DATA PREPARATION FOR DATA MINING BASED ON NEURAL NETWORK: A CASE STUDY ON GERMAN CREDIT CLASSIFICATION DATASET

Maria Ulfah Siregar
Teknik Informatika Fakultas Sains & Teknologi UIN Sunan Kalijaga Yogyakarta
(Email: ulfahtc96@yahoo.com)

Abstract

This paper will give detailed data description and preparation of German Credit Classification dataset, before it is used for further processes in data mining or data warehouse. Data preparation is the longest and most difficult part of data mining process. In general, readily available data is usually dirty and sometimes no quality data is available. There are five parts in data description and preparation that are going to be given in this paper. The first part is the name of the dataset and the number of examples and their types of attributes. In the second part, some examples from good and bad class as are given in the form of tables. Then, a data preliminary process is carried out to detect missing values from each attribute. Next, the result of statistical data analysis is displayed on charts or categories tables from each attribute. The last part is preprocessing, which comprise of data cleaning, integration and transformation. Based on the results obtained, three out of twenty attributes are deleted: Attribute 10, Attribute 18 and Attribute 20. So, the final data is smaller than the original one. Moreover, data is distributed more normally and in suitable patterns, which is hoped to be helpful for further processes.

Keywords: Data preparation, data, missing value, data cleaning, data integration, data transformation, statistical analysis.

A. Introduction

Data preparation is an important issue for pattern recognition, information retrieval, machine learning, web intelligence, data warehousing, data mining, and other areas of research that concerns with data. It has been understood that data preparation practically takes approximately 80% of the total data engineering effort (Zhang et al., 2003), so data preparation is a crucial part in data-based research. Yet, data preparation comprises techniques concerned with analyzing raw data so as to yield quality data (Zhang et al., 2003).

In this paper, data is prepared for further usage in data mining based on neural network. Data is explored in data mining to get the pattern of data and finally to predict something (Lukawiecki, 2007). Data mining is the search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amount of data (Sukumar). In other word, data mining can be used to extract knowledge from data (Negnevitsky, 2002).

In the real world, data tend to be incomplete, noisy, and inconsistent. In other word, data is dirty (Han & Kamber). Incomplete data means there are lacks of attribute values, lacks

of certain attributes of interest or only aggregate data. Meanwhile, noisy data means data containing either errors or outliers. And, inconsistent data means data containing discrepancies in codes or names. In addition, sometimes there are no quality data, and this will result in no quality mining results. We may say that data must be in 'clean' forms before it is processed.

Data preparation is the longest and most difficult part of the data mining process (Borysowich, 2007). It starts with data selection, maximization of quality of data, data mining approaches selection, data mining methods selection, and data transformation. However, this paper will only cover data cleaning, data integration, data transformation, data reduction, and data discretization.

Data cleaning handles data that are incomplete, noisy, and inconsistent. It will also fill in missing values, smooth noisy data, identify outliers, and correct data inconsistency. Meanwhile, data integration combines data from multiple sources to form a coherent data store, like data warehouse for example. Metadata, correlation analysis, data conflict detection, and resolution of semantic heterogeneity contribute towards smooth data integration. Data reduction obtains reduced representation in volume but produces the same or similar analytical results. Data discretization is part of data reduction but with a particular importance, especially for numerical data.

In this paper, data mining will derive some criteria classifying customer into good or bad classification. In the future, this classification can be used to process acceptation or rejection of loan requests from customer. For 'good customer', bank almost certainly process and accept his/her loan requests because of his/her lower credit risk. Otherwise, bank may need extra data or further assessment. From this data, we are looking for 'clues' that bring us to the decision of whether a customer can be classified into the good or the bad class.

Customers, which are classified into good class, have certain data that make them classified into this class. We can adjust some relationships such as has good history in previous credits, has property which safe the loan, has stable account, has good job, and etc so as he/she can be classified into one of two these classes .

Credit risk management is an important and crucial part of financial and regulation institutions due to measurement and attenuation of unexpected losses for the credit display in a portfolio (Mihail et al.). The recent tarnishing bankruptcies in US and others countries proof its importance and existence. By using credit risk management, institutions will be aware in their decisions to whether to sanction a loan request or not. The assessment of the loan proposal requires a lot of expertise, experience and systematic approach (Kulkarni & Sreekantha, 2008).

