UNIT 2

Data Preprocessing

**Que 1: what is data preprocessing?**

It is a process of converting the data into appropriate and clear form for data mining. The input data should be of good quality, because the quality of knowledge provided by data mining is proportional to the input data.

**Que 2: What do you mean by aggregation? What are the advantages of aggregation?**

Ans: Combining of 2 or more objects is known as aggregation. The large data set Quantitative attributes may be aggregated by either taking sum or mean of attribute values.

Advantages of aggregation

* Aggregation makes data smaller, smaller data sets require less processing time.
* It can provide higher view of data i.e. Data can be viewed at higher level of abstraction (summarized data can be viewed).
* Attribute at higher level of abstraction have less variability than attribute at lower level of abstraction.

Consider an example:

Below example records the sales in 3 branches of a supermarket (day wise report).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TID** | **Item** | **Store location** | **Date** | **price** |
| 1001 | Colgate tooth brush | Bhimavaram | 1/3/2013 | 30 |
| 1002 | Amul Butter | Tanuku | 2/4/2013 | 20 |
| 1003 | hp Computer mouse | Bhimavaram | 2/4/2013 | 200 |
| 1004 | Pepsodent toothpaste | Tanuku | 3/4/2013 | 50 |
| 1005 | Night lamp | Tanuku | 3/3/2013 | 100 |
| 1006 | Thums up | Bhimavaram | 4/4/2013 | 65 |
|  | Fig:1 |  |  |  |
|  |  |  |  |  |
|  | Location | Month | Sales |  |
|  | Bhimavaram | March | 98.3 |  |
|  | Tanuku | April | 56.6 |  |

Fig :2

Fig1 represents the original data set and fig 2 represents the aggregated data set. TID and Items are removed and daily sales are converted into monthly sales.

**Que 3: What is sampling technique? What are the different techniques for sampling?**

Ans: This technique is inherited by data mining from statistics.

It is very expensive and time consuming to process all the data. Sampling technique can be used to consider only a subset of data inserted of whole data for analysis.

A Sample is said to be repetitive if it has the same properties as original data.

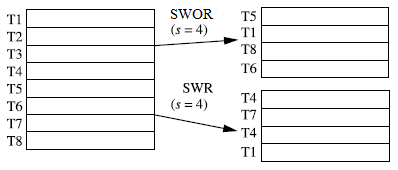
**Sampling Approaches**

There are mainly three techniques for sampling:

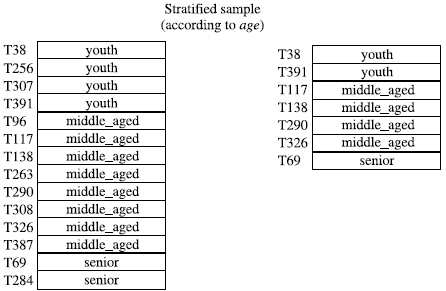
* Sampling without replacement (SWOR)
* Sampling with replacement(SWR)
* Stratified sampling

**Sampling without replacement (SWOR):** An item once selected from the data set, it is removed from the original population (data set).Consider the below example: No item is sampled more than once.

**Sampling with replacement (SWR):** An item once selected from the data set, it is again kept in the same place constituting original population. Means, a same item may be sampled more than once. Consider the example below. Item T4 is sampled 2 times.



**Stratified sampling:** The number of objects drawn from each group is proportional to the size of the group in original population. Consider the below example. The ratio of age groups in the original population is same as that of the samples. (4:8:2) is proportional to (2:4:1).



**Problems with sampling:**

* If the sample size is less then, some important patterns may be missed. And if, sample size is more then they eliminate the advantages of sampling (i.e. less time consuming and less storage space)

**Que 4: Write a short note on progressive sample.**

Progressive sample method is used to determine the sufficient size of the sample. This approach starts with a small sample, and then increase the sample size until a sample of sufficient size has been obtained.

The correct sample size can be found in the following way:

The accuracy of predictive model increases with respect to the size of sample. At a point the accuracy doesn’t increase. This point is known as leveling off point. Another sample is considered from the same original data and increase the small sample to the same size. Now, the closeness of this with the leveling point is measured.

**Que5: What do you mean by dimensionality reduction? What are the advantages of it? Explain the curse of dimensionality?**

Ans: Dimensionality reduction refers to the creation of new attributes that are combination of the old attributes.

Advantages of dimensionality reduction

1. Dimensionality reduction eliminates irrelevant attributes and reduces noise in the data
2. Many Data mining algorithm work better with data having less number of dimensions (attributes).
3. Reduction of dimensions leads to a more understandable data model.
4. Reduction of dimensions allows data to be visualized easily.
5. Time and memory required by the data mining algorithm is reduced with reduction in dimensionality.

