Ramya.U *et al*, International Journal of Computer Science and Mobile Applications,
Vol.6 Issue. 12, December- 2018, pg. 1-7 **ISSN: 2321-8363**
Impact Factor: 5.515

Widespread Manner of Scrutiny on Data Pre-Processing Techniques

Ramya.U[1], Hema Priya.K.E[2], Lakshmi Priya.P[3], Tharani.K[4]
[1]Assistant Professor, Department of CSA, Sri Krishna Arts and Science College, ramyau@skasc.ac.in
[2]Assistant Professor, Department of CSA, Sri Krishna Arts and Science College, hemapriyake@skasc.ac.in
[3]Department of CSA, Sri Krishna Arts and Science College, lakshmipriyap16bcc125@skasc.ac.in
[4]Department of CSA, Sri Krishna Arts and Science College, Tharanik16bcc154@skasc.ac.in

ABSTRACT: Data pre-processing technique is one important method in data mining. But in middle-of-the-road this process is over and over again not painstaking as an important aspect in data mining process. Only if a dataset is cleaned then the supplementary step can be done. This paper deals with the analysis on data pre-processing technique.

KEYWORDS: Data pre-processing, Data cleaning, Data Integration, Data transformation, Data reduction.

1. INTRODUCTION:

Data pre-processing is one of the important aspect in data mining process. While gathering the data from real world there may be a lot of attributes missing, incomplete and inconsistent values. This leads to ambiguous results.. So data pre-processing techniques are applied to the gathered data and it is moved for the further data mining techniques.

2. DATA PRE-PROCESSING METHODS:

Unprocessed data may highly inclined to noise, missing values, and inconsistent data. So the unprocessed data should undergo some of the pre-processing techniques to improve the quality of the data .But data-Pre-processing is considered one of the most difficult method in data mining techniques .Because it deals with the grounding and transformation of dataset. Some of the data pre -processing techniques are listed below.

- Data cleaning
- Data integration
- Data transformation
- Data reduction

These techniques are applied to improve the quality of the dataset and then it is taken for data mining techniques.

3. DATA CLEANING

Data mining is a process of gathering the knowledge from the large dataset. It is actually a process of refining so while doing data mining there may be a lot of inconsistent, noisy and incomplete data which contains errors and some missing values. Missing values often occur due to loss of updates and the data which is not considered as that much important are not entered .In the dataset the inconsistent data occurs due to deleted entries. And modifications of the data have not been updated. The tools used for data collection may be faulty. The errors in data transmission also may occur. There are many methods used in the data cleaning process to find the missing values, noisy data and inconsistent data.

3.1. Binning methods:

Binning methods are used to smooth the sorted data by consulting its environs .Then the arranged values are disseminated into amount of bins. Binnig methods perform local smoothing technique by consulting its neighbour.

3.2. Clustering:

Clustering is a form of grouping the similar objects. By grouping the similar objects the objects that are not similar(outliers) can be easily identified.

For instance diabetes dataset is taken and it is being pre-processed to find the patients ranging from their age get affected to diabetes. So that it will result in the appropriate age of getting affected to diabetes. A simple K-means clustering is used to group the patients based on their age value.

Final cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(768.0)	(500.0)	(268.0)
preg	3.8451	3.298	4.8657
plas	120.8945	109.98	141.2575
pres	69.1055	68.184	70.8246
skin	20.5365	19.664	22.1642
insu	79.7995	68.792	100.3358
mass	31.9926	30.3042	35.1425
pedi	0.4719	0.4297	0.5505
age	33.2409	31.19	37.0672
class	tested_negative	tested_negative	tested_positive

