SJCIT CS 035

Christy Varghese

28/06/2021
Monday.

DATA MINING AND WAREHOUSING
CS 402.

(2) (b) $A = \{116, 234, 486, 544\}$.

min - max normalization

new_min$_A$ = 0

new_max$_A$ = 1

min$_A$ = 116

max$_A$ = 544.

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

$$V_{116} = \frac{116 - 116}{544 - 116}(1-0)+0 = \underline{\underline{0}}$$

$$V_{234} = \frac{234 - 116}{544 - 116}(1-0)+0 = \frac{118}{428} = \underline{\underline{0.276}}$$

$$V_{486} = \frac{486 - 116}{544 - 116} \times 1 = \frac{370}{428} = \underline{\underline{0.8644}}$$

$$V_{544} = \frac{544 - 116}{544 - 116} \times 1 = \underline{\underline{1}}$$

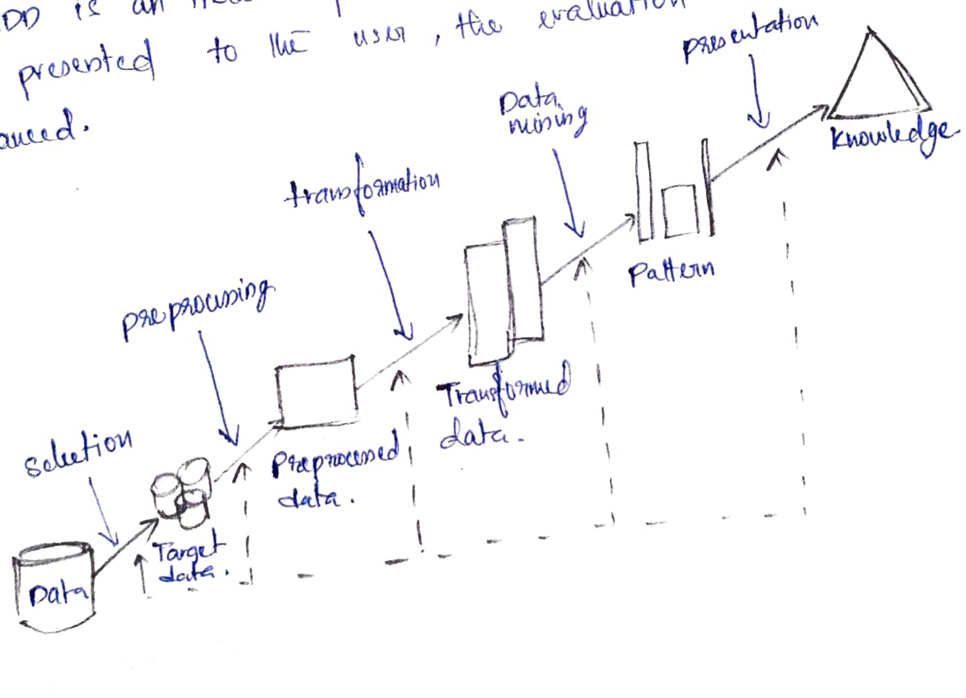
(1) (a) Knowledge Discovery in Databases (KDD).

