Data Mining and warehousing

Converting DBMS to data warehousing - data pre-processing., ETL tool is used (extract, Transform, Load)

* Why data mining - All DM will be used in ML
  There is explosive growth of data from terabyte to petabyte.

→ activity of user - help in recommendation system.
Ex) case study of 2003 is considered, when internet not used so much

* Pattern retrieve करना है data set से - analytic used.
* Evolution of Sciences

① Data collection & database creation (before DBMS whatever you learnt) → Primitive file processing.
1960s & earlier. No GUIs, window. so, file के अंदर data store करते थे। so, data retrieve करने के लिए application programing lang like C++, Java में program करना पड़ा। problem if, C++ के user ने & अब यदि C के user ने access किया। so, intermediate code लिखना पड़ा। so, problem. (1970 - 1980 early)

so ② DBMS comes — Hierarchical & network DS.
└ Relational DBS — last ran wal
(ER, indexing, SQL, report,

Scanned by TapScanner

~~First~~ subquery को पहले process करना (अधिकतम no of tuples कम transfer करना है)
↑
query processing, transaction, online transaction processing

(3) Advanced Database system (mid 80s - present)
spatial - geoinformatics, satellite, sequence, structural
semantic web - meaning of whatever you type in google
Page is ranked on the basis of hit ratio.

(4) Advanced data analysis - OLAP → for retrieving data from data warehouse. OLAP tool use data mining algo.
classification, clustering, outlier, association & correlation.

multidimensional model of datawarehouse -
* In cube, sales are shown x - item, y - time, z = city

* data is classified in various categories.

Ex) 2003 case study, no net, phone, no OS, RAM with less memory, suppose person start e-commerce he started 4 major shop in India. He need barcode scanner. He took frontend .dot & mysql in backend, developed interface & database dataset बनाया ID, product name, salesperson. 4 जगह software लगाया,
In 2005, frontend netbeans, backend SQL server. Product ID, attribute = barcode. table created after scope of analytics come. summarized report बनाओ, dataset पर है & database मे है। link देने चाहे, csv मे convert करो इस data को

multiple dataset है। with this help find in which region sale is less
1000×1000×10 products per day you create on server side
summarized info or result is stored in separate store which is called data warehouse (no write can be done, only read). Only those attributes stored which are important for decision making.

* temporal - time factor हो वहाँ like weather forecasting data
* spatial → space related, sequence - weather forecasting all frames in sequence.

* Data mining → (knowledge discover from data)
Processed data → information,
combining multiple info together & finding pattern → knowledge.
→ extraction of interesting

* Data mining as step in process of knowledge discovery -

* Basic flow of ML ~~or~~ data analytic part -

(1) databases (multiple datasource ex SQL, SAP, ERP) - data fetch करो। इन sources से।
(2) Data warehouse (सारा data 3D or 4D मे होता है) only sale is stored in cube वाला labelling is done.

Scanned by TapScanner

store sale in data warehouse.

ETL is used to transform database to data warehouse.
(↑ Relational database)

→ Data warehouse → 3 types of schema is stored.
→ we need sale, from ~~city~~, item, time, location.
select count of item[4] from table where city = ' ' and time = ' '.

sale is ~~stored in this way (multi-dimensional)~~

1 quarter = 3 month.
3 array is defined है। 1 array for Jan other arr of feb. march

400 is sale of 150, 100, 150 (Jan, Feb, March)

Jan also divided into 4 week, [ ] [ ] [ ]
week also divided into daily basis

(all segregate coz query can be made on any level.)

* data sale daily base से segregate करना है। sale को different abstraction level पे store कर रहे। we're storing sale in different level

* Datawarehouse always single value stored in 1 cube

Database की info metadata के अंदर रहती है। ~~with help of metadata~~, ~~with error~~

we clean data first, last 5 years चाहिए बाकी years को remove करा। now we keep only city, time, item, sale.

* Now, in datawarehouse (→ multi dimensional) value linked with 3 values

(3) Datawarehouse में (array) OLAP tool implement करेंगे so that patterns find करें using data mining algo

(4) on basis of patterns, knowledge base is created

* Roll up करना city को। → countries में convert किया
drill down - quarter ~~month~~ से month में आओ।

* Decision making historical data से करते है
~~To make product selling~~

* Normal database से knowledge पहुंचना है through

(1) Database (different sources are there) ~~(fact)~~

(2) Data warehouse (1 DW) बनेगा from different ~~source~~ databases.
ETL used to transform database to ~~datawarehouse~~

Schema in DW:
- Star
- Snowflake
- ~~Constellation~~ (fact constellation)

(S) table:
- Factual table
- Dimension table

depend on ...

Page No. | Date:

Cube में city की concept hierarchy बनाओ।
city store base पे ही sale होती है
city break into street.