The following are some of benefits of credit risk measurement (Kulkarni & Sreekantha, 2008):

1. It facilitates informed credit decision consistent with Bank's Risk appetite
2. It provides ability to price products on the basis of risk
3. It facilitates dynamic provisioning and minimizes impact of losses

- A174: management/self-employed/highly qualified employee/officer

- Attribute 18: number of people being liable to provide maintenance for
- Attribute 19: telephone
  - A191: none
  - A192: yes, registered under the customers name
- Attribute 20: foreign worker
  - A201: yes
  - A202: no

3. Source of Data Before Preprocessing

In order to give a deeper understanding on the dataset, five examples of upper lines for each class are depicted in separate tables below.

Table 1 : Examples from good class

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
No-account	42	duly-till-now	radio-tv	7166	unknown	seven-years	2	married-male	None	4
No-account	18	duly-till-now	radio-tv	1126	unknown	one-year	4	female-divorced	None	2
less-200DM	24	duly-till-now	furniture	4351	unknown	four-years	1	female-divorced	None	4
0DM	12	duly-till-now	education	1200	unknown	four-years	4	female-divorced	None	4
No-account	12	duly-till-now	radio-tv	1963	less100DM	seven-years	4	single-male	None	2

Table 2 : (Continued)

A12	A13	A14	A15	A16	A17	A18	A19	A20	Target
building-society	29	None	rent	1	skilled	1	no	yes	good.
real-estate	21	None	rent	1	skilled	1	no	yes	good.
building-society	48	None	own	1	unskilled-resident	1	no	yes	good.
building-society	23	Bank	rent	1	skilled	1	no	yes	good.
car	31	None	rent	2	management	2	no	yes	good.

Table 3 : Examples from bad class

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
less-200DM	12	duly-till-now	repairs	639	less100DM	four-years	4	single-male	none	2
less-200DM	24	duly-till-now	Used-car	11560	less100DM	four-years	1	female-divorced	none	4
0DM	9	duly-till-now	radio-tv	1366	less100DM	one-year	3	female-divorced	none	4
0DM	48	duly-till-now	Used-car	10297	less100DM	seven-years	4	single-male	none	4
0DM	24	duly-till-now	furniture	3345	less100DM	over-seven	4	single-male	none	2

Table 4 : (Continued)

A12	A13	A14	A15	A16	A17	A18	A19	A20	Target
car	30	none	own	1	skilled	1	yes	yes	bad.
car	23	none	rent	2	management	1	yes	yes	bad.
Building-society	22	none	rent	1	skilled	1	yes	yes	bad.
none	39	stores	free	3	skilled	2	no	yes	bad.
Building-society	39	none	rent	1	management	1	no	yes	bad.

4. List of Attributes With Their Types (Categorical, Nominal, or Continuous)

In this dataset, there are only two types of attributes, categorical and continuous.

- A62: 100 d" ... < 500 DM
- A63: 500 d" ... < 1000 DM
- A64: ... e" 1000 DM
- A65: unknown/no savings account

- Attribute 7: present employment since
  - A71: unemployed
  - A72: ... < 1 year
  - A73: 1 d" ... < 4 years
  - A74: 4 d" ... < 7 years
  - A75: ... e" 7 years
- Attribute 8: installment rate in percentage of disposable
- Attribute 9: personal status and sex
  - A91: male : divorced/separated
  - A92: female: divorced/separated
  - A93: male : single
  - A94: male : married/widowed
  - A95: female: single
- Attribute 10: other debtors/guarantors
  - A101: none
  - A101: co-applicant
  - A103: guarantor
- Attribute 11: present residence since
- Attribute 12: property
  - A121: real estate
  - A122: if not A121: building society savings agreement/life insurance
  - A123: if not A121/A122: car or other, not in attribute 6
  - A124: unknown/no property
- Attribute 13: age in years
- Attribute 14: other installment plans
  - A141: bank
  - A142: stores
  - A143: none
- Attribute 15: housing
  - A151: rent
  - A152: own
  - A153: for free
- Attribute 16: number of existing credits at this bank
- Attribute 17: job
  - A171: unemployed/unskilled – non-resident
  - A172: unskilled – resident
  - A173: skilled employee/official

4. Lending decisions can be taken with minimum time, thus lending business volume can increase substantially.

B. Method

Data preparation is initiated with the following data descriptions:

1. Name of Dataset

German Credit Classification dataset, obtained from the UCI (University of California, Irvine) Machine Learning Repository, was used in this study. We chose this dataset because of its number of examples is sufficient and its values for each attribute are complete or available.