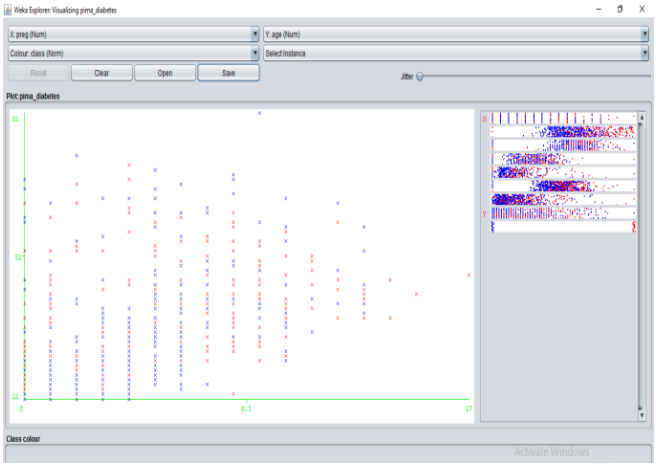


Figure 3.2.1

3.3. Computer-Human assessment:

Outliers can also be identified by combining the human and computer assessment methods.

3.4 Regression

Another form of cleaning method is regression .Data can be smoothed by fitting the data to a function. Linear regression is one form of method of regression it indulges in finding the best line to fit two variables.

4. DATA INTEGRATION:

Data integration is actually a process of analysis task It is used to combine the data from multiple sources into a lucid data store .Issues that occur in data integration is that the data that is gathered from multiple sources are not likely to be the same that is the data won't be matched up, this type of problem is called as entity identification problem. Another important issue is data redundancy .Data redundancy can be occurred due to inconsistent data, missing data and noisy data.

5. DATA TRANSFORMATION:

Data transformation is a process of converting or consolidating the data into the forms that is apposite for mining .Some of the data transformation techniques are

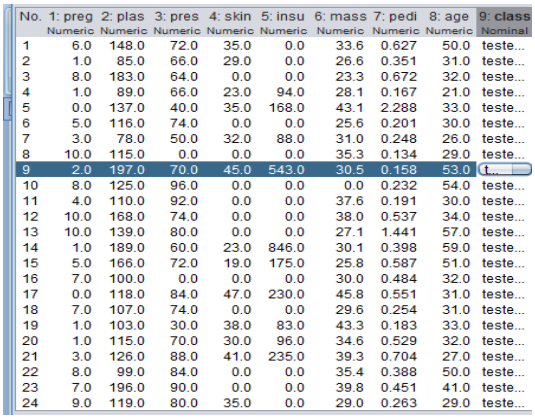
- Normalization
- Smoothing
- Aggregation
- Generalization of data

5.1 Normalization

Normalization is a process of reducing the data redundancy .Scaling of values to fall within a specified range, such as -1.0 to 1.0 or 0 to 1.0.For instance normalization done to diabetes dataset.

 ISSN: 2321-8363
Impact Factor: 5.515

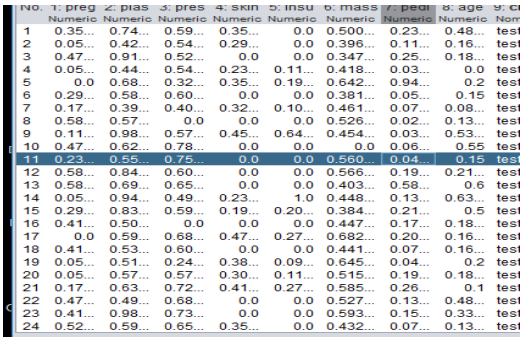
Before Normalization



No.	1: preg	2: plas	3: pres	4: skin	5: insu	6: mass	7: pedi	8: age	9: class
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	teste...
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	teste...
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	teste...
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	teste...
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	teste...
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	teste...
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	teste...
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	teste...
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	t...
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	teste...
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	teste...
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	teste...
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	teste...
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	teste...
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	teste...
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	teste...
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	teste...
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	teste...
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	teste...
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	teste...
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	teste...
22	8.0	99.0	84.0	0.0	0.0	35.4	0.388	50.0	teste...
23	7.0	196.0	90.0	0.0	0.0	39.8	0.451	41.0	teste...
24	9.0	119.0	80.0	35.0	0.0	29.0	0.263	29.0	teste...

Figure 5.1.1.