* DW is subject oriented.

(4) Data mining algo से patterns find करो। operate on DW (OLAP tool भी कहते हैं) on basis of pattern we develop ~~pattern~~, knowledge.

* जितनी dimensional link करेंगे उतने pattern पता लगेंगे।

Lecture-3

* Datamining → algo deployed in Machine Learning lang using R, Python.

* mining algo operate on—
(1) Data stream (youtube) & sensor data (weather forecasting)
(2) Temporal
(3) Spatial data
(4) Multimedia databases
(5) Text database, www

* DataMining Function— (1) Class/concept description (Generalization), unit-3

Data mining categorize tasks into—

operating query on ...

Page No. | Date:

(1) descriptive → characterize properties of data in target dataset
(2) Predictive → current data से induction in order to make predictions.

* Whatever algo we create are object oriented concept, real time problem we're solving

works on basis of attendance
→ If we summarizing dataset of class → data characterised
→ 1 class को दूसरे class compare करना → discrimination.

• Suppose coke, Pepsi are selling. If sale of coke ↑ then pepsi ↑. this association can be find out on basis of discrimination.

(2) Mining Frequent Patterns, → On E-commerce you do shopping, suggestions are made (Association, correlation.)

Market Product 1 – most sale
2 – less sale
} so, they keep combo where they kept items (1) & (2) together so that (2) can get easily sold.

* Transactional data में check occurrence. count of item1, count of item2 & count of item1 & 2.

• Subsequence → पहले ~~lapto~~ occurrence एक साथ नहीं हो रही। पहले laptop फिर webcam फिर memory card.

• Frequent pattern → एक साथ sale milk & bread.

* Substructure → milk < full cream / less cream. In Railway Station

in all cities use full cream milk, all juice shop near hospital for all city → substructure.
include frequent set is subsequence

→ support & confidence factor in frequent set.
out of 100, 50 milk & bread buy की, then
~~confidence~~
support = 50% (actual में हो गया जितनों buy कर लिया)
confidence - 50% (जो होगा next day)

• उस association rule को वहां पहुंचने के लिए min कितना observe करना - min threshold

③ classification - (supervised)
ML train then we predict [class level पता है / जो pattern पहले बार आ रहे]

(decision tree, neural, ... classif)

clustering - unsupervised.

After 5 attributes 10, 12, Btech %, communication skill, technical, placement → we train model on basis of these attributes then we predict at what company you will be placed.

class level - output
कितने हैं जिनके 65 से upper हैं %. loop लगाया
check infosys में कितने placed हैं

* If we don't have class level then it comes in unsupervised learning

---

*
yes (10%, 7CS)
yes (12%, 7CS)
yes (Btech%, 7CS)
(अमुक)

internal node - condition होती है
branches - outcome
class A/class B - leaf

② Neural (numeric data में)
hidden layer there, weight हैं

④ * Cluster analysis - when no class label, we need to find pattern, we use this
→ we apply when no cla on basis of similarities, dissimilarity we make clusters

* K mean algo → जितने clusters बनाने (we select 3 dots | representative. then we calculate Euclidean distance of this point from all the points (n-1). जो close होगा उसे 1 boundary में ले लिया) calculated mean value then make cluster head again calculated distance) now, cluster will change.

• 1 class के अंदर minimize inter class
maximize similarity in 1 class

• outside cluster | elements are outlier || fraud detection |. unauthorized access

Scanned by TapScanner

in all cities use full cream milk, all juice shop near hospital for all city → substructure.
include frequent set & subsequence

→ support & confidence factor in frequent set.
out of 100, 50 milk & bread buy की, then
~~confidence~~
support = 50%. (actual में ho gya | जितना buy कर लिए)
confidence = 50% (जो होगा next day)

- उस association rule को वहां पहुँचने के लिए min कितना observe करना — min threshold

③ classification – (supervised)
ML train then we predict [class level पता है / जो pattern पहले बार आ रहे

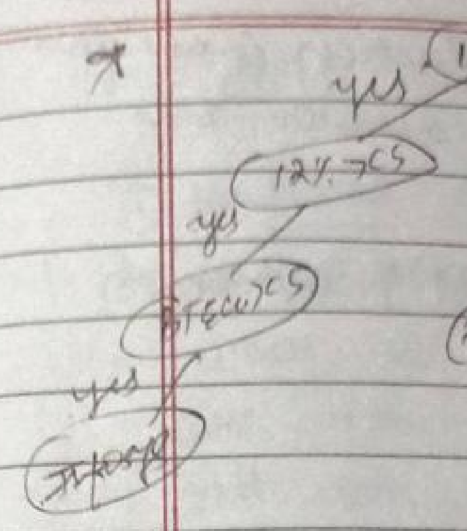
decision tree / Neural, Bayesian classifier

clustering – unsupervised.

