MUST

Wisdom & Virtue

MIRPUR UNIVERSITY OF SCIENCE AND TECHNOLOGY (MUST), MIRPUR

DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

DATA MINING

BCS-3605

Lecture 15

Dr Yasir Mehmood

(Assistant Professor)

Agenda of Today's Lecture

- ***Data Reduction***
  - ***A-Dimensionality Reduction***
    - ***1-Wavelet Transformation***
    - ***2-Principal Component Analysis***
    - ***3-Attribute Subset Selection***

Data Reduction

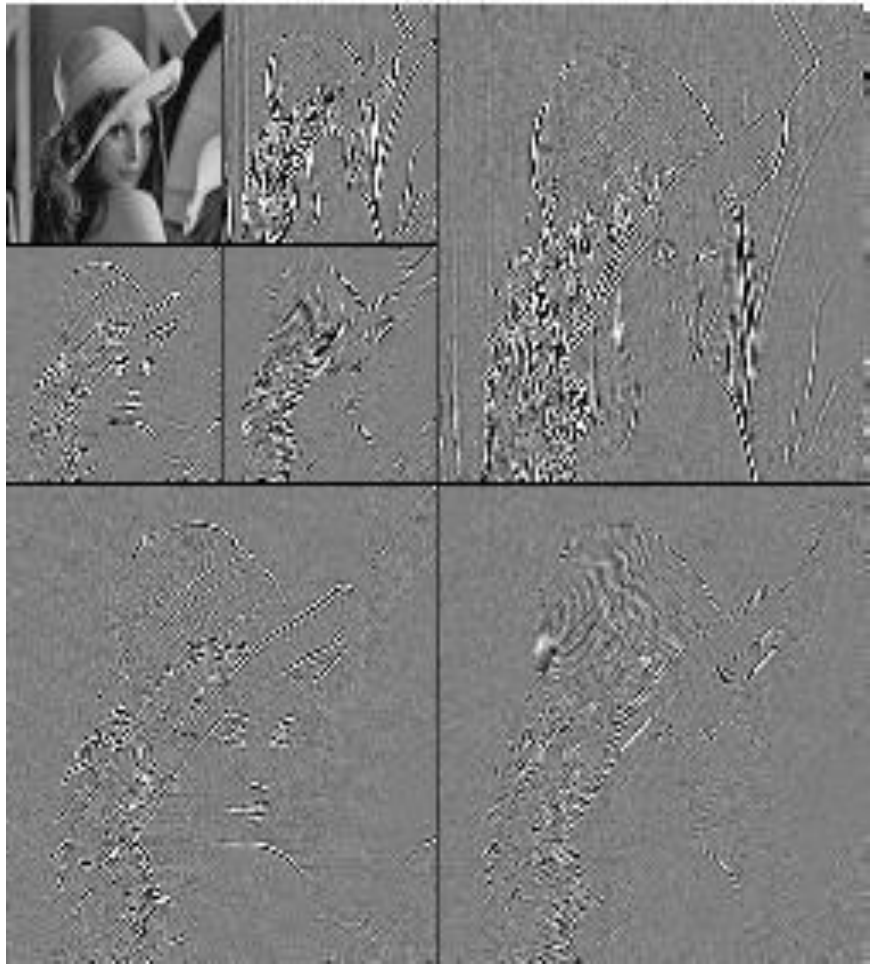
- Obtain a reduced representation of huge data set that produces the same analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
  - Dimensionality reduction (remove unimportant attributes)
    - Wavelet transforms
    - Principal Components Analysis (PCA)
    - Feature subset selection, feature creation
  - Numerosity reduction (replace by alternative smaller form)
    - Parametric, estimate the data and instead of actual data, storing the parameters only, includes Regression and Log-Linear models
    - Non-parametric, storing reduced representations of the data, Histograms, clustering, sampling, Data cube aggregation
  - Data compression (apply transformation to compress the data)
    - Lossless (If the original data can be *reconstructed* from the compressed data without any information loss)
    - Lossy(we can reconstruct only an approximation of the original data)

A-Dimensionality Reduction

- **Curse of dimensionality**
  - When dimensionality increases, volume of space increases and data becomes sparse.
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- **Dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization
- **Dimensionality reduction techniques**
  - Wavelet transforms
  - Principal Component Analysis
  - Supervised and nonlinear techniques (e.g., feature selection)

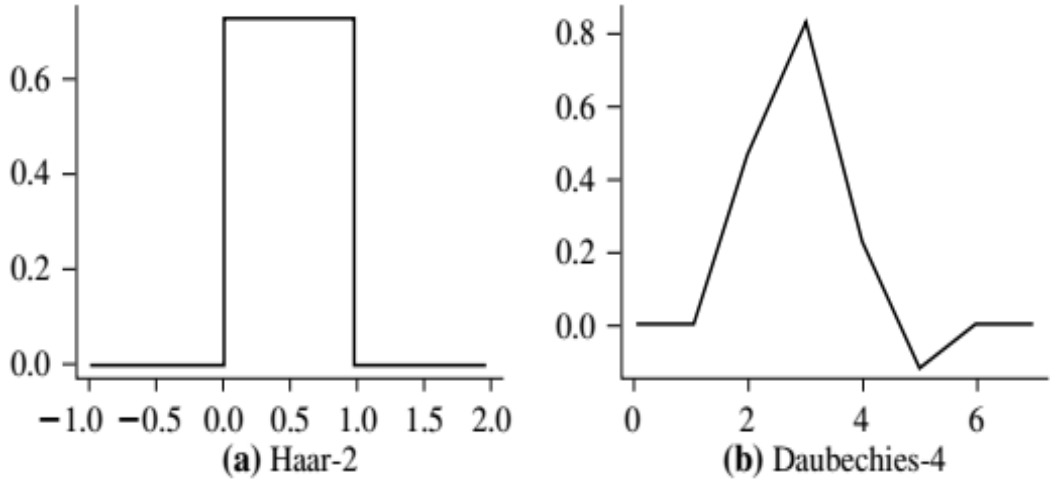
1-Wavelet Transformation

- Decomposes a signal into different frequency subbands
  - Applicable to n-dimensional signals
- Data are transformed to preserve relative distance between objects at different levels of resolution
- Allow natural clusters to become more distinguishable
- Used for image compression



1-Wavelet Transformation Cont'd

- Discrete wavelet transform (DWT) is a linear signal processing, multi-resolution analysis, transform data into wavelet coefficients
- Store only a small fraction of the strongest of the wavelet coefficients
- Lossy compression that retains the coefficients larger than the certain threshold and set all coefficients to 0
- The sparse representation is computationally efficient and remove noise without smoothing
  - Haar-2, Daubechies-4, 6
- Original data can be constructed by applying the inverse of the DWT used.
- Good for sparse, ordered, and high dimensional attributes.



1-Wavelet Transformation Cont'd

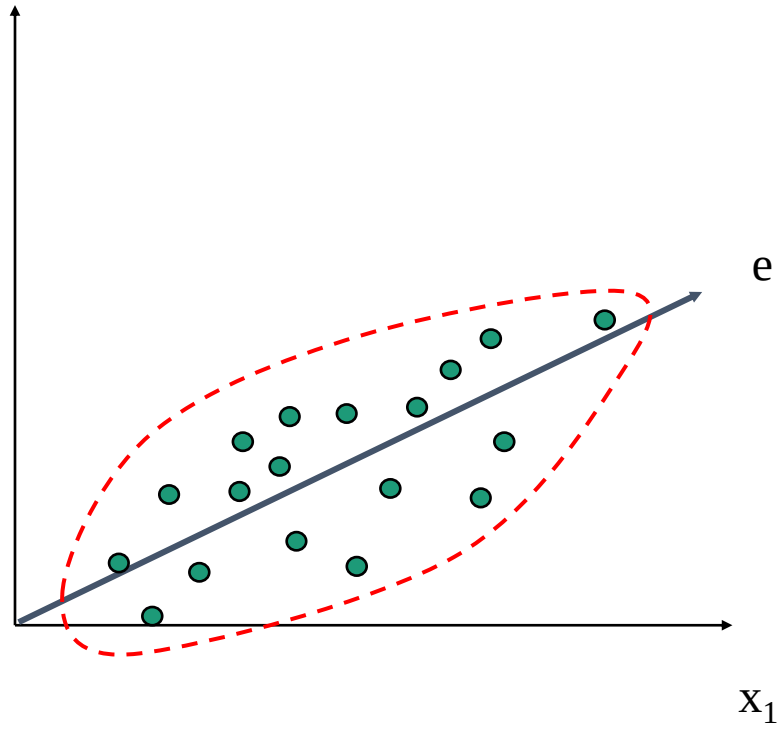
- The DWT is closely related to the *discrete Fourier transform (DFT)*, it provide a more accurate approximation of the original data for an equivalent approximation, the DWT requires less space than the DF

- Hierarchical pyramid algorithm that halves the data at each iteration:

  - Length, L, of input data must be an integer power of 2 (padding with 0's, when necessary)
  - Each transform has 2 functions: smoothing, difference
  - Applies to pairs of data, resulting in two set of data of length L/2
  - Applies two functions recursively, until reaches the desired length
  - Selected values are designated the wavelet coefficients of the transformed data.