Curse of dimensionality (disadvantages of having more dimensions in data)

1. Data analysis becomes difficult as the dimensionality of data increases.
2. If data is having more number of dimensions, it is very difficult to create a classification model due to fewer objects.
3. In the case of clustering, the density and distance between the objects would be more. This makes clustering difficult.

**Que6 How to select the best set of features (attributes, dimensions).**

**(or)**

**Explain feature subset selection?**

Ans: In feature subset selection, only a subset of all dimensions is used. This is specially used when there are large numbers of redundant and irrelevant dimensions in the dataset.

Note:

**Redundant features**: Has same information (same values) in more than one dimension.

**Irrelevant features:** Hasdimensions irrelevant to data mining tasks.

**Techniques for eliminating irrelevant and redundant features:**

1. **Common sense**: Some irrelevant and redundant dimensions can be removed using common sense or having a sound command on domain.
2. **Brute-force approach**: Try all possible feature subsets as input to data mining algorithm and select the subset which produces best results. But this method will not work if numbers of attributes (features) are more.
3. **Embedded approaches**: Feature selection occurs naturally as part of the data mining algorithm. The algorithm will decide which features to include and which to ignore.
4. **Filter approaches:** attributes (features) are selected before data mining algorithm is run using some independent approaches.
5. **Wrapper** **approaches**: The data mining algorithm itself is used to determine the attribute subset.

**Que7: What do you mean by feature creation? What are the three methodologies for creating new attributes?**

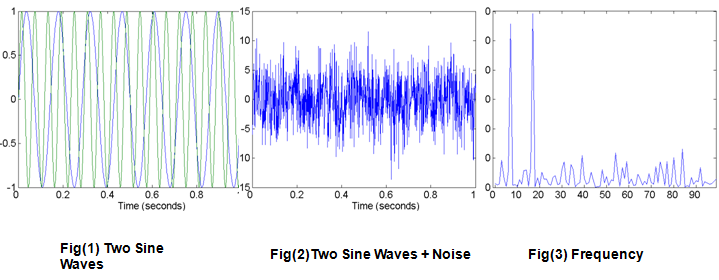
**Ans:** Feature creation is a process of creating new set of attributes that can capture information more efficiently than the original attributes.

The three methodologies for creating new attributes are:

1. Feature extraction.
2. Mapping the data to a new space
3. Feature construction.

**Feature extraction:** The creation of new attributes from original raw data is known as Feature extraction. This method is mainly used in image processing. Consider a set of images stored in the form of pixels, we want to classify them as containing human faces or not. If new attributes are created containing information about certain edges and colors, then this attributes helps us to better classify these images.

**Mapping the data to a new space:** It is the process of viewing the data in different angles to reveal important and interesting features in it. This method is mainly used in Time series data. Consider the below example. Fig 1 contains two time series data with out noise. And fig (2) contains two time series data with noise. By Appling Fourier transformation, time series data in fig (2) is converted into frequency information presented in fig (3).

****

**Feature construction:** Sometimes the attributes in the original data sets have the necessary information, but still these attributes are not suitable for data mining algorithm. In this situation, New attributes are constructed from the original attributes which as more suitable for data mining algorithm.

Example: Consider an original data set containing mass and volume information of various metals. In this case it is more meaningful to construct density information (density =mass/volume) rather than mass and volume.

|  |  |  |
| --- | --- | --- |
| Metal | Mass | volume |
| Metal A | 16.23 | 15.68 |
| Metal B | 17.89 | 14.34 |
| Metal C | 18.33 | 13.67 |

Fig A

|  |  |
| --- | --- |
| Metal | Density |
| Metal A | 1.0350 |
| Metal B | 1.2475 |
| Metal C | 1.3408 |

Fig B

**Que 8: Explain Discretization and Binarization? What are the different types of discretization?**

**Ans:**

Binarization: The process of converting continuous and discrete attributes into binary attributes is known as binarization.

The two techniques for binarization:

1) If there are m categorical values, then uniquely assign each ordinal value to an integer (0,m-1). Then convert these integers into binary numbers

In the below example there are 5 categorical values (awful, poor, ok, good, great), the uniquely assign each value to an integer (awful=0, poor=1, ok=2, good=3, great=4). Then convert these integers into binary numbers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Categorical value | Integer value | X1 | X2 | X3 |
| Awful | 0 | 0 | 0 | 0 |
| Poor | 1 | 0 | 0 | 1 |
| OK | 2 | 0 | 1 | 0 |
| Good | 3 | 0 | 1 | 1 |
| Great | 4 | 1 | 0 | 0 |

Note: If the categorical values are ordinal then values should be stored in a sequence.