2. Brief Description

The number of examples in the dataset is 1000. The dataset is classified into two classes: good and bad class. The good class has 700 examples whereas the bad one has 300. The dataset has 20 attributes, each labeled from A1 to A20. Seven of the attributes are of continuous (numerical) types, while the other 13 are of categorical types. The description of each attribute is given as follow:

- Attribute 1: status of existing checking account
  - A11: ... < 0 DM
  - A12: 0 d" ... < 200 DM
  - A13: ... e" 200 DM/salary assignments for at least 1 year
  - A14: no checking account.
- Attribute 2: duration of loan in month
- Attribute 3: credit history
  - A30: no credits taken/all credits paid back duly
  - A31: all credits at this bank paid back duly
  - A32: existing credits paid back duly till now
  - A33: delay in paying off in the past
  - A34: critical account/other credits existing (not at this bank).
- Attribute 4: purpose
  - A40: car (new)
  - A41: car (used)
  - A42: furniture/equipment
  - A43: radio/television
  - A44: domestic appliances
  - A45: repairs
  - A46: education
  - A47: vacation
  - A48: retraining
  - A49: business
  - A410: others
- Attribute 5: credit amount
- Attribute 6: savings account/bonds
  - A61: ... < 100 DM

Table 5 : Attributes and its types

Status (qualitative - categorical).	Residence-time (numerical): continuous.
Duration (numerical): continuous.	Property (qualitative – categorical).
Credit-history (qualitative – categorical).	Age (numerical): continuous.
Purpose (qualitative – categorical).	Installments (qualitative – categorical).
Credit (numerical): continuous.	Housing (qualitative – categorical), values that accepted: rent, own, free.
Savings-account (qualitative – categorical).	Existing-credits (numerical): continuous.
Employment (qualitative – categorical).	Job (qualitative – categorical).
Installment-rate (numerical): continuous.	Liable-people (numerical): continuous.
Personal-status (qualitative – categorical).	Telephone (qualitative – categorical).
Debtors (qualitative – categorical).	Foreign (qualitative – categorical).

After we examine the whole data, it is found that there are no missing values for all attributes and objects. Detection of missing values and fill in is one-step of data cleaning. Missing values here refers to some data that may be coded as a blank or assigned a special value that cannot possibly occur. This error may occur if the data was not obtained from observation or if it was lost.

The next step in this study is statistical analysis of the data. For categorical types of data, the analysis can be easily carried out based on frequency. After we obtain frequency tables for each categorical attribute, we can plot the frequencies onto some graphical tools. In this paper, we use help from chart of Excel and choose bar chart to represent them. But here, we just describe one attribute, A1, in detail. The other attributes will be depicted in table forms.

- **A1**

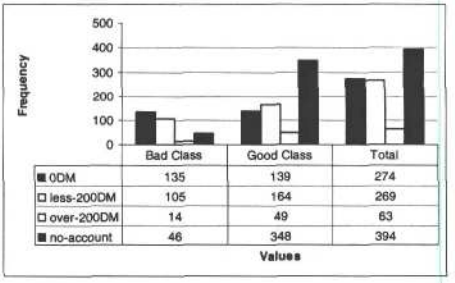

Figure 2 : Attribute status before preprocessing

Although data distribution is not normal, we did not find any outliers. Outliers are data from the same attribute which are very different from the rest of the whole data being observed. Outliers can be detected by examining frequency table, plot of chart, or cluster analysis. If cluster analysis is used, the outliers will lay outside of clusters. To obtain a normal distribution for the data, one of the alternatives is to combine the category of less-200DM and over-200DM. The rest of the attributes of categorical types are listed in following table:

Table 6 : Attributes and its suggestions

Attribute	Suggestions
A3	Categories *all-paid-duly* and *bank-paid-duly* can be combined into one category
A4	• *Vacation* category should be better ignored and deleted. • Categories *retraining*, *domestic-app* and *repairs*, can be integrated into one category, which is *others*.
A6	*less 1000DM* and *over1000 DM* can be combined into new category, *greater than or equal 500DM*
A7	• Categories *one-year* and *four-years* can be combined into new category, *less than four years*. • Still, categories *seven-years* and *over-seven* can also be combined into new category, *greater than or equal four years*.
A9	Integrate its categories into just two categories, one for *male* and one for *female*
A10	Deleted
A12	No modification
A14	*bunk* and *stores* categories can be combined into one category
A15	*rent* and *free* can be combined into one category.
A17	*unemployed/unskilled – non-resident, unskilled – resident,* and *management/self-employed/highly qualified* employee/officer are combined into one category.
A19	No modification
A20	Deleted

For continuous data type, it is difficult to do straight analysis by frequency, especially if the data varies strongly. To solve this problem, it is better to put them into interval categories, as summarized in the following table:

Table 7 : Analyzation of continuous types attributes

Attribute	Analyze
A2	Data '72' might be an outlier. Its value has very large difference from the others. In addition, its frequency is only one, so small. This object might come from another population, which has large duration.
A5	Examining attribute credit, data '18424' has range that far from previous data, while the other values differ smoothly each other.
A8	The data distribution is quite good. Moreover, no such outliers are found.
A11	Same with A8
A13	Same with A8
A16	Although the distribution is out of normal, there is no such outlier found.
A18	The distribution of data of this attribute is not normal. We decide by deleting this attribute will not interfere other attributes.

Univariate examination sometimes is not sufficient to identify all errors exist in the data. Therefore, attributes should be cross-related in order to determine whether there is any relationship between attributes which is commonly known not to be true.

For categorical attribute, we can use cross-tabulation to do this. Here, we perform cross-tabulation to the "job" and "employment" attributes since the relationship between them is generally known.

Table 8 : Cross-tabulation between job and employment attributes

		Attribute employment				
		unemployed	one-year	four-years	seven-years	over-seven
Attribute job	unemployed-non-resident	16	5	1	0	0
	unskilled-resident	1	44	80	34	41
	skilled	12	107	230	119	162
	management	33	16	28	21	50

From table 8, it can be seen that the five unemployed-non-resident who having their job for one year, one unemployed-non-resident who has his/her job for four years, the one unskilled-resident employee who is unemployed, 12 skilled employees who are unemployed, the 33

employees with management job who are unemployed, are known not to be true relationships. All of these errors are difficult to find out if it is just based on univariate examining.

For continuous attribute, scatter graph can be used to analyze the correlation between attributes (Kannan). There are two steps in determining the existence and degree of linear association between two attributes (Afifi & Azen, 1979). To do this, first, we can plot the points of (x_1attr_1, y_1attrr_1), (x_2attr_1, y_2attrr_1), ..., (x_nattr_1, y_nattrr_1) in the x-y plane. The resulted graph is called a scatter gram/scatter graph. After that, the simple coefficient of correlation can be calculated using the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}}$$

The correlation coefficient value (r) lays between the interval of -1 d" r d" 1. The closer the correlation coefficient value to -1 or 1, the stronger the relationship between the attributes. The positive value of r means that when the value of attribute 1 increases, the value of attribute 2 also increases, which proves that they have a positive correlation. Meanwhile, the negative value of r means that when the value of attribute 1 increases, the value of attribute 2 decreases. The extreme cases occur when the $r = \pm 1$, which indicate a perfect linear association between the two attributes, where if the value of attribute 1 is given, the value of attribute 2 can be determined exactly.

The below diagram, Figure 2, shows that duration and credit is related as positive linear. And, it is also reflected from its correlation value, 0.62, which is almost 1. There are up to 15 pairs of correlation among continuous attributes that can be drawn from the dataset. However, only one correlation displayed graphically in this paper, while the other 14 correlations are summarized in a correlation matrix, which is depicted in Table 9.