* Data cleaning also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

christy

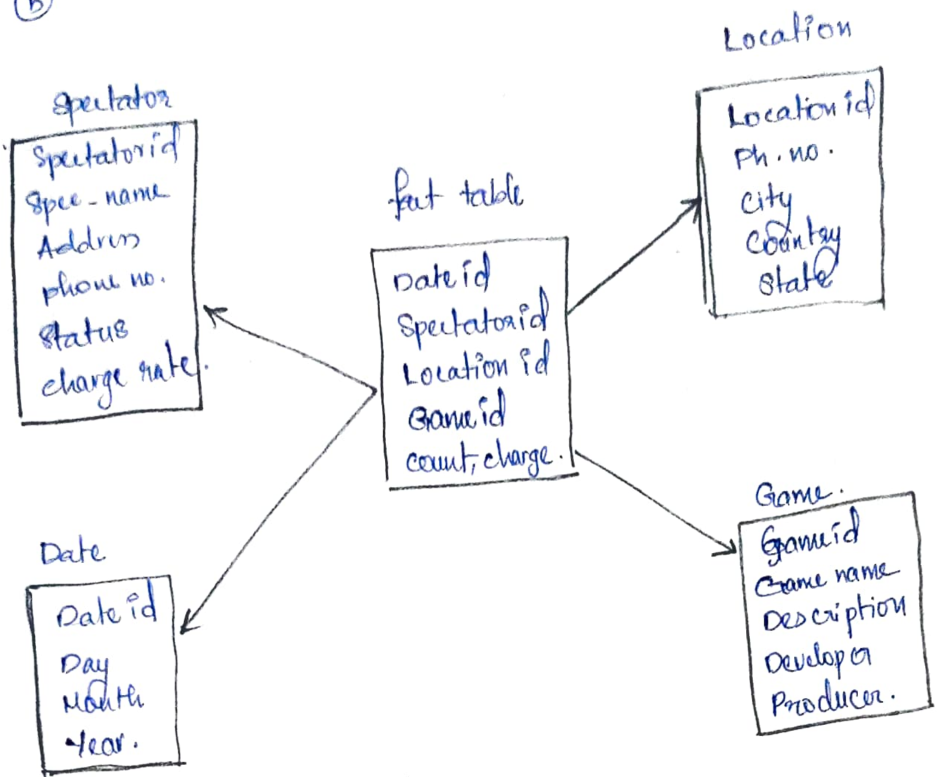
SJC17CS035

* **Data integration** at this stage multiple data sources, often heterogeneous, may be combined in a common source.
* **Data mining** it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
* **Data selection** at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
* **Data transformation** it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
* **Pattern Evaluation** strictly interesting patterns representing knowledge are identified based on given measures.
* **Knowledge representation** is the final phase in which the discovered knowledge is visually represented to the user.
* Data selection & data transformation can also be combined where the consolidation of data is result to the selection.
* KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced.



SJC17CS035

① . ⓑ



Star schema.

② ⓐ Sampling can be used as a data reduction technique because it allows a large data set to be represented by a much smaller random sample. Suppose that a large data set D, contains N tuples.

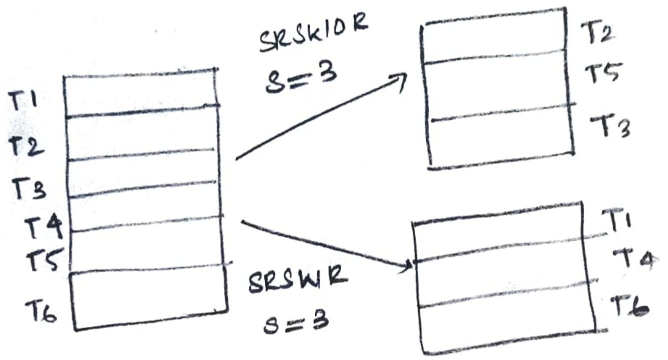
ⓘ Simple random sample without replacement (SRSWOR) of size s: This is created by drawing s of the N tuples from (D≤N) where the probability of drawing any tuple in D is 1/N, i.e; after all tuples are equally likely to be sampled.

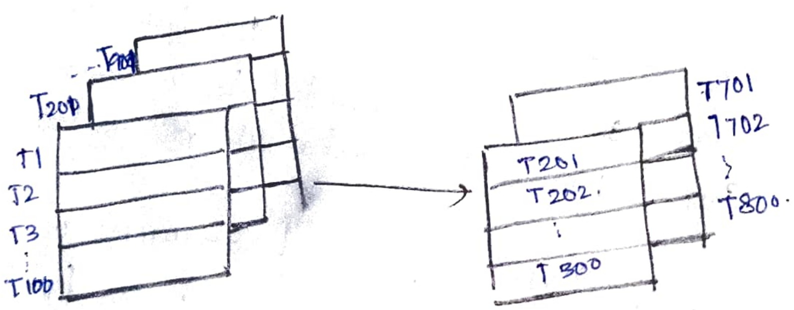
ⓘⓘ Simple random sample with replacement (SRSWR) of size s: This is similar to SRSWOR, except that each time a tuple is drawn from D, it is recorded and then replaced. i.e; after a ~~tu~~ tuple is drawn, it is placed back in D, so that it may be drawn again.

Christy

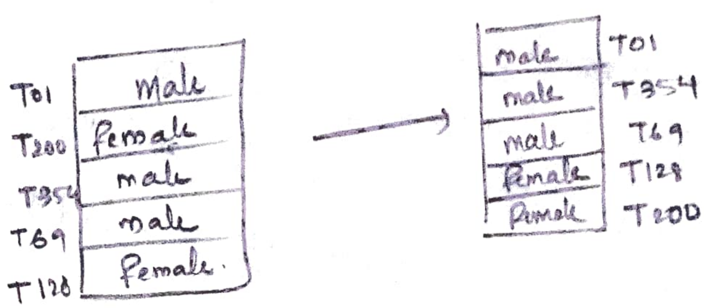
(iii) cluster sample : If the tuples in D are grouped into M mutually disjoint "clusters" then an SRS of s cluster can be obtained, where s ≤ M.

