# Chapter Publication

## Domain: Data Mining and Warehousing

## Subdomain: Data Preprocessing


Author By:

**Mrs. Nilam S. Patil**

**Assistant Professor,**

**D. Y. Patil College of Engineering Akurdi Pune,**

**Teaching Experience: 17+ Years**

**PhD Pursuing From Vel Tech University, Chennai**



All data related to this chapter will copyright@aspirepublishers.com
Publication Date: December 2018
Web: www.aspirepublishers.com
Email: info@aspirepublishers.com

# INDEX

# CHAPTER 1

# KNOWLEDGE DISCOVERY FROM DATA(KDD)

## 1.1 KNOWLEDGE DISCOVERY FROM DATA

Sequence of iterative step:-

1. Data cleaning (to remove noise and in-consistence data)
2. Data integration (multiple data source are combined)
3. Data selection (data relevant to analysis task are retrieve from the database)
4. Data transformation (data are transformed and consolidated into forms appropriate for mining by performing summery or aggregate)
5. Data mining (essential method are applied to extract data pattern)
6. Pattern evaluation(to identify truly interesting pattern representing knowledge based on intelligent method)
7. Knowledge representation (visualization and knowledge representation technology are used to present knowledge to user)

Data preprocessing where data is prepared for mining. The dm step intract user or knowledge base.
New interesting pattern are presented to the user and may br stored new knowledge in knowledge base.



## 1.2 DATA MINING

it is process of discovering interesting patten and knowledge from large amount of data the data source include database ,data,warehouse,the web other system data or data stream into system dynamically

### 1.2.1 Data Mining Task Primitive

data mining task can be specified in the form of a data query, which is i/p to dm system dm query is defined in terms of dm task primitive these primitive allows the user to interactively communicate dm system during discovering in order to direct the mining process or examine the finding from different angle or primitive


1) The set of task relevant data to mined :-

-it include the potion the database or set of data in which user is interested. it include db attributes or data ware house dimension.

2) The kind of knowledge to be mined:-

-it include dm function to be formed as characterization ,discrimination ,association and correlation analysis ,classify, prediction, clustering, analysis, evolution analysis.

3) The background knowledge to be used in the discovery process:-

knowledge about for domain to be is useful for guiding kd process and for evaluating the patterns formed.

Concept Hierarchy:-

they are popular from of background knowledge ,which allows data to be mined of multiple levels of abstraction.

4) Interesting measures and should for pattern evaluation they may be used to guide the mining process or offer discovering to evaluate the discovered pattern .

Different kinds of knowledge has different discovery patterns.

interesting measure for association mining are support and confidence rules where suppot and confidence as below user specified considered .

5) Expected representation for visualizing the discovered pattern :-

it refer to the forms in which many discovered knowledge is to be displayed ,which may include rules, table, charts, graph, decision, tree, cubes etc.

Data :Information and Knowledge



### 1.2.2 Data

Data is raw, unrecognised facts that need to be processed .its is simple ,random unless its organized.

ex: student test score

Information :- When data is processed organized structures or presented in given contract to make its useful its allows information

ex:-average score a class

Attributes types :- binary ,nominal ,ordinal and numeric attributes discrete vs continues attributes

- data set are made up of data object

-data object represent an entity - object db customer ,store,sales ,patient ,university

-data object are described by attributes

-stored in db ,they are called tuples row of data

### 1.2.3 Attributes

its is a data field representing character or features of a data object

attributes = dimension = features = variable

data ware housing - dimension

machine learning = features

statisticians- variable

data mining and database professional - attributes



### 1.2.4 Attributes vector

A set of attributes used to describe a given object is called a attribute vector (features vector)

Unvariate : distribution of data involving are attributes

Bivariate : distribution of data involving 2 attributes

Types of an attributes is detemined by the set of possible values - nominal, binary, numeric

### 1.2.5 Nominal attributes

- values of a nominal attributes are symbol ,or names of thing

-each values represent same kind of category code or state and so nominal attributes

are also referred to as categorical

-the values do not have any meaningful order

-in computer system ,values ,are called enumeration

ex:-

| object | Attributes | value |
|---|---|---|
| person | hair color | black, brown, blond |
| person | maritial status | single, married |

Nominal attributes represent by symbol /number/black /brown

- they do not have any meaningful order not quantitative so we cant final mean median mode(commonly occurring value)



### 1.2.6 binary attributes

only 2 values as 1 or 0

1:present

0:absent

boolean values: true,false

Ex:-

medical test is attributes

values as test is positive and negative

**symmetric:** both values are equally important

no preference gender:male/female

**asymmetric:**

Both values are not equally important

Positive

Negative

### 1.2.7 ordinal

Attributes having meaningful order or among them ,but magnitude between successive values is not known.

Ex:- customer satisfaction

1 vary dissatisfied

2 natural

3 satisfied

4 vary, nominal, binary, attributes

### 1.2.8 Numeric attributes

it is quantitative ,measurable quantity represented in integer / real values. They can be interval scaled or ratio scaled.

**Interval scaled attributes**

- They are measured an a scale of equals size units

- The values of interval scaled attributes hares order can be negative and positive

- Providing ranking of values such attributes allow us to compare and qualify the difference between value.