2) In this technique only the presence of item is considered. Here the number of binary attributes is equal to the number of categorical values.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Categorical value | Integer value | X1 | X2 | X3 | X4 | X5 |
| Awful | 0 | 1 | 0 | 0 | 0 | 0 |
| Poor | 1 | 0 | 1 | 0 | 0 | 0 |
| OK | 2 | 0 | 0 | 1 | 0 | 0 |
| Good | 3 | 0 | 0 | 0 | 1 | 0 |
| Great | 4 | 0 | 0 | 0 | 0 | 1 |

Discretization:

Process of converting continuous attributes into categorical attributes is known as Discretization. This technique is used for the data used for classification and association analysis. Discretization involves two subtasks:

1. How many categorical values to include.
2. How to map continuous attributes to categorical attributes.

For example consider a student table sorted in percentages order

|  |  |  |  |
| --- | --- | --- | --- |
| Student name | Percentages |  | grade |
| A | 45.8 |  | Second class |
| B | 47.9 |
| C | 66.7 |  | First class |
| D | 65.6 |
| E | 62.5 |
| F | 77.8 |  | destination |
| G | 80.6 |

There are two types of discretization 1) Unsupervised discretization

2) Supervised discretization

**Unsupervised discretization**: In this type of discretization, domain knowledge (class information) is not used to convert continuous attributes into categorical attributes. Rather, they are discretized using techniques like 1) Equal width 2) Equal depth 3) K means etc...

**Equal width**: It divides the attribute values into equal intervals. Consider a set of attribute values(2,3,4,9,8,15,16,23,26,28,21,22)

**Bin1(1-10)-2,3,4,9,8**

**Bin 2(11-20)-15, 16**

**Bin3(20-30)-23,26,28,21,22**

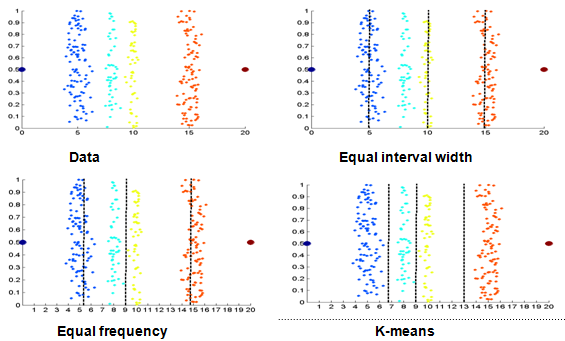
**Equal depth**: It divides the attribute values into equal parts. First assign them in an order(2,3,4,8,9,15,16,21,22,23,26,28)

**Bin1: 2,3,4,8**

**Bin 2: 9,15,16,21**

**Bin3: 22,23,26,28**

In the below example, dark dotes are outliers. K-means performs the best

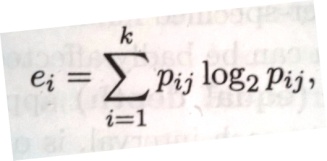
****

Supervised Discretization:

Discretization methods which use class labels are known as supervised discretization. Supervised discretization produces better results than unsupervised discretization.

Entropy measure is a common method of discretization.

Entropy = ei=

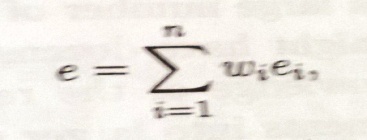


Where, k=number of class labels, and i=(1,2,3..)

mij= number of values of class j in ith interval

And, pij=mij/mi.

The total entropy is denoted by



**Variable Transformation or attribute transformation**

When transformation is performed on all values of a attribute, then it is known as variable transformation. This technique is mainly used when the attribute values are very large, or when the magnitude is more

Uses of variable transformation

1) It is used to transform larger values into smaller values

2) It is used to reduce the dependency of values on its units

3) It is also used when the magnitude of the values are important then the values.

**Simple functions**

In this, simple mathematical function is applied to all values of the variable (attribute) individually. For example consider a variable(x) it is converted by taking its or ex or 1/X or sin X or |x| or log X etc..

|  |  |
| --- | --- |
| Weight (x) | Transformed weight |
| 1000 | 31.66 |
| 100 | 10 |
| 350 | 18.70 |
| 245 | 15.65 |

This technique is mainly used when the value range is very large.

Note: In some cases the transformations changes the nature of the variable. For example the values {1, 2, 3} if transformed into {1/1, 1/2, 1/3} then the order changes i.e. (1<2<3) but (1/1> 1/2> 1/3).

**Normalization or standardization**

In this, formulas like mean and standard deviation are used. These methods are mainly used to reduce the dependency of values on its units. For example weight of people can be measured in kg or pounds. In order to consider weights irrespective of measures these techniques are helpful.

If the mean of attribute values and Sx standard deviation then, X|= (X-)/SX Creates a new variable.

The mean and standard deviation are affected by outliers. In this case median and absolute standard deviation can be used.

Absolute standard deviation=

Xi= value of the variable

= median

m= number of values of an attribute.