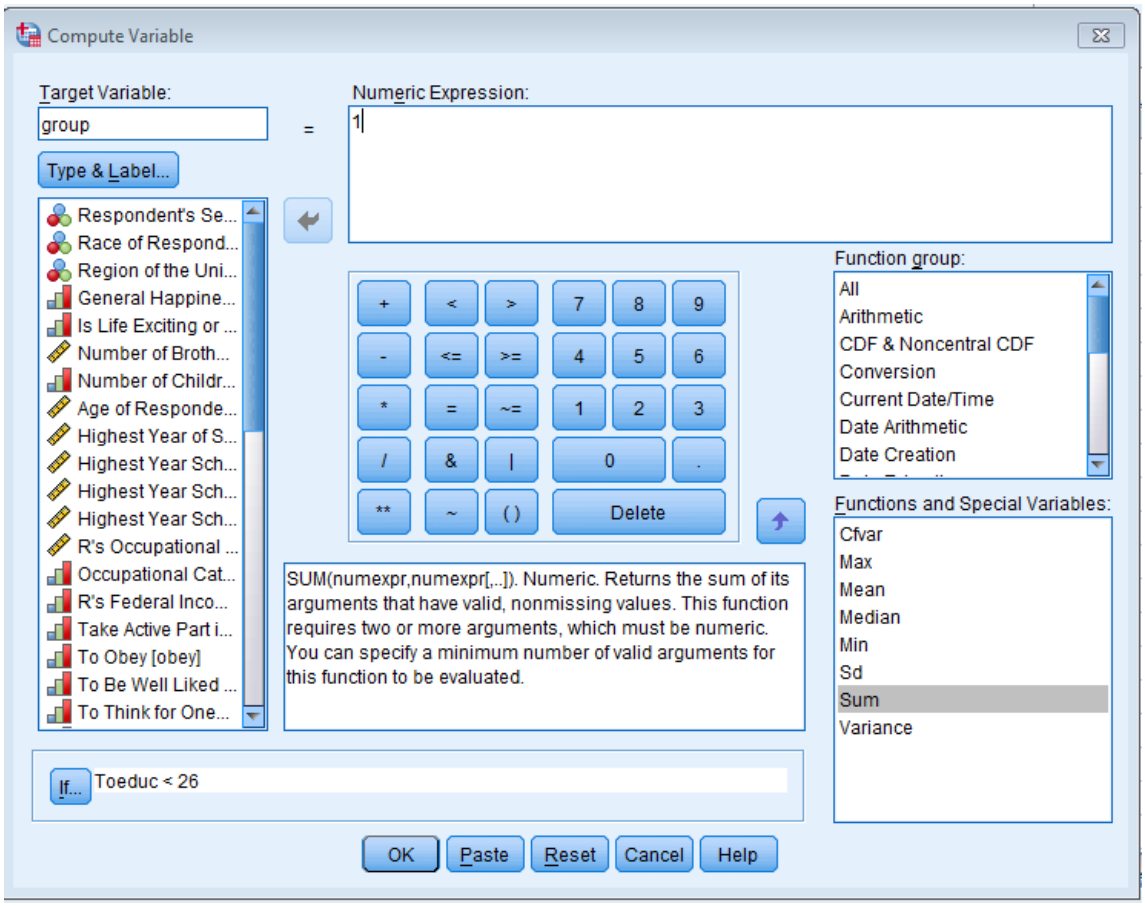


Figure 12. If Cases Configuration

Now check the Data Editor and the output Viewer and ensure there is no missing value in ‘group’.

```
COMPUTE Toeduc=SUM(educ,paeduc,maeduc,speduc).
VARIABLE LABELS  Toeduc 'Total education year'.
EXECUTE.
IF  (Toeduc < 26) group=1.
VARIABLE LABELS  group 'Group for Toeduc'.
EXECUTE.
IF  (Toeduc >= 26 & Toeduc < 36) group=2.
VARIABLE LABELS  group 'Group for Toeduc'.
EXECUTE.
IF  (Toeduc >= 36 & Toeduc < 46) group=3.
VARIABLE LABELS  group 'Group for Toeduc'.
EXECUTE.
IF  (Toeduc >= 46 & Toeduc < 56) group=4.
VARIABLE LABELS  group 'Group for Toeduc'.
EXECUTE.
IF  (Toeduc  >=  56) group=5.
VARIABLE LABELS  group 'Group for Toeduc'.
EXECUTE.
```