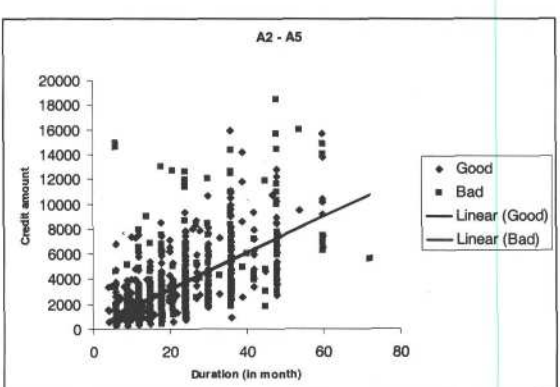

Figure 2 : Scatter graph of attributes A2 and A5

Table 9 : Correlation matrix for attribute A2 to A16

	A2	A5	A8	A11	A13	A16
A2	1					
A5	0.624984	1				
A8	0.074749	-0.2713157	1			
A11	0.034067	0.028926323	0.04930237	1		
A13	-0.03614	0.032716417	0.05826568	0.26641918	1	
A16	-0.01128	0.020794552	0.02166874	0.08962523	0.149253582	1

After statistical analysis of the data, the next step is data preprocessing, which begins with data cleaning. If there are missing values, we can either fill in the missing values manually or delete the object with missing values found in more than one attributes. The value filled in can be obtained from the global constant, the attribute mean, or the most probable value. However, since the analysis indicates that the dataset does not contain any missing value, this process was not carried out.

If the data varies largely, we can smooth the noise in the data using binning technique. To do this, firstly, we need to sort the data, and then distribute it into a number of bins. After that, the data in each bin is smoothed either by bin means, bin medians or bin boundaries. For categorical attributes, some of the categories can be combined to obtain a normal distribution. In addition, a few of the attributes considered meaningless are deleted. Results are depicted in Table 10.

Table 10 : Categorical attributes after preprocessing

Attribute	Existence	Massaged	Detail Real Value	Frequency
A1	Exist	0	less 0DM	274
		0.5	greater than or equal 0DM	332
		1	no-account	394
A3	Exist	0	paid-duly	89
		0.333	duly-till-now	530
		0.667	delay	88
		1	critical	293
A4	Exist	0	new-car	234
		0.167	used-car	103
		0.333	furniture	181
		0.5	radio-tv	280
		0.667	education	50
		0.833	business	97
		1	others	55
A6	Exist	0	less 100DM	603
		0.333	less 500DM	103
		0.667	greater than or equal 500DM	111
		1	unknown	183
A7	Exist	0	unemployed	62
		0.5	less four years	511
		1	greater than or equal four years	427
A9	Exist	0	male	690
		1	female	310
A10	Not exist			
A12	Exist, no need further preprocessing	0	real-estate	282
		0.333	building-society	232
		0.667	car	332
		1	none	154
A14	Exist	0	none	814
		1	Exist	186
A15	Exist	0	own	713
		1	either rent or free	287
A17	Exist	0	skilled	630
		1	unemployed-non-resident or unskilled-resident or management	370
A19	Exist, no need further preprocessing	0	No	404
		1	Yes	596
A20	Not exist			

For continuous attributes, we will describe them one by one separately. The duration attribute is partitioned into 3 bins using equal-depth (frequency) partitioning method. Three intervals are chosen because they give almost the same number of examples. The equal-depth partitioning method is used here since the data is skewed. After the data was partitioned into 3 bins, it was smoothed by bin means which can be seen as follows:

Table 11 : Duration attribute (A2)

Interval	Number of examples	Bin means
4 – 12	359	9.86
13 – 24	411	20.4
25 – 72	230	39.05

The credit attribute is partitioned into 7 bins to accommodate the same number of examples in each bin. After partitioned and smoothed by bin means, we obtain the following:

Table 12 : Credit attribute (A5)

Interval	Number of examples	Bin means
250 – 1000DM	116	723.534
> 1000 & ≤ 1500DM	190	1280.395
> 1500 & ≤ 2000DM	126	1746.571
> 2000 & ≤ 3000DM	188	2460.697
> 3000 & ≤ 4000DM	134	3491.284
> 4000 & ≤ 7000DM	141	5381.965
> 7000DM	105	9854.143

The installment-rate attribute is partitioned into 4 bins using Equal-width (distance) partitioning method. By substituting N, B (max value) and A (min value) in the width interval formula with 4(four), 4 (four) and 1 (one) respectively, we obtain:

$$W = \frac{B - A}{N}$$

$$= \frac{4 - 1}{4} = 0.75$$

Since the data in each interval is already uniformed, we do not need to perform data smoothing.