Ex: Temperature attribute

1) Each day outdoor temperature is object rendering values 20c is 5c higher than 15c

2) calender dates: year 2002 eight year higher than 2010

- There is no true zero point for temperature and calendar dates. year 0 doesn't mean beginning if time . Internal scaled attributes are numeric. we can compute mean, median, mode.

**Ratio scaled** Numeric attribute with an inherent zero point if measurement is ratio scaled, a value can be multiple(ratio) of other values values are ordered We can compute, mean, mode, median, difference.

Ex. Temperature in kelvin

0 k = -273.15c



### 1.2.9 Discrete vs Continuous

**Discrete :**

Finite, countably finite set of value which may or may not be represented as integer.

Ex. Attributes with numeric value as binary

An attributes can be countably infinite but values can be put in a one to one correspondence of natural numbers

Ex. Customer_ID : countably infinite

Number of customers may be infinite but actual set of values are countable.

**Continuous :**

Attribute which is not discrete is continuous.

numeric attribute = continuous attribute.

-continuous values are real numbers ,floater pant variable.

-Number Value are real integers or real value.

## 1.3 DATA TRANSFORMATION

Min max normalization

Z-Score Normalization

Decimal scaling

Data are transformed or consolidated so that resultry mining process may be 1) more efficient, 2)practices found may be easier to understand.



### 1.3.1 Strategies

#### 1.3.1.1 Smoothing :

Remove noise from the data.

Techniques : binary, regression, cluster.

#### 1.3.1.2 Attribute construction (feature construction )

new Attributes are constructed and added from the given set of attribute to help mining process.

#### 1.3.1.3 Aggregation

Summary or aggregation operations are applied to data. Ex. Daily sales data may be aggregated so as to compute monthly and annual total amount. This is used in

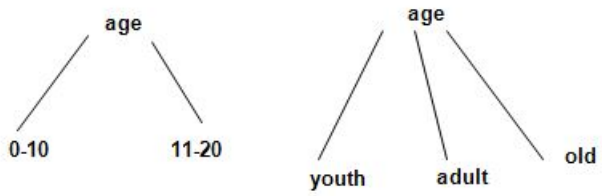
constructing data cube for data analysis at multiple abstraction levels.

1.3.1.4 Normalization :

The attribute data are scaled so as to fall within a smaller range such as -1.0 or 0.0 to 1.0

1.3.1.5 Discretization :

Raw value of a numeric attribute (age) are replaces by interval values (0-10,11-20 etc ) or conceptual labels (youth, adult, senior) The Labels are can be recursively organized into higher level concepts, resulting concept hierarchical, the numeric attribute.




1.3.1.6 Concept hierarchy generation for nominal data :

Ex. street can be generalized to higher levels concepts like city or country. Many hierarchies for nominal attributes are complicit within the database schema and can be automatically defined at the schema definition level.



## 1.4 MIN MAX NORMALIZATION

Measurement unit can affect data analysis change measurement from meter to height ,kg to pounds weight may generate different results.

- Normalizing the data attempts to give all attributes an equal weight.
- Its useful for classification algorithms like nearest neighbourhood classification/clustering.

**Neural NN back propagation algorithm for classification**

- Normalizing the input values for each attributes measurement in training tuples will help to speed up learning phase.

- Distance based metric normalization help to present the attributes with initially large from outweighing with initial smaller range (binary)
- Its also helpful if no prior knowledge of data.

### 1.4.1 Normalization Technique

#### 1.4.1.1 Min max normalization

Minmax normalization is a normalization strategy which linearly transforms x to y= (x-min)/(max-min), where min and max are the minimum and maximum values in X, where X is the set of observed values of x.

It can be easily seen that when x=min, then y=0, and

When x=max, then y=1.

This means, the minimum value in X is mapped to 0 and the maximum value in X is mapped to 1. So, the entire range of values of X from min to max are mapped to the range 0 to 1.

$$v' = \frac{v - min_F}{max_F - min_F}(new\_max_F - new\_min_F) + new\_min_F$$



#### 1.4.1.2 Z-score normalization

values for an attribute A, are normalized based on mean (i.e average) and standard deviation of A, A value, Vi of A is normalizes to Vi' by A measure of an observations distance from the mean.

The distance is measured in standard deviation units.

If a z-score is zero, its on the mean.

If a z-score is positive, its above the mean.

If a z-score is negative, its below the mean.

If a z-score is 1, its 1 SD above the mean.

If a z-score is 2, its 2 SDs below the mean.

$$z=\frac{X-\mu}{\sigma} \text{ or } z=\frac{X-\overline{X}}{SD}$$

1.4.1.3 Normalization by decimal scaling

This is generally used in data mining, but is one of the techniques used wherever there is a need to normalize data from disparate sources.

When you have a range of numbers like 50, 250, 400, you can do this:

Take the maximum number of digits. Here it is 3 (400 has 3 digits)

Calculate power of 10. $10^3 = 1000$.

Divide each number by 1000.

The results would be 0.05, 0.25 and 0.4.

1.4.1.4 Discretization by binning

These methods can also be used for data reduction and concept hierarchy generalization. -Binning does not used and class info so unsuspended Discretization.

Binning : sorted data by consulting its neighbourhood

# REFERENCES

Reference: https://docs.rapidminer.com/latest/studio/operators/cleansing/binning/discretize by bins.html
https://t4tutorials.com/min-max-normalization-of-data-in-data-mining/
https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/