Figure 13. Calculation of qualified variable Results in Viewer

	work4	work5	work6	work7	work8	work9	prob1	prob2	prob3	prob4	Toeduc	group
1	2	2	2	2	2	2	.	.	.	.	36,00	3,00
2	2	2	2	2	2	2	.	.	.	.	24,00	1,00
3	2	2	2	2	2	2	.	.	.	.	30,00	2,00
4	2	2	2	2	2	2	.	.	.	.	42,00	3,00
5	2	2	2	2	2	2	1	.	.	.	31,00	2,00
6	2	2	2	2	2	2	4	.	.	.	50,00	4,00
7	2	2	2	2	2	2	5	7	.	.	16,00	1,00
8	2	2	2	2	2	2	.	.	.	.	36,00	3,00
9	2	2	2	2	2	2	.	.	.	.	42,00	3,00
10	0	0	0	0	0	0	.	.	.	.	36,00	3,00
11	2	2	2	2	2	2	.	.	.	.	38,00	3,00
12	0	0	0	0	0	0	.	.	.	.	34,00	2,00
13	2	2	2	2	2	2	1	.	.	.	32,00	2,00
14	1	2	2	2	2	2	1	.	.	.	24,00	1,00
15	2	2	2	2	2	2	1	.	.	.	51,00	4,00
16	0	0	0	0	0	0	.	.	.	.	36,00	3,00
17	2	2	2	2	1	2	.	.	.	.	47,00	4,00
18	2	2	2	2	2	2	.	.	.	.	50,00	4,00
19	2	2	2	2	2	2	.	.	.	.	20,00	1,00
20	2	2	2	2	2	2	.	.	.	.	19,00	1,00
21	0	0	0	0	0	0	.	.	.	.	28,00	2,00

Figure 14. Calculation of qualified variable Results in Data Editor

5 Summary

It is well-known that knowledge is fairly important and precious in this information age. For companies, discovering new knowledge earlier than competitors means a great competitive advantage and profits. Therefore, data mining emerged and has been improved rapidly to generate knowledge from huge amounts of data. However, as there are always redundant, missing, uncertain and inconsistent data in the real world, data mining cannot be executed before preprocessing. In the whole data mining process, data preprocessing is a large part spending up to 70% of project time and has a tremendous influence on the correction of the final result.

In this thesis, we focused on basic concepts and procedures in data mining as well as the importance and tasks of data preprocessing. For the practical part, it is obvious that by using software SPSS, the four major tasks in data preprocessing including data cleaning, integration, transformation and reduction can be easily carried out with no need for programming and can even make some predictions before the actual data mining.

REFERENCES

Fayyad U, Piatetsky-Shapiro G and Smyth P, (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3): 37.

SPSS Company. (2005) 'SPSS Data Mining Tips'.

Weir, N., Fayyad, U. M., Djorgovski, S. G., & Roden, J., (1995). The SKICAT system for processing and analyzing digital imaging sky surveys. Astronomical Society of the Pacific, 1243-1254.

Zhu X, Davidson I., (2007). Knowledge Discovery and Data Mining: Challenges and Realities. New York, NY: Hershey. p. 18.

O'Brien, J. A., Marakas, G. M., (2006). Management information systems. McGraw-Hill Irwin.

China BI, (2011). Analysis the difference between data mining and knowledge discovery. Available at: http://www.vsharing.com/k/2011-9/A649204.html (Accessed 5 June 2013)

Watson, H. J., Gray, P., (1997). Decision support in the data warehouse. Prentice Hall Professional Technical Reference.

Langley, P., Simon, H. A., (1995). Application of Machine Learning and Rule Induction. Communications of the ACM, (38:11), 55.64

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J., (2005). The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer, 27(2), 83-85.

Han, J., Kamber, M., and Pei, J., (2006). Data mining: concepts and techniques. Morgan kaufmann.

Fernandez, G., (2003). Data mining using SAS applications. CRC.

Pyle, Dorian, (1999). Data preparation for data mining. Vol. 1. Morgan Kaufmann.

Abbott, D. W., Matkovsky, I. P., and Elder IV, J. F., (1998, October). An evaluation of high-end data mining tools for fraud detection. In Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on (Vol. 3, pp. 2836-2841). IEEE.

Grossman, R. L. (2001). Data mining for scientific and engineering applications (Vol. 2). Springer.

KON, (2012). Introduction to Data Mining – Chapter 1: Introduction http://4.bp.blogspot.com/-7fni3gRPWTc/T_0IG9u5OCI/AAAAAAAAZfI/Z4SRIJuNTMs/s1600/1_3.png (Accessed 5 June 2013)

Xiaoli Geng, (2011). The application of data mining methods. Turku University of Applied Sciences.

Jackson, Joyce, (2002). Data mining: A conceptual overview. Communications of the Association for Information Systems, 8, 267-296.

Leo, W. R. (1994). Techniques for nuclear and particle physics experiments: a how-to approach. Springer Verlag.

Appendix

The Output of SPSS

```
GET
  FILE='Z:\SPSS\1991_U.S._General_Social_Survey.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
SORT CASES BY sibs(A).
SORT CASES BY sibs(A) sex(A).
CROSSTABS
  /TABLES=sex BY happy
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT
  /COUNT ROUND CELL.
```

Crosstabs

Notes

Output Created		29-APL-2013 14:28:57
Comments		
Input	Data	Z:\SPSS\1991_U.S._General_Social_Survey.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	1517
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics for each table are based on all the cases with valid data in the specified range(s) for all variables in each table.
Syntax		CROSSTABS /TABLES=sex BY happy /FORMAT=AVALUE TABLES /STATISTICS=CHISQ /CELLS=COUNT /COUNT ROUND CELL.
Resources	Processor Time	00:00:00,00
	Elapsed Time	00:00:00,02
	Dimensions Requested	2
	Cells Available	174762

[DataSet1] Z:\SPSS\1991_U.S._General_Social_Survey.sav

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Respondent's Sex * General Happiness	1504	99,1%	13	0,9%	1517	100,0%

Respondent's Sex * General Happiness Crosstabulation

Count

		General Happiness			Total
		Very Happy	Pretty Happy	Not Too Happy	
Respondent's Sex	Male	206	374	53	633
	Female	261	498	112	871
Total		467	872	165	1504

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7,739[a]	2	,021
Likelihood Ratio	7,936	2	,019
Linear-by-Linear Association	4,812	1	,028
N of Valid Cases	1504		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 69,44.

