CHAPTER-5

DATA PREPROCESSING

5.1 Why preprocess data

5.2 Data cleaning:

5.3 Noisy data

5.4 Inconsistent Data:

5.5 Data Integration and Transformation

5.6 Data reduction:

5.7 Data Cube Aggregation:.

5.8 Dimensionally Reduction:

5.9 Data Compression

5.10 Review Questions

5.11 References

5. DATA PREPROCESSING

5.1 Why preprocess data?

Incomplete, inconsistent and noisy data are commonplace properties of large real-world databases. Attributes of interest may not always be available and other data was included just because it was considered to be important at the time of entity. Relevant data may not sometimes be recorded. Further more, the recording of the modifications to the data may not have been done. There are many possible reasons for noisy data (incorrect attribute values). They could have been human as well as computer errors that occurred during data entry. There could be inconsistent in the naming conventions adopted. Sometimes duplicate tuples may occur.

Data cleaning routines work to "clean" the data by filling in the missing values, smoothing noisy data, identifying and remo0ving outliers, and resolving inconsistencies in the data. Although mining routines have some form of handling noisy data, they are always not robust. If you would like to include files from many sources in your analysis then requires data integration. Naming inconsistencies may occur in this context. A large amount of redundant data may confuse or slow down the knowledge discovery process. In addition to data cleaning steps must taken to remove redundancies in the data.

Sometimes data would have to be normalized so that it scaled to a specific range e.g. [0.0, 1.0] in order to work data mining algorithms such as neural-networks, or clustering. Furthermore, you would require aggregating data e.g. as sales per region-something thet is not part of the data transformation methods need to be applied to the data.

Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same or almost the same analytical results. There are a number of strategies for data reduction-data compression, numerosity reduction, generalization; data reduction will be discussed in detail later.

5.2 Data cleaning:

Missing values: Sometimes you may find that many tuples do not value in them for some particular attributes. How do you fill in the missing values for these attributes?

1. **Ignore the tuples**: This is usually done when the label is missing. This method though is not very effective. It is especially poor when the percentage of missing values is high.
2. **Fill in the missing values manually**: In general this method is very time consuming and it may not be possible when the volume of the data set is very large.
3. **Use a global constant to fill in the missing values**: Replace all missing values by the same constant such as label like "unknown". Though this method is quite simple it may not be recommended in some cases since all those data of the same class label may be treated together as a class.
4. **Use the attributes mean to fill up the missing values**: For example suppose the average income is Rs.12, 000 uses this value to fill in the missing values for income.
5. **Use the most probable value to fill in the missing values**: This may be determined with the regression, inference based tools using Bayesian formalism or decision tree induction.

5.3 Noisy data:

Noise is the random error or variance in a measured variable. Given a numeric attribute such as say price-what mechanisms are involved to smooth out the data. The following are some smoothing techniques:

1. **Binning:** Binning methods smooth out stored data by consulting the neighborhood values around the noisy data. The stored values are distributed into number of "buckets" or bins.

Because binning methods consult the local neighborhood they do local smoothing. In smoothing by means each value in the bin is replaced by the mean value of the bin. Smoothing bins by medians each value in the bin is replaced by the median value for that particular bin. In smoothing by bin boundaries, the largest and smallest value in the bin are consider tobe bin boundaries, then each bin value is replace by the closest bin value to its boundary. In general, the greater the width the larger the effect of smoothing.

2. **Clustering:** Outliers may be detected by clustering, when the similar values are organized into “clusters”. Those values that fall outside the cluster are considered to be outliers.
3. **Combined computer and human inspection:** outliers may be identified by a combination of computer and human inspection. Outlier patterns may be informative (identifying useful data exceptions) or “garbage”. Patterns, which have “surprise” value content, may be output to a list. The human can then sort through the garbage patterns, which can be excluded from use in subsequent data mining.
4. **Regression:** Data can be smoothed by fitting a data to a function, such as with regression. Linear regression gives the best fit line of two variables. Multiple linear regressions is an extension of linear regression where more than two variables are invoked.

5.4 Inconsistent Data:

There may be inconsistencies that are recorded for some data in certain transactions. Errors made at data entry may be made using a paper trace. Knowledge engineering tools may be used for detecting violations of data constraints. There can also be inconsistencies arising from integrating data from different databases.

5.5 Data Integration and Transformation

Data Integration:

It is unlikely that the data analysis task will invoke data integration, which combines data from multiple sources into a coherent data store, as in the data warehousing. The sources may include multiple databases, data cubes, or flat files.

