Min-Max Normalization and Example

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

Min-max normalization. Suppose that the minimum and maximum values for the attribute *income* are \$12,000 and \$98,000, respectively. We would like to map *income* to the range [0.0, 1.0]. By min-max normalization, a value of \$73,600 for *income* is transformed to $\frac{73,600-12,000}{98,000-12,000}(1.0-0)+0 = 0.716$. ∎

Z-Score Normalization and Example

$$v' = \frac{v - \bar{A}}{\sigma_A},$$

z-score normalization Suppose that the mean and standard deviation of the values for the attribute *income* are \$54,000 and \$16,000, respectively. With z-score normalization, a value of \$73,600 for *income* is transformed to $\frac{73,600-54,000}{16,000} = 1.225$. ■

Decimal Scaling and Example

$$v' = \frac{v}{10^j}$$

Decimal scaling. Suppose that the recorded values of A range from -986 to 917. The maximum absolute value of A is 986. To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., $j = 3$) so that -986 normalizes to -0.986 and 917 normalizes to 0.917. ■

Problem

Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order)

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

(a) Use min-max normalization to transform the value 35 for age onto the range [0.0,1.0].
(b) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.
(c) Use normalization by decimal scaling to transform the value 35 for age.
(d) Comment on which method you would prefer to use for the given data, giving reasons as to why.

Problem

Use the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000

(a) min-max normalization by setting min = 0 and max = 1
(b) z-score normalization (stdev=316.22)

Data	Min-Max Normalized Values	Z-Score Values
200	?	?
300	?	?
400	?	?
600	?	?
1000	?	?

Data Reduction

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.

Prepared By : Dr K RAJENDRA PRASAD, PROFESSOR, DEPT. OF CSE , RGMCET (Autonomous), Nandyal

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

Prepared By : Dr K RAJENDRA PRASAD, PROFESSOR, DEPT. OF CSE , RGMCET (Autonomous), Nandyal

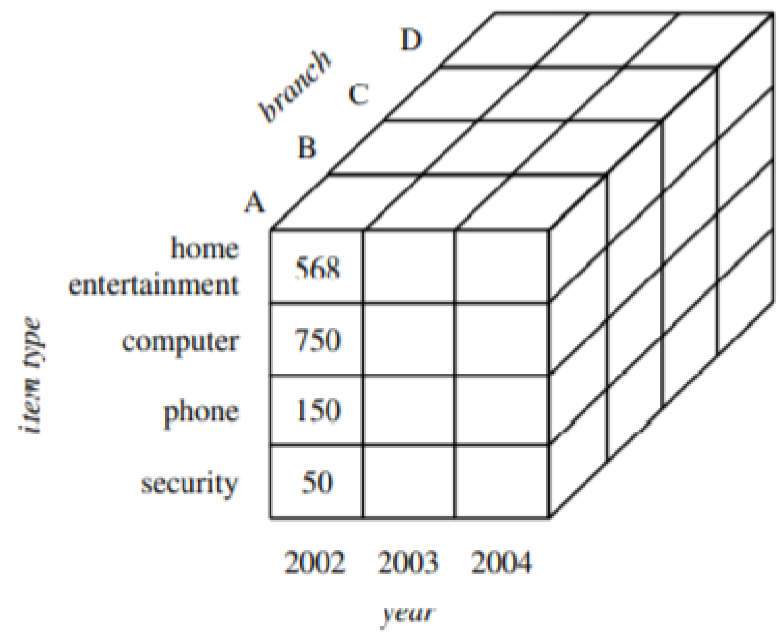
Data Cube Aggregation : Example

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.”

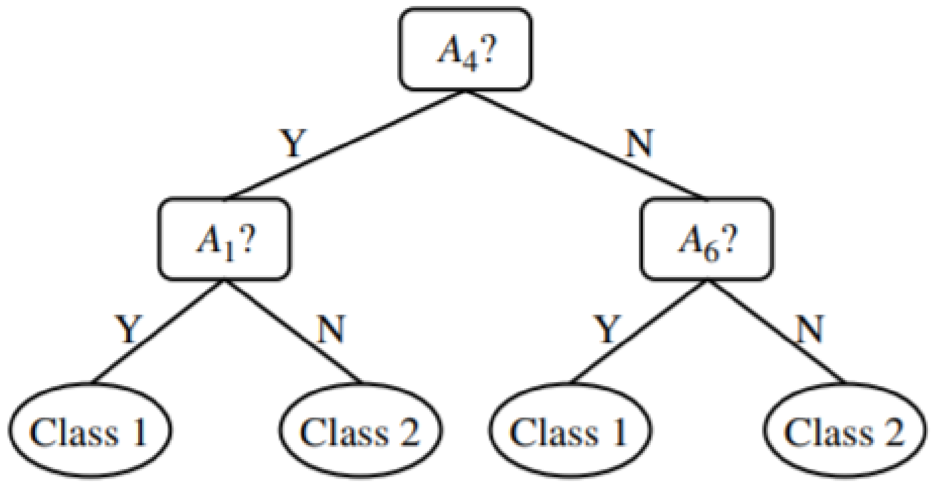
Reduces the dataset size

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

Prepared By : Dr K RAJENDRA PRASAD, PROFESSOR, DEPT. OF CSE , RGMCET (Autonomous), Nandyal

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: $\{A_1, A_4, A_6\}$