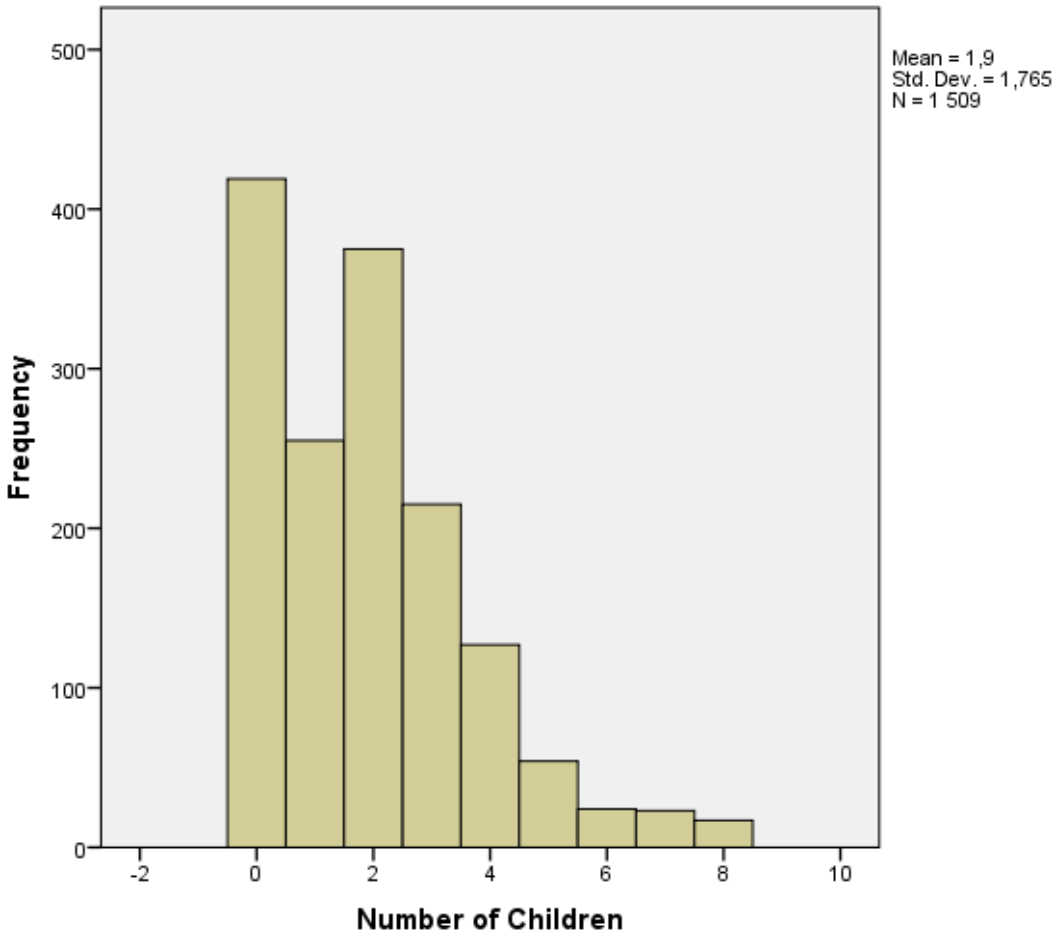
```
GRAPH
  /HISTOGRAM=childs.
```

Graph

Notes

Output Created		29-MAY-2013 14:30:50
Comments		
Input	Data	Z:\SPSS\1991_U.S._General_Social_Survey.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	1517
Syntax		GRAPH /HISTOGRAM=childs.
Resources	Processor Time	00:00:00,87
	Elapsed Time	00:00:00,85

[DataSet1] Z:\SPSS\1991_U.S._General_Social_Survey.sav




```
COMPUTE Toeduc=SUM(educ,paeduc,maeduc,speduc).
VARIABLE LABELS  Toeduc 'Total education year'.
EXECUTE.
IF  (Toeduc < 26) group=1.
VARIABLE LABELS  group 'Group for Toeduc'.
EXECUTE.
IF  (Toeduc >= 26 & Toeduc < 36) group=2.
VARIABLE LABELS  group 'Group for Toeduc'.
EXECUTE.
IF  (Toeduc >= 36 & Toeduc < 46) group=3.
VARIABLE LABELS  group 'Group for Toeduc'.
EXECUTE.
IF  (Toeduc >= 46 & Toeduc < 56) group=4.
VARIABLE LABELS  group 'Group for Toeduc'.
EXECUTE.
IF  (Toeduc  >=  56) group=5.
VARIABLE LABELS  group 'Group for Toeduc'.
EXECUTE.
```