After 5 attributes 10, 12, Btech %, communication skill, technical, placement → we train model on basis of these attributes then we predict at what company you will be placed.

class level – output
कितने हैं जिनके 65 से upper है % 100p लगाया check infosys में कितने placed हैं

* If we don't have class level then it comes in unsupervised learning.



internal node – condition लगाते हैं
branches – outcome
class A / class B – label

② Neural (Numeric data में)
hidden layer there, weight है

(4) * Cluster analysis – when no class label, we need to find pattern, we use this.
→ we apply when no cla on basis of similarities, dissimilarity we make clusters

* Kmean algo → जितने clusters बनाने (we select 3 dots) representative. then we calculate Euclidean distance of this point from all the points (n-1). जो close होगा उसे 1 boundary में ले लिया | calculated mean value then make cluster head again calculated distance) now, cluster will change.

- 1 class के अंदर minimize inter class maximize similarity in 1 class

- outside cluster elements are outlier || fraud detection, unauthorized access

* If numerical value - prediction (numerical data)
categorical (alphabet) - decision tree

* outlier → object that does not comply with general behaviour of data

* Sequential pattern mining -

*

* Data Warehouse Usage

Three data warehouse applications

① Information processing
② Analytical processing.
③ Data mining

* From OLAP to OLAM

① High quality of data in data warehouses
② Available info processing
Whenever we have to fetch data from warehouse, we use OLEDB.

* OLAP server architectures -

① Relational OLAP - (ROLAP) - close to relational DBMS
② Multidimensional OLAP (MOLAP) - storing multi dimensional array. If we aggregate data then many models are there. Multiple abstraction levels are there, all data stored at different abstraction level.
③ Hybrid → combo of ROLAP & MOLAP
④ specialised SQL servers - tool depend upon company, company अपने domain के data के tool provide करती)

• We ~~were~~ using Data warehouse वहाँ से data fetch करने के लिए → ROLAP

• Specialisation - I.T. में different level में store करते हैं ex UG → BTech → CSE, Civil.

• Relational database → multidimen

P-90

Unit 2

Data Preprocessing

① Data cleaning — 1 Numerical
② Data integration — 2 Numerical
③ Data reduction (data set is very large, hold that data which is helpful)
④ Data transformation — 2,3 numerical
↳ normalize, concept hierarchy.

we use ETL tool to implement all ④ part.
↓
transform data

Informatica, multiple companies provide ETL

ETL → relational data को factual & descriptive में convert.

① fill in missing value, smooth noisy data [binning ... remove outlier & resolve inconsistencies.

① whenever dataset is combined from sources, data is faulty.
- incomplete
- noise → error at data में
- inconsistent → rating 1,2,3 now A, B, C
- intentional → as everyone's data.

- Data is not available always.

* How to handle missing data

① ~~Ignore tuple where data mission.~~

② Fill missing value manually.
③ Fill it automatically with
→ global constant (not proper)
→ attribute mean (total mean लिया → overall % निकाले सबको)
→ attribute mean for all samples belonging to same class. Eg. जिसका 12th में grade A है उन सबको। साथ डाल दिया avg निकाला उस class में fill कर दिया

④ most probable value → कोई भी classification model

- Noisy data
Noise - random error or variance in measured variable.
दो table को ... में attribute different है तो data ... को

→ To handle noisy data
① Binning
② Regression
③ Clustering
④ Combined computer & human inspection

★ Data Transformation Strategies →
We transform data from / form to other, transform transactional data → factorise & dimensions data (value to coefficient)

① Smoothing → Techniques binning, regression, clustering. (value to cluster)
remove noise from data

→ clustering → datapoint that are not in particular cluster is noise & are removed. (median point.
Ex → boys hostel, girls, localities, pg. are of different clusters. Those who are out of clusters are outliers.
• 1 cluster - whole feature of that datapoint
यह pattern की value store करके रखते हैं
यह not individual.

② attribute construction → new attributes are constructed, added from given set of attributes
Ex) sale attribute

③ aggregation - In datawarehouse, sale is divided according to location, annually.
sale is stored on basis of time also.
means storing on different abstraction level.

Numerical

④ Normalization - attribute data are scaled so as to fall within smaller range.
→ large range को transform करके smaller range में।
Ex) graph में 1 cm drawn 1cm is nb.