Table 13 : Installment-rate attribute (A8)

Interval	Member of examples
1 – 1.75	136
1.76 – 2.5	231
2.51 – 3.25	157
3.26 – 4	476

The residence-time attribute is partitioned into 4 bins, also using Equal-width (distance) partitioning method. And the same as before, we can substitute N, B (max value) and A (min value) in the width interval formula with 4(four), 4 (four) and 1 (one) respectively to obtain:

$$W = \frac{B - A}{N}$$

$$= \frac{4 - 1}{4} = 0.75$$

For this attribute, we do not need to perform data smoothing since the data in each interval is already uniformed.

Table 14 : residence-time attribute (A11)

Interval	Member of examples
1 – 1.75	*130*
1.76 – 2.5	308
2.51 – 3.25	149
3.26 – 4	413

Same as before, age attribute is also partitioned into 3 bins. The reasons are not only due to the same number of member examples in each interval but also the age. After partitioned into 3 bins and smoothed by bin boundaries, the following table is obtained:

Table 15 : Age attribute (A13)

Interval	Number of member	Data before smoothing	Data after smoothing
19 – 30	411	{19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30}	19 = 149 30 = 262
31 – 40	315	{31, 32, 33, 34, 35, 36, 37, 38, 39, 40}	31 = 177 40 = 138
41 – 75	274	{41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 70, 74, 75}	45 = 226 75 = 48

The existing-credits attribute is partitioned into 2 bins using Equal-depth (frequency) partitioning method. Then, we perform data smoothing using bin boundaries.

Table 16 : Existing-credits attribute (A16)

Interval	Member of examples	Data
≤ 1	633	1 = 633
2 - 4	367	2 = 333 4 = 34

After preprocessing is done, the results in the dataset are different from the original ones, as summarized in the following tables:

Table 17 : Examples from good class after preprocessing

A1	A2	A3	A4	A5	A6	A7	A8	A9
1	39.05	0.333	0.5	9854.143	1	1	2	0
1	20.4	0.333	0.5	1280.395	1	0.5	4	1
0.5	20.4	0.333	0.333	5381.965	*1*	*0.5*	*1*	*1*
0	9.86	0.333	0.667	1280.395	1	0.5	4	1
1	9.86	0.333	0.5	1746.571	0	1	4	0

Table 18 : (Continued)

A11	A12	A13	A14	A15	A16	A17	A19	Target
4	0.333	19	0	1	1	0	0	1
2	0	19	0	1	1	0	0	1
4	0.333	41	0	0	1	1	0	1
4	0.333	19	1	1	1	0	0	1
2	0.667	40	0	1	4	1	0	1

Table 19 : Examples from bad class after preprocessing

A1	A2	A3	A4	A5	A6	A7	A8	A9
0.5	9.86	0.33	1	723.534	0	0.5	4	0
0.5	20.4	0.33	0.167	9854.143	0	0.5	1	1
0	9.86	0.33	0.5	1280.395	0	0.5	3	1
0	39.05	0.33	0.167	9854.143	0	1	4	0
0	20.4	0.33	0.333	3491.284	0	1	4	0

Table 20 : (Continued)

A11	A12	A13	A14	A15	A16	A17	A19	Target
2	0.667	19	0	0	1	0	1	0
4	0.667	19	0	1	4	1	1	0
4	0.333	19	0	1	1	0	1	0
4	1	31	1	1	4	0	0	0
2	0.333	31	0	1	1	1	0	0

If data is originated from multiple sources, data integration is needed to combine such data into a coherent data store. And since the dataset used in this study comes from the same source, UCI Machine Learning, it does not require integration.

The next step is data transformation. In this step, the data is transformed or consolidated into forms, which is appropriate for mining processes. As transformation runs, we apply min-max normalization to the examples so that their value can propagate into mining process. Min-max normalization is well explained by this formula:

$$v' = \frac{v - \min}{\max - \min}(new\_\max - new\_\min) + new\_\min$$

After transformation, the value of each attribute lays between 0 and 1. This is common to neural network, although there are other set of values between -1 and 1.

C. Results

Finally, we come to results as show in the following tables. The results are given as five examples from upper line dataset for Good and Bad class.

