Business Intelligence

ADDIS ABABA SCIENCE AND TECHNOLOGY UNIVERSITY
UNIVERSITY FOR INDUSTRY

Week 5

Data Integration and Preparation

- Data Transformation
- Data reduction
- Feature Selection
- Data Discretization and concept hierarchy generation

Tilahun Melak(PhD)

April, 2026

Objectives

At the end of this lecture students will be able to :

- Understand data transformation techniques
- Understand data reduction techniques
- Explain feature selection methods
- Explain data discretization and concept hierarchy generation

Data Transformation

- Data transformation is the process of transforming or consolidating data into a form appropriate for mining which is more appropriate for measurement of similarity and distance
- This involves
  - Smoothing
  - Aggregation
  - Generalization
  - Normalization
  - Attribute/feature construction

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Transformation…

- **Smoothing:** concerned mainly to remove noise from data using techniques such as binning, clustering, and regression
- **Aggregation:** summarization, data cube construction
- **Generalization:** concept hierarchy climbing (from low level into higher level)

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Transformation…

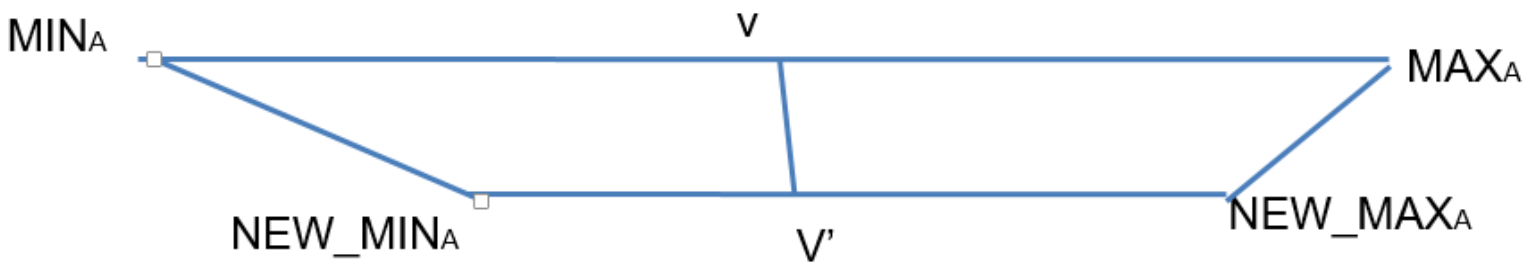
- **Normalization:** scaled to fall within a small, specified range
  - Used mainly
    - for classification algorithms such as neural network,
    - distance measurements such as clustering, nearest neighbor approach
  - Exist in various forms

1. min-max normalization
2. z-score normalization
3. normalization by decimal scaling
4. Attribute/feature construction

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Transformation: Normalization

- **min-max normalization**
  - Perform a linear transformation on the original data into a range specified range of min and max value



$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Transformation: Normalization…

- z-score (zero mean) normalization
  - A value will be normalized based on the mean and standard deviation of the original data

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Transformation: Normalization…

- Normalization by decimal scaling
  - Normalizes by moving the decimal point of values of attribute A.
  - The resulting value ranges from -1 to +1 exclusive

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that $\text{Max}(|v'|)<1$

- For example, given the data set V= 132, -89, 756, -1560, 234, -345 and 1234
- The value of v' becomes in the range from -1 to +1 if j=4
- In that case V' = 0.0132, -0.0089, 0.0756, -0.156, 0.0234, -0.0345, and 0.1234

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Transformation: Attribute construction

- Attribute construction is the process of driving new attributes from the existing attributes.
- Attribute construction is important to improve performance of data mining as the derived attribute will have more discriminative power than the base attributes
- Enable to discover missing information or information hidden within the data set
- For example
  - Rate of telephone charge can be derived from billable amount and call duration
  - area can be derived from width and height

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Reduction

- Data sources may store terabytes of data
- Complex data analysis/mining may take a very long time to run on the complete dataset
- Data reduction tries to obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same or better) analytical results

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Reduction…

- Data reduction strategies includes
  1. **Data cube aggregation**
     - Apply data cube aggregation to have summarized data to work with DM functionalities
  2. **Attribute subset selection**
     - Select only relevant attribute from the set of attributes that the data set has

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Reduction…

- Data reduction strategies
  3. Numerosity reduction
     - **Regression and log-linear models** (once analyzed store only the parameters that determine the regression and the independent variable values)
     - **Histograms:** store the frequency of each bin in the histogram rather than each element in detail
     - **Clustering:** store the cluster labels rather than the individual elements
     - **Sampling:** use representative sample data rather than the entire population of data

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Reduction…

- Data reduction strategies includes
  4. **Dimensionality reduction**
     - ***Huffman coding*** (text coding or compression algorithm)
     - ***Wavelet transforms*** (transforming usually image data into important set of coefficients called wavelets)

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Reduction…

- Data reduction strategies includes
  4. **Dimensionality reduction**
     - ***Principal component*** analysis (representing N sequence of M dimensional data by using small parameters having either N or M elements which ever is possible)
       - It can be used for a series image data reduction where N can be the number of pixels and M the important information on the pixel at the same location
       - If each parameter has N elements, we use it for generating every image from the parameters and the reduced image coefficients
       - If each parameter has M elements, we use it for generating every pixel of all the image from the parameters and the reduced pixel coefficient values

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data reduction strategies: Data Cube Aggregation

- Data cube aggregation and using it for data mining task reduces the data set size significantly
- For example, one can aggregate sales amount specified at each year and quarter into the sum of the sales amount per year
- Multiple levels of aggregation in data cubes further reduce the size of data to deal with
- One should select appropriate levels of aggregation
- Use the smallest representation which is sufficient to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data reduction strategies: Attribute subset selection

- Removes irrelevant attribute by attribute relevance analysis

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close possible as to the original distribution given the values of all features
  - reduce # of patterns in the patterns, so that it will be easier to understand

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data reduction strategies: Attribute subset selection…

- Let us assume we have **d** set of attributes in the data set.

- This set has 2^d sub sets of attributes and dimensionality reduction refers to selection of the subset which has the minimum number of elements in it and represent the pattern as close as possible with the original attributes

- Hence dimensionality reduction refers to attribute subset selection

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data reduction strategies: Attribute subset selection…

- Several heuristic for attribute subset (feature) selection exists
- Four of them are:
  - step-wise forward selection
  - step-wise backward elimination
  - combining forward selection and backward elimination
  - decision-tree induction algorithm

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data reduction strategies: Attribute subset selection…

- step-wise forward selection
  - Start with empty set
  - The best single-feature is picked first
  - Then next best feature will be selected conditioned by the first, ...
  - Stop when the selected feature set closely represent the entire features

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data reduction strategies: Attribute subset selection…

–step-wise backward elimination

- Start with all the feature set elements
- The feature which is most irrelevant will be discarded first
- Then next most irrelevant feature will be discarded and repeated, ...
- Stop when removing the next candidate attribute for removal affects the pattern significantly

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data reduction strategies: Attribute subset selection…

- combining forward selection and backward elimination
  - At each step, the procedure selects the best feature and remove the most irrelevant

- decision-tree induction algorithm
  - This algorithm generate a decision tree using some of the attributes
  - The attributes used in building the decision tree will be taken as attributes that represents closely the entire attributes

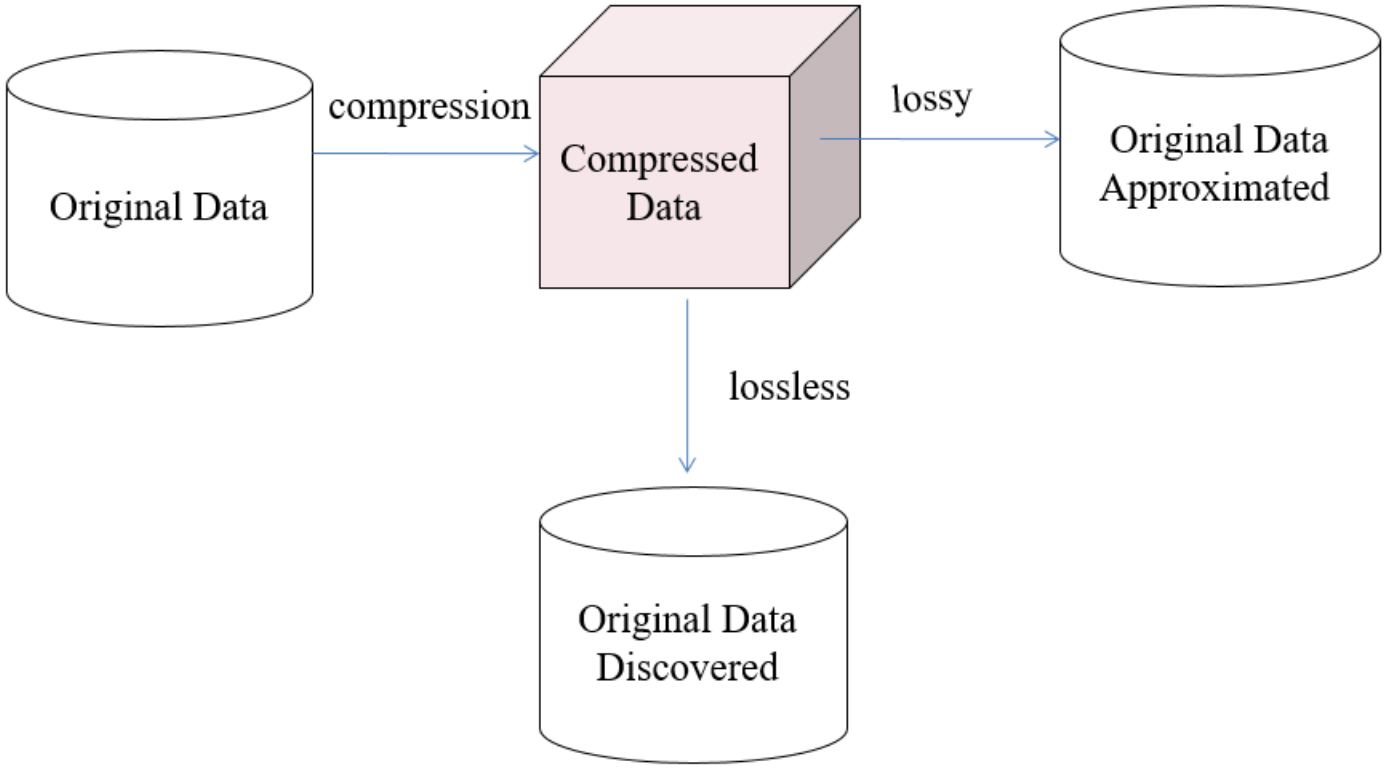
Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data reduction strategies: Dimensionality reduction

- Tries to compress the data using encoding scheme such as
  - minimum length encoding,
    - Huffman Encoding
  - wavelet encoding,
  - principal component analysis, etc

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data reduction strategies: Dimensionality reduction…



Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data reduction strategies: Dimensionality reduction…

- Compression can be made on data such as string, audio, and video
- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
  - But only limited manipulation is possible without expansion

- Wavelet transformation and principal component analysis are two of the most common dimension reduction approaches which are lossy

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data reduction: Numerosity Reduction

- Refers to replacing the data by alternatives smaller form of the same data representation
- This can be parametric or non-parametric methods

- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - Some examples include: Gaussian distribution, regression and log-linear models
- Non-parametric methods
  - Do not assume models
  - Find alternate smaller representation and discard the data
  - Major families: histograms, clustering, sampling

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Discretization and concept hierarchy generation

- Data discritization refers to transforming the data set which is usually continous into discrete interval values

- Concept hierarchy refers to generating the concept levels so that data mining function can be applied at specific concept level

- Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of attribute into intervals

- Interval labels can be used to replace actual data values

- This leads to concise, easy to use, knowledge level representation of mining result

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Discretization and concept hierarchy generation…

- Discretization:
  - Divide the range of a continuous attribute into intervals
  - Reduce data size by discretization
  - Prepare for further analysis
  - Some classification algorithms only accept categorical attributes.
- Concept hierarchies
  - Reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).
  - Concept hierarchy generation refers to finding hierarchical relation in the data set and build the concept hierarchy automatically

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Discretization and concept hierarchy generation for numeric data

- It is difficult and laborious to specify concept hierarchies for numerical attributes because of the wide diversity of possible data ranges and frequent update of data values

- Concept hierarchies for numerical data can be constructed automatically based on the data discretization

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Discretization and concept hierarchy generation for numeric data…

- The following are methods for discretization and concept hierarchy generation
  - Binning :
    - as shown in the previous sections this will discretize the attribute and doing it recursively in a top down approach will generate concept hierarchy
  - Histogram analysis
    - as shown in the previous sections this will discretize the attribute and doing it recursively in a top down approach will generate concept hierarchy

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Discretization and concept hierarchy generation for numeric data…

- The following are methods for discretization and concept hierarchy generation
  - Clustering analysis
    - as shown in the previous sections this will discretize the attribute and doing it recursively in a top down or bottom up approach will generate concept hierarchy
    - The bottom up approach will start by generating larger number of class and incrementally merge some of them forming the concept hierarchy from the bottom up till the top

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Discretization and concept hierarchy generation for numeric data…

- Some more methods for discretization and concept hierarchy generation of numeric data
  - Entropy-based discretization
  - Segmentation by natural partitioning
  - Interval merging by χ^2 Analysis
  - Clustering

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Summary

- In todays lecture we have discussed about;
  - Covers key data preprocessing techniques
  - Introduces data transformation and data reduction
  - Explains feature selection methods
  - Discusses data discretization and concept hierarchy generation

References

- Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.