(i) min-max normalization -

Quiz

Ex) attribute A value है 100, 200, 300, 400.
$min_A = 100$
$max_A = 400$
Transform these value in range of -2 - -8
यानि कि 100 - 400
new $min_A = -2$, new $max_A = -8$
$v_i' = -2 - -8$
$v_i = 100 - 400$

Soln

Put these value in formula $v_1 = 100, v_2 = 200, v_3 = 300, v_4 = 400$

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

$$v_1' = \frac{100 - 100}{400 - 100} (-8 - -2) + -2$$

$$v_1' = 0 + -2 = 0.2$$

$$v_2' = \frac{200 - 100}{400 - 100} (-8 - -2) + -2$$
$$= \frac{100}{300} \times 0.6 + -2$$
$$= 0.4$$

V_3', V_4' calculate

$V_3' = \frac{300-100}{300}(0.6) + 0.2$

$= \frac{200}{300} \times 0.2 + 0.2$

0.6

$V_4' = \frac{400-100}{300}(0.6) + 0.2$

$\boxed{V_4' = 0.8}$

Ans)

(ii) Z-score normalisation → when large dataset is there than min, max value maybe outliers then we use this.

Zero-mean normalization

we use mean value, standard duration.

mean of $\bar{A} = \frac{100+200+300+400}{4} = 250$

$\sigma_A = \sqrt{\frac{1}{n}\left((100-250)^2 + (200-250)^2 + (300-250)^2 + (400-250)^2\right)}$

$= \sqrt{\quad}$

$\boxed{V_1' = \frac{V_1 - \bar{A}}{\sigma_A}}$

$V_1' = \frac{100-250}{\sigma_A}$, $V_2' = \frac{200-250}{\sigma_A}$

$V_3' = \frac{300-250}{\sigma_A}$, $V_4' = \frac{400-250}{\sigma_A}$

Book → Data Mining concepts & Techniques

Page No.
Date:

* Data reduction → data size को reduce.
transactional database ⟶ warehouse (data reduce) + only important data है there

(i) dimensionality reduction -
1 value (cell) linked with 3 dimension
If you need only 1 dimension rest have to be discarded i.e. dimensionality reduction.

↳ obtain reduced data smaller in volume hence maintain integrity of dataset.

↳ process of removing attributes.
whenever you plot graph, you reduce 3 dim to 3dm
Major pattern, coefficient is same but data is reduced

* same all members in dataset are not responsible for pattern generation, we took those only which is reduce

(ii) Numerosity reduction → replace original data vol^m by smaller forms of data / represent

↳ parametric
↳ non-metric → histogram, clust

↳ वो format में data को represent करते हैं

↳ parameters store करते हैं not data
ex → linear regression m store करते हैं $\hat{E}$ not x, y.

↳ storing data but in reduced form.

(iii) Data compression → reduced or compressed representation of data.
↳ lossless → original data from compressed data
↳ lossy

* Dimensionality Reduction — जो component use नहीं है उनको reduce करते

1) Wavelet Transform —

Discrete wavelet transform (DWT) —

we have data vector X, we converted it into X', of wavelet coefficient. For this multiply data vector with wavelet coefficient i.e. $X = (x_1, x_2 \ldots x_n)$

→ higher degree term को रखेंगे | smaller coefficient remove. ex $n^2 + 2n + 1 \Rightarrow n^2$ (99% pattern can be accessed.

* In ML, packages are there name Haar 2, Daubechies 4, Daubechies 6 for reduction of dimension.

Algo

step 1) Length L, of input data vector must be integer power of 2, this is met by padding data vector with zeros as $(L \geq n)$

2) Each transform involve 2 function,
1 → data smoothing (sum or avg)
2 → weighted difference which find detailed features of data

as we do in aptitude series ques (find pattern)

→ length get reduced by 2.

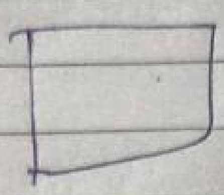
3) functions are applied to pairs of data point and
1 side → low-frequency version of I/p data
2 side → High frequency.

* If data in matrix form, then all data must be in orthonormal (90°)

⇒ highest coefficient से पता लगेगा |

→ used for higher dimension.

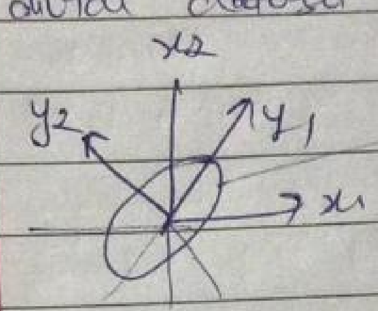
*

Ex) 

matrix find sum 1 & 2 value, 3 & 4 value form another matrix.
find difference 1 & 2, 3 & 4 another matrix.
In this way 1 single value form, pattern

(K-L method.

2) PCA (Principal component analysis) — converting data vector divide dataset to orthogonal form



pattern along with y_1 & y_2 compo spread हो रहा | reduce x_1 & x_2 as इसके basis spread होता

i) searches $k<n$ vectors that can best use to represent data.
suppose 4 vectors there, then reduce to best needed i.e. 2 vectors.

step i) I/p data are normalized, so that each attribute falls within same range.

(ii) compute

→ used for lower dimension.