2-Principal Component Analysis

- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



2-Principal Component Analysis Cont'd

- Also called Kerhunen-Loeve or K-L method, searches for *k* n-dimensional orthogonal vectors (*principal components*) that represents data where $k < n$
- The data is projected into smaller space with reduced dimensions
  - Normalize input data: Each attribute falls within the same range
  - Compute *k* orthonormal (unit) vectors, i.e., *principal components*
  - Each input data (vector) is a linear combination of the *k* principal component vectors
  - The principal components are sorted in order of decreasing "significance" or strength
  - Since the components are sorted in decreasing order of "significance", the size of the data can be reduced by eliminating the *weak components*, i.e., those with low variance.
- Applied to ordered and unordered attributes, and can handle sparse and skewed and high dimensionality data
- Works for numeric data only

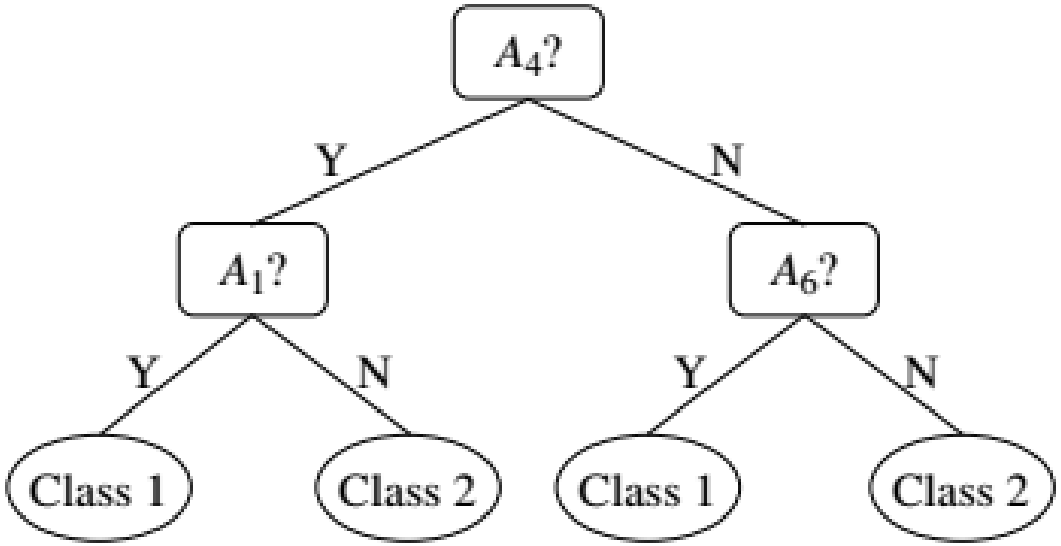
3-Attribute Subset Selection

- Another way to reduce dimensionality of data
- Find a minimum set of attributes to make the patterns easier to understand.
- Reduces the data set size by removing redundant or irrelevant attributes
  - Redundant attributes
    - Duplicate much or all of the information contained in one or more other attributes
    - E.g., purchase price of a product and the amount of sales tax paid
  - Irrelevant attributes
    - Contain no information that is useful for the data mining task at hand
    - E.g., students' ID is often irrelevant to the task of predicting students' GPA

3-Attribute Subset Selection Cont'd

- For n attributes, 2^n subsets. Greedy heuristic methods
  - Stepwise forward selection
    - start with empty and at each iteration, best attribute is determined and added into the reduced set
  - Stepwise backward elimination,
    - start with full set and at each step it removes the worst attribute remaining in the set
  - Combined,
    - select best attribute and remove worst among the remaining attributes.
  - Decision tree,
    - includes only relevant attributes

3-Attribute Subset Selection Cont'd

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

Reference

Data Mining Concepts and Techniques Third Edition

3.4 Data Reduction

THANKS