Table 21 : Examples from good class

A1	A2	A3	A4	A5	A6	A7	A8	A9
1	1	0.333	0.5	1	1	1	0.333333	0
1	0.361083	0.333	0.5	0.060988	1	0.5	1	1
0.5	0.361083	0.333	0.333	0.510199	1	0.5	0	1
0	0	0.333	0.667	0.060988	1	0.5	1	1
1	0	0.333	0.5	0.112045	0	1	1	0

Table 22 : (Continued)

A11	A12	A13	A14	A15	A16	A17	A19	Target
1	0.333	0	0	1	0	0	0	1
0.333333	0	0	0	1	0	0	0	1
1	0.333	0.392857	0	0	0	1	0	1
1	0.333	0	1	1	0	0	0	1
0.333333	0.667	0.375	0	1	1	1	0	1

Table 23 : Examples from bad class after transformation

A1	A2	A3	A4	A5	A6	A7	A8	A9
0.5	0	0.33	1	0	0	0.5	1	0
0.5	0.361083	0.33	0.167	1	0	0.5	0	1
0	0	0.33	0.5	0.060988	0	0.5	0.666667	1
0	1	0.33	0.167	1	0	1	1	0
0	0.361083	0.33	0.333	0.303129	0	1	1	0

Table 24 : (Continued)

A11	A12	A13	A14	A15	A16	A17	A19	Target
0.333333	0.667	0	0	0	0	0	1	0
1	0.667	0	0	1	1	1	1	0
1	0.333	0	0	1	0	0	1	0
1	1	0.214286	1	1	1	0	0	0
0.333333	0.333	0.214286	0	1	0	1	0	0

D. Conclusion

This paper has explained basic data preparation, which can be used to identify the good data and the bad one. The dataset used in this study was German Credit Classification. As summarized in previous sections, some outliers were found. However, these outliers can be treated either by deleting the objects present in only one attribute, or by deleting the attributes.

Therefore, as can be seen from results, three attributes have been deleted. We assume that those attributes are less significant to contribute for loan sanction. These attributes are other debtor/guarantors, the number of people being liable to provide maintenance for, and foreign worker. Data preprocessing also showed that there are relationship or correlation between attributes, as exemplified by the correlation between the duration of loan and the credit amount. Its correlation can be deduced from its correlation coefficient value which is near to 1.

From result tables capturing a few of data after preprocessing, values for its attributes are laid between 0 and 1, which is a very small number. We can say that data is smoother and therefore suggest that it has suitable patterns for data mining process based on neural network.

Although some customers have bad credit history or risky credit experience, which is categorized as critical or delay on attribute credit-history, it is found that it is still possible for them to obtain the credit, as they can be included into the good class.

Finally, the dataset that used in this paper may help in the decision-making about customer's loan request. Therefore, experts involved on credits, loan, bank should be highly qualified so that they would be able to understand the data well.

REFERENCES

Afifi, A.A., & Azen, S.P. 1979. Statistical Analysis: A Computer Oriented Approach, Second Edition, Academic Press.

Borysowich, Craig. 2007. Preparing Data for Data Mining. Accessed from http://it.toolbox.com/blogs/enterprise-solutions/preparing-data-for-data-mining-17755. 6th of February 2009.

Han, Jiawei & Kamber, Micheline. Data Mining: Concepts and Techniques. Accessed from http://www.cs.sfu.ca/~han/bk/a2dbminer.ppt. 6th of February 2009.

Hofmann, H. 2007. UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

Kannan, Kavitha. Data Mining Report on Iris and Australian Credit Card Dataset. Tugas Ilmiah. School of Computer Science and Information Technology. Universiti Putra Malaysia.

Kulkarni, R.V. & Sreekantha, D.K. 2008. Knowledgebase System Design For Credit Risk Evaluation Using Neuro - Fuzzy Logic. *Proc. Of NCKM.*

Lukawiecki, Rafal. 2007. Introduction to Data Mining. Accessed from http://download.microsoft.com/documents/uk/technet/postevent/04-04-2008/2_Introduction_to_Data_Mining.pptx. 6th of February 2009.

Mihail, N., Cetină, I., and Orzan, G. Credit Risk Evaluation. Theoretical and Applied Economics.

Negnevitsky, Michael. 2002. Artificial Intelligence: A Guide to Intelligent Systems, First Edition, Addison-Wesley.

Sukumar, Rajagopal. Data Mining. Accessed from http://sirius-software.com/sug99/datmin.ppt. 6th of February 2009.

Zhang, S., Zhang, C.Q., and Yang, Q. 2003. Data Preparation For Data Mining. Applied Artificial Intelligence, 17:375–381.