Strategies of Data Reduction

Data Cube Aggregation

Attribute Subset Selection

Dimensionality Reduction

Numerosity Reduction

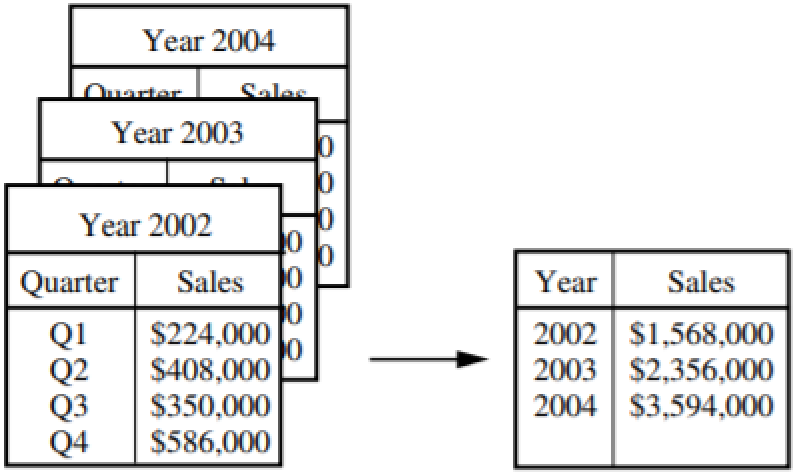
Discretization and Concept Hierarchy Generation

Data Cube Aggregation

Data Cube Aggregation

Aggregation operations are applied to the data in the construction of the data cube

Data Cube Aggregation : Example



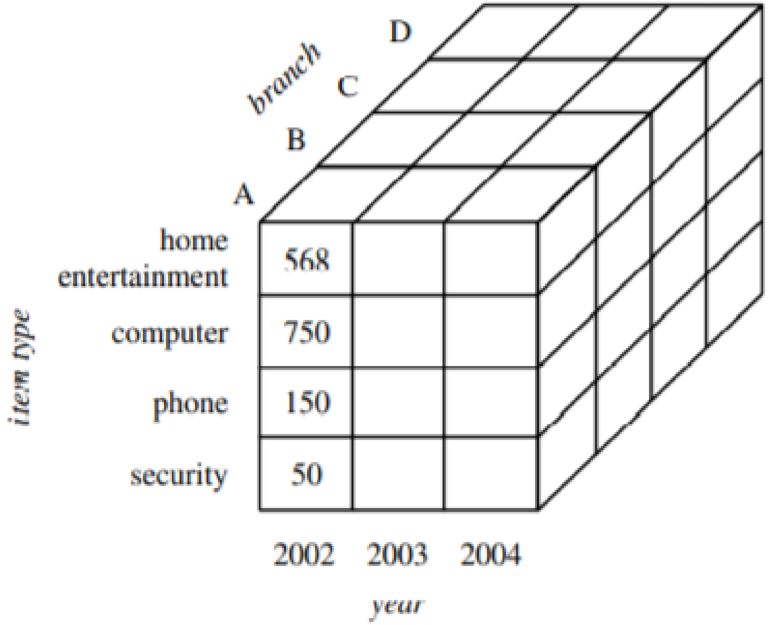
Year 2004

Year 2003

Year 2002

Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
2002	$1,568,000
2003	$2,356,000
2004	$3,594,000



Attribute Subset Selection

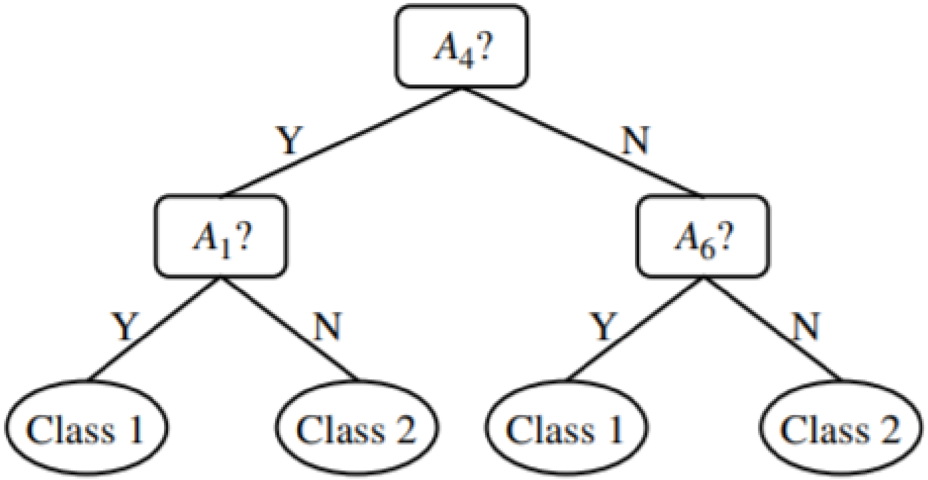
“where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.”

Reduces the dataset size

Minimum set of attributes

Attribute Subset Selection

1. Stepwise forward selection
2. Stepwise backward elimination
3. Combination of forward selection and backward elimination
4. Decision tree induction

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: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  A_4? Y N A_1? A_6? Y N Y N Class 1 Class 2 Class 1 Class 2 => Reduced attribute set: $\{A_1, A_4, A_6\}$

Dimensionality Reduction

Data encoding or transformation methods are applied – to obtain either a reduced or compressed representation of the original data

Lossless methods

Lossy methods

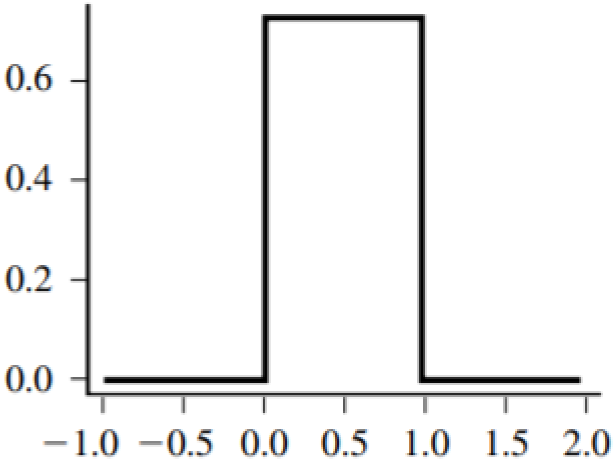
Effective methods for Lossy Dimensionality Reduction

Wavelet Transformation

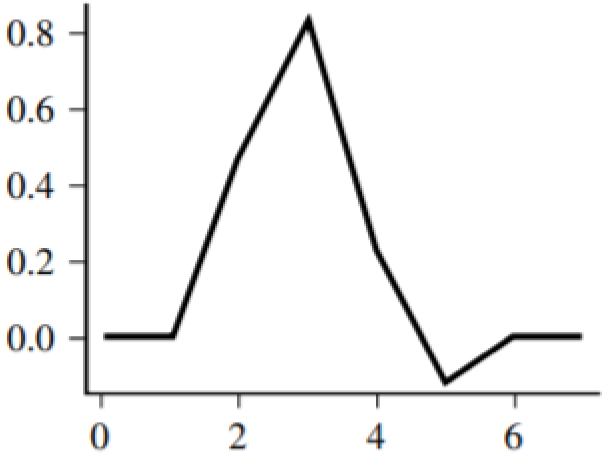
Principal Components Analysis

Wavelet Transforms

The discrete wavelet transform (DWT) is a linear signal processing technique that, when applied to a data vector X, transforms it to a numerically different vector, X0 , of wavelet coefficients



(a) Haar-2



(b) Daubechies-4

Principal Components Analysis

The original data are thus projected onto a much smaller space, resulting in dimensionality reduction.

Unlike attribute subset selection, which reduces the attribute set size by retaining a subset
of the initial set of attributes

searches for k n-dimensional orthogonal vectors
that can best be used to represent the data, where k ≤ n.

Principal Components Analysis

The basic procedure is as follows:

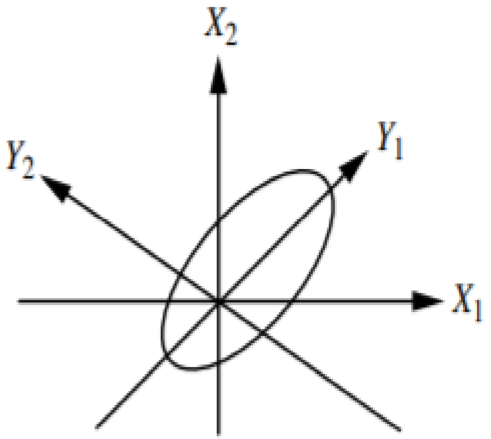
1. The input data are normalized, so that each attribute falls within the same range

2. PCA computes k orthonormal vectors that provide a basis for the normalized input data.

3. The principal components are sorted in order of decreasing “significance” or strength.

4. Because the components are sorted according to decreasing order of “significance,” the size of the data can be reduced by eliminating the weaker components

Principal Components Analysis



Principal components analysis. $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are the first two principal components for the given data.