After Normalization



No.	1: preg	2: plas	3: pres	4: skin	5: insu	6: mass	7: pedi	8: age	9: Cl
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nom
1	0.35...	0.74...	0.59...	0.35...	0.0	0.500...	0.23...	0.48...	test
2	0.05...	0.42...	0.54...	0.29...	0.0	0.396...	0.11...	0.16...	test
3	0.47...	0.91...	0.52...	0.0	0.0	0.347...	0.25...	0.18...	test
4	0.05...	0.44...	0.54...	0.23...	0.11...	0.418...	0.03...	0.0	test
5	0.0	0.68...	0.32...	0.35...	0.19...	0.642...	0.94...	0.2	test
6	0.29...	0.58...	0.60...	0.0	0.0	0.381...	0.05...	0.15	test
7	0.17...	0.39...	0.40...	0.32...	0.10...	0.461...	0.07...	0.08...	test
8	0.58...	0.57...	0.0	0.0	0.0	0.526...	0.02...	0.13...	test
9	0.11...	0.98...	0.57...	0.45...	0.64...	0.454...	0.03...	0.53...	test
10	0.47...	0.62...	0.78...	0.0	0.0	0.0	0.06...	0.55	test
11	0.23...	0.55...	0.75...	0.0	0.0	0.560...	0.04...	0.15	test
12	0.58...	0.84...	0.60...	0.0	0.0	0.566...	0.19...	0.21...	test
13	0.58...	0.69...	0.65...	0.0	0.0	0.403...	0.58...	0.6	test
14	0.05...	0.94...	0.49...	0.23...	1.0	0.448...	0.13...	0.63...	test
15	0.29...	0.83...	0.59...	0.19...	0.20...	0.384...	0.21...	0.5	test
16	0.41...	0.50...	0.0	0.0	0.0	0.447...	0.17...	0.18...	test
17	0.0	0.59...	0.68...	0.47...	0.27...	0.682...	0.20...	0.16...	test
18	0.41...	0.53...	0.60...	0.0	0.0	0.441...	0.07...	0.16...	test
19	0.05...	0.51...	0.24...	0.38...	0.09...	0.645...	0.04...	0.2	test
20	0.05...	0.57...	0.57...	0.30...	0.11...	0.515...	0.19...	0.18...	test
21	0.17...	0.63...	0.72...	0.41...	0.27...	0.585...	0.26...	0.1	test
22	0.47...	0.49...	0.68...	0.0	0.0	0.527...	0.13...	0.48...	test
23	0.41...	0.98...	0.73...	0.0	0.0	0.593...	0.15...	0.33...	test
24	0.52...	0.59...	0.65...	0.35...	0.0	0.432...	0.07...	0.13...	test

Figure 5.1.2.

Before normalization the values ranges from 1 to 250 .But after normalization the values are scaled to fall within 0 to 1 measure.

5.2 Smoothing

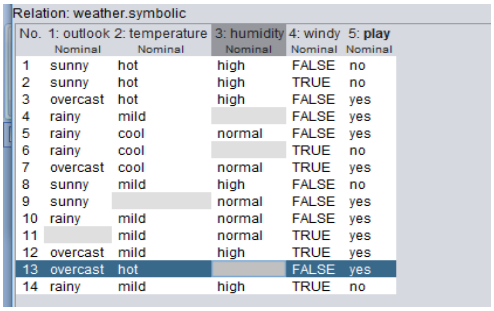
Smoothing is a technique used to remove the noisy data . Smoothing technique includes

- Binning
- Clustering
- Regression

Mostly in many dataset there are loss of updates that is there may many missing values . Before data mining all the missing values should be replaced. If the missing values are not found it will result in inappropriate results.

For example in weather nominal dataset there are many missing values .Before mining the missing values should be replaced.