There are a number of issues to be considered during data integration. Schema integration can be tricky. The entity identification problem matches entities from different sources e.g., customer-id in one database and cust_number in one database. Metadata that i8s available can be used help avoid errors during integration.

Some redundancies can be identified by correlation analysis. In addition to detecting redundancies between attributes, duplication should also be identified at the tuple level. A third important issue in data integration is the detection and resolution of data value conflicts. For example, for the same real world entity, attribute values from different sources may differ. This may be due to differences in representation, scaling or encoding e.g., the price in different hotels may be in different currency units.

Careful integration of the data from multiple sources can help reduce and avoid redundancy.

Data Transformation:

In data transformation, the data are real transformed or consolidated into forms appropriate for mining. Data transformation cans involve the following:

- **Smoothing**: which works to remove noise from data?
- **Aggregation:** where summary or aggregation operations are applied to aggregate the data.
- **Generalization:** which works to make low-level “primitives” into higher level concepts using concept hierarchies e.g., categorical attributes like street may be made into city or country.
- **Normalization:** where the attribute data are scaled so as to fall within as small specified range e.g., -1.0 to 1.0 or 0.0 to 1.0
- **Attribute construction:** (or feature construction0, where new attributes are constructed and added from a given set to help the data mining process.

5.6 Data reduction:

Normally data used for data mining is huge. Complex analysis on data mining on huge amounts of data can take a very long time, making such analysis impartibly or infeasible.

Data reduction techniques can be applied to obtain a reduced representation of the data set that is smaller in volume yet closely maintains the integrity of the original data. That is, mining on the educed set should be efficient and yet produce the same or almost the same analytical results.

5.7 Data Cube Aggregation:

Data cube store multidimensional aggregated information. Each cell holds an aggregate data value, corresponding to the data in the multidimensional space. Concept hierarchies may exist for each attribute, allowing analysis of data at multiple levels of abstraction. For example, a hierarchy for branch could allow branches to be grouped into regions, based on their address. Data cubes provide fast access to pre computed summarized data thereby benefiting online analytical processing as well as data mining.

The cube created at the lowest level of abstraction is referred to as the base cuboids. Data cubes created at various levels of abstraction are called as cuboids, so that a data cube may refer to a lattice of cuboids. Each higher level of abstraction further reduces the resulting data size. The lowest level cuboids should be useable or used for data analysis.

5.8 Dimensionally Reduction:

Data sets may contain hundreds of attributes for analysis most of which may be irrelevant to the data mining task, or redundant. Although it may be possible for a domain expert to pick out certain attributes, this cans a difficult and a time consuming task. Leaving out relevant attributes or keeping irrelevant attributes may cause confusion for the mining algorithm employed. The redundant data may slow down the mining process.

Dimensionally reduction reduces the data set size by removing such attributes (dimensions) from it. Typically methods for attribute selection are applied. The goal of attribute subset selection is to find the minimum number of attributes so that the resulting probability pattern of the data class is as close as the original distribution obtained using all the attributes. Mining on a reduced set has an additional benefit. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand.

Basic heuristic methods of attribute selection include the following techniques:

1. **Stepwise forward selection:** The procedure starts with an empty set of attributes. The best of the original attributes is determined and added tom the set. At each subsequent iteration or step, it removes the worst attribute remaining original attributes is added to the set.
2. **Stepwise backward elimination:** The procedure starts with the full set of attributes and at each step, it removes the worst attribute remaining in the set.
3. **Combination of forward selection and backward elimination:** The stepwise forward selection and backward elimination methods can be combined so that at each step, the procedure selects the best attributes and removes the worst from among the remaining attributes.

If the mining task is classification, and the mining algorithm itself is used to determine the attribute subset, then it is called a wrapper approach, otherwise a filter approaches. In general the wrapper approach is leas to greater accuracy since it optimizes the evolution measure of the algorithm while removing attributes. However, it requires much more computation than a filter approach.

5.9 Data Compression:

In data compression, data encoding or transformations are applied so as to obtain reduced or "compressed" representation of the original data. If the original data can be reconstructed from the compressed without loss of information the data compression technique is called lossless. If, instead we can reconstruct a partial approximation to the original data it is called lossy. Here two popular and effective methods of data a compression, which are lossy, are explained.

5.10 Review questions

1 Why we preprocess data

2 Expalin aboutData cleaning:

3Explain about Data Integration and Transformation

4 Explain about Data Compression.

5.11 References

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition