Techniques : Data Reduction

1. Data cube aggregation
2. Attribute subset selection
3. Dimentionality Reduction
4. Numerosity reduction
5. Discretization and concept hierarchy generation

---

(1) Data cube aggregation

2017

Product	sales
H1	500
H2	200

⇒

year	sales
2017	700
2018	600

} storing data year wise apply aggregation

2018

Product	sales
H1	300
H2	300

(2) Attribute subset selection

irrelevant att. may be discarded/Remove | original

Redundant att

Sr. No	Roll. No	Name	gender	Religion	Marks

Compare student performance

select best / relevant att

a_1	a_2	a_3	a_4	a_5

⇒

a_2	a_4	a_5

⇒

selected

gender	Marks

must follow original dataset

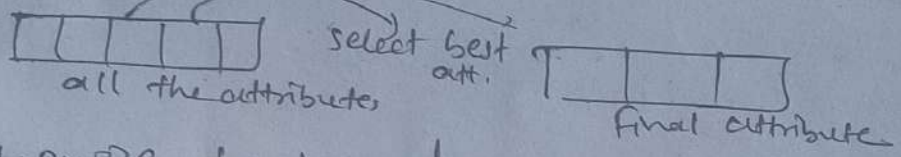
- Exhaustive search tech (identify each & every combination)
- greedy method
- Best heuristic method

VP → Verb
VP → Verb NP
VP → VP PP
PP → Prep NP

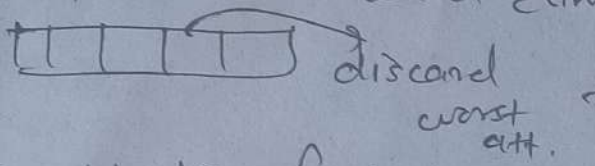
Attribute subset selection : Retain original Data

Basic heuristic methods of attribute selection :

1) Stepwise forward solution

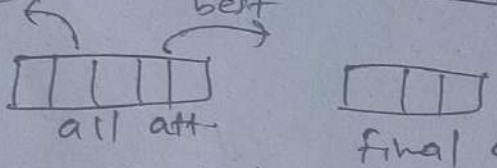


2) stepwise backward Elimination



then form final att.

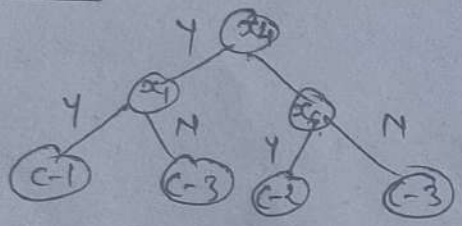
3) Combination of both 1) & 2)



4) decision tree induction : different algo. ⇒ classification

$(x_1, \ldots x_6)$

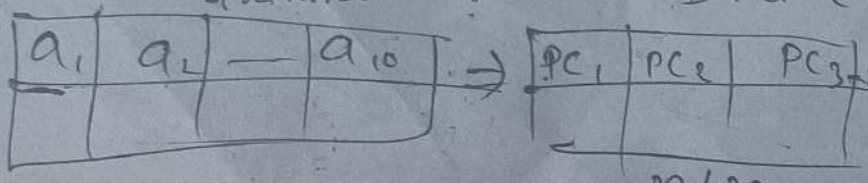
Retain 3 att
discard all other att.



Dimentionality Reduction

↳ principle component analysis – $N^D \rightarrow n^D$ where $n \leq N$

Transform data so Data change



④ Numerosity Reduction

↳ | a | b | c |

$y = x_1 + ax_2 + bx_3 + cx_4$

⇒ equation form through Regression

↳ clustering : grouping data

just store cluster center instead of storing comple dataset

⑤ data discretization & concept hierarcy generation

(0 — — 100)

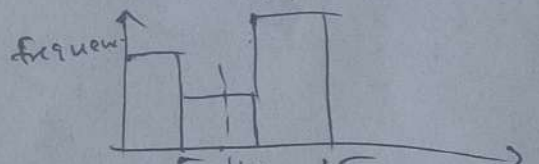
(0-10) (20-30) —

break into small parts

- Best heuristic method

↳ histogram

2, 2, 5, 5, 10, 10, 12, 12, —



↳ sampling

randomly select few data

10, 20, 30, 50, 50, 80, 90

young

middle age

senenter

group into higher level

Feature Selection / Attribute Selection Techniques

why? — Raw data training → model generate → low efficiency
↓ why?
Not relevant features / Noise

optimal features / relevant features

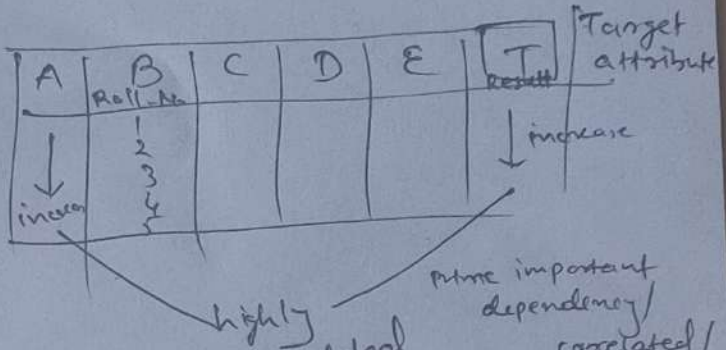
⇓

- Subset of features pass to training
  - generate model

(1) Filter Method : relevant of att.

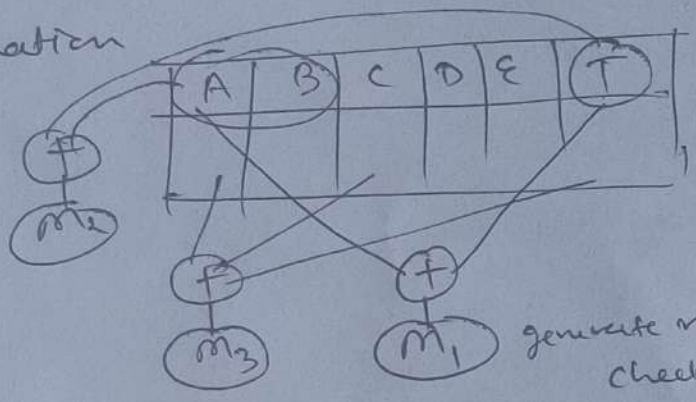
- Information gain
- chi-square Test
- correlation coef.

Ex



(2) Wrapper Method

- Recursive feature Elimination
- Genetic Algo.



take a subset of att. & merge with Target & generate multiple model
- A
- AB
- ABC
- AC

generate model & check accuracy & error rate

DB :- highly computationally expensive ↳ overfitting problem.

(3) Embedded methods } both (2) + (3) ⇒ find attribute usefulness

- Decision Tree

less computationally expensive
overfitting ~~very less~~ does not encounter

sunny cloudy rainy