Before Replacing Missing values

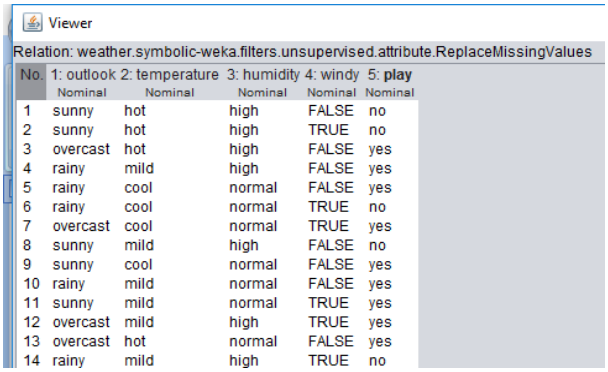


Relation: weather.symbolic

No.	1: outlook Nominal	2: temperature Nominal	3: humidity Nominal	4: windy Nominal	5: play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild		FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool		TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny		normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11		mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot		FALSE	yes
14	rainy	mild	high	TRUE	no

Figure 5.2.1

After replacing missing values



Viewer

Relation: weather.symbolic-weka.filters.unsupervised.attribute.ReplaceMissingValues

No.	1: outlook Nominal	2: temperature Nominal	3: humidity Nominal	4: windy Nominal	5: play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Figure 5.2.2

5.3 Aggregation:

Overall the dataset is aggregated or summarized by applying aggregation operations on the dataset.

5.4 Generalization of data:

In this type of transformation method the lower level of data is being replaced by higher level of concepts with the help of concept hierarchies.

6. DATA REDUCTION:

Data reduction techniques are helpful in representing the dataset in a reduced form because analyzing the huge amount of data .It is a process of either reducing the volume of the dataset or dimensions of the data set. There are n number of methods that smooth the progress of the breakdown on reducing the volumes and dimensions of the dataset which gives a high knowledge of data with appropriate results. The strategies of data reduction are

- Data cube aggregation
- Dimension reduction
- Data compression

- Numerosity reduction
- Discretization and concept of hierarchy generation

6.1 Data cube aggregation

Aggregation operations are being applied to the data and the data cubes are built

6.2 Dimension Reduction

Dimension reduction is the concept of reducing the irrelevant or redundant attributes.

6.3 Data compression

Data compression is a technique of reducing the size of the data set that can be used for principle component analysis

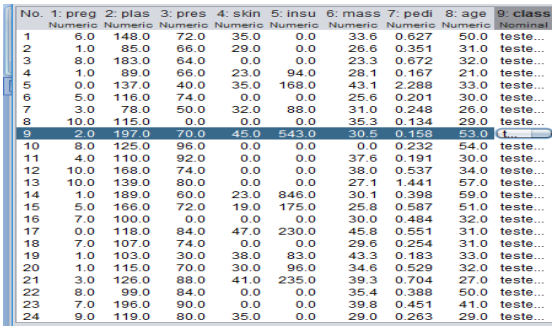
6.4 Numerosity reduction

Numerosity reduction is the concept of replacing the data with smaller data representations

6.5 Discretization and concept of hierarchy generation

Raw data values are replaced by higher conceptual levels.

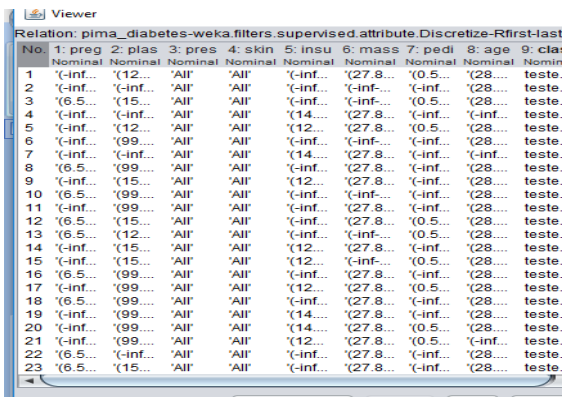
Before Discretization



No.	1: preg	2: plas	3: pres	4: skin	5: insu	6: mass	7: pedi	8: age	9: class
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	teste...
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	teste...
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	teste...
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	teste...
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	teste...
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	teste...
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	teste...
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	teste...
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	1
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	teste...
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	teste...
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	teste...
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	teste...
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	teste...
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	teste...
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	teste...
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	teste...
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	teste...
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	teste...
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	teste...
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	teste...
22	8.0	99.0	84.0	0.0	0.0	35.4	0.388	50.0	teste...
23	7.0	196.0	90.0	0.0	0.0	39.8	0.451	41.0	teste...
24	9.0	119.0	80.0	35.0	0.0	29.0	0.263	29.0	teste...

Figure 5.4

After Discretization



Viewer

Relation: pima_diabetes-weka.filters.supervised.attribute.Discretize-Rfirst-last

No.	1: preg	2: plas	3: pres	4: skin	5: insu	6: mass	7: pedi	8: age	9: cla
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nomin
1	'(-inf...	'(12...	'All'	'All'	'(-inf...	'(27.8...	'(0.5...	'(28....	teste.
2	'(-inf...	'(-inf...	'All'	'All'	'(-inf...	'(-inf-...	'(-inf...	'(28....	teste.
3	'(6.5...	'(15...	'All'	'All'	'(-inf...	'(-inf-...	'(0.5...	'(28....	teste.
4	'(-inf...	'(-inf...	'All'	'All'	'(14....	'(27.8...	'(-inf...	'(-inf...	teste.
5	'(-inf...	'(12...	'All'	'All'	'(12...	'(27.8...	'(0.5...	'(28....	teste.
6	'(-inf...	'(99....	'All'	'All'	'(-inf...	'(-inf-...	'(-inf...	'(28....	teste.
7	'(-inf...	'(-inf...	'All'	'All'	'(14....	'(27.8...	'(-inf...	'(-inf...	teste.
8	'(6.5...	'(99....	'All'	'All'	'(-inf...	'(27.8...	'(-inf...	'(28....	teste.
9	'(-inf...	'(15...	'All'	'All'	'(12...	'(27.8...	'(-inf...	'(28....	teste.
10	'(6.5...	'(99....	'All'	'All'	'(-inf...	'(-inf-...	'(-inf...	'(28....	teste.
11	'(-inf...	'(99....	'All'	'All'	'(-inf...	'(27.8...	'(-inf...	'(28....	teste.
12	'(6.5...	'(15...	'All'	'All'	'(-inf...	'(27.8...	'(0.5...	'(28....	teste.
13	'(6.5...	'(12...	'All'	'All'	'(-inf...	'(-inf-...	'(0.5...	'(28....	teste.
14	'(-inf...	'(15...	'All'	'All'	'(12...	'(27.8...	'(-inf...	'(28....	teste.
15	'(-inf...	'(15...	'All'	'All'	'(12...	'(-inf-...	'(0.5...	'(28....	teste.
16	'(6.5...	'(99....	'All'	'All'	'(-inf...	'(27.8...	'(-inf...	'(28....	teste.
17	'(-inf...	'(99....	'All'	'All'	'(12...	'(27.8...	'(0.5...	'(28....	teste.
18	'(6.5...	'(99....	'All'	'All'	'(-inf...	'(27.8...	'(-inf...	'(28....	teste.
19	'(-inf...	'(99....	'All'	'All'	'(14....	'(27.8...	'(-inf...	'(28....	teste.
20	'(-inf...	'(99....	'All'	'All'	'(14....	'(27.8...	'(0.5...	'(28....	teste.
21	'(-inf...	'(99....	'All'	'All'	'(12...	'(27.8...	'(0.5...	'(-inf...	teste.
22	'(6.5...	'(-inf...	'All'	'All'	'(-inf...	'(27.8...	'(-inf...	'(28....	teste.
23	'(6.5...	'(15...	'All'	'All'	'(-inf...	'(27.8...	'(-inf...	'(28....	teste.

Figure 5.5

7. CONCLUSION:

Thus data pre-processing improves the generalizability of the model. And it is to be considered as the one of the major factors in data mining .By doing data pre-processing mostly all the irrelevant, inconsistent and noisy data can be reduced and it reflects in the quality of the knowledge gathered.

REFERENCES:

[1] Winter school on "Data Mining techniques and tools for knowledge in agriculture data set".

[2] www.wikipedia.com

[3] www.cs.ccsu.edu

[4] www.techopedia.com

[5] www.mimuw,edu.pl