(iv) strafied sample : If D is divided into mutually disjoint parts called strata, a straified sample of D is generated by obtaining an SRS at each stratum. This helps ensure a representative sample, especially when data are skewed.





cluster sample (s=2)



stratified sample according to male, female.

(7)

(b) $A = (22, 1, 42, 10)$ $B = (20, 0, 36, 8)$.

(i) ~~Hamill~~ Euclidean distance.

$$D_E(A,B) = \sqrt{(22-20)^2 + (1)^2 + (42-36)^2 + (10-8)^2}$$
$$= \sqrt{4+1+36+4}$$
$$= \sqrt{45}$$
$$= \underline{6.708}$$

(ii) ~~Ha~~Manhattan distance

$$D_M(A,B) = |22-20| + |1-0| + |42-36| + |10-8|$$
$$= 2+1+6+2$$
$$= \underline{11}.$$

(7) (a) minimum support = 3 confidence = 80%

TID	items bought
T100	{M; O; N, K; E, Y}
T200	{D, O, N, K, E, Y}.
T300	{M A, K, E}
T400	{M, U, C, K, Y}.
T500	{C, O, O, K, I, E}.

$C_1 \rightarrow$

itemset	count		
M	3.	D	1
O	~~4~~ 3.	A	1
N	2	U	1
K	5	C	2
E	4	I	1
Y	3		

[illegible]

Since the minimum support count is 3.

SJC17CS035

L_4 will be ⇒

Itemset	Support count
M	3
O	~~4~~ 3
K	5
E	4
Y	3

cluster 1.

{M, O, K, E, Y}

C_2 ⇒

itemset	support count
(M, O)	1
(M, K)	3
(M, E)	2
(M, Y)	2
(O, K)	~~4~~ 3
(O, E)	~~4~~ 3
(O, Y)	2
(K, E)	4
(K, Y)	3
(E, Y)	2

L_2 ⇒

itemset	support count
(M, K)	3
(O, K)	~~4~~ 3
(O, E)	~~4~~ 3
(K, E)	4
(K, Y)	3

C_3 ⇒

itemset	support count
(M, O, K)	1
(M, K, E)	2
(M, K, Y)	2
(O, K, E)	3
(O, K, Y)	2

c Christy

$L_3 \Rightarrow$ itemset	support count
(O, K, E)	3.

∴ The frequent itemset (O, K, E) is obtained by Apriori Algorithm.

min. confidence = 80%

consider L3 = (O, K, E)

Association rule formed from (O, K, E).

{O, K} ⇒ E confidence = $\frac{3}{3}$ = 100%

{O, E} ⇒ K confidence = $\frac{3}{3}$ = 100%

{E, K} ⇒ O confidence = $\frac{3}{4}$ = 75%.

∴ ~~strong~~ association ⇒

E ⇒ {O, K} confidence = $\frac{3}{4}$

K ⇒ {O, E} confidence = $\frac{3}{5}$

O ⇒ {K, E}. confidence = $\frac{3}{3}$ = 100%.

∴ strong association ⇒ {O, K} ⇒ E
{O, E} ⇒ K
O ⇒ {K, E}

chetty

SJC17CS035

4

(b) $P(\text{headache} = Y \mid \text{runny nose} = N \mid \text{fever} = Y) = ?$

$P(\text{flu} = Y) = 1/2$
$P(\text{flu} = N) = 1/2$

$P(\text{headache} = Y \mid \text{flu} = Y) = 2/3$
$P(\text{headache} = Y \mid \text{flu} = N) = 1/3.$

$P(\text{runny nose} = N \mid \text{flu} = Y) = 1/3.$
$P(\text{runny nose} = N \mid \text{flu} = N) = 2/3.$
$P(\text{fever} = Y \mid \text{flu} = Y) = 2/3.$
$P(\text{fever} = Y \mid \text{flu} = N) = 1/3.$

$P(\text{headache} = Y \mid \text{runny nose} = N \mid \text{fever} = Y \mid \text{flu} = Y)$

$= 2/3 \times 1/3 \times 2/3 \times 1/2 = \underline{.074}$

$P(\text{headache} = Y \mid \text{runny nose} = N \mid \text{fever} = Y \mid \text{flu} = N)$

$= 1/3 \times 2/3 \times 1/3 \times 1/2 = \underline{0.037}.$

max (0.074, 0.037) = $\underline{0.074}$

So the patient is likely to have flu = Yes.

		P. -ve	+ve
A	-ve	TN	FP
	+ve	FN	TP

~~4 (a)~~

5 (b) $\text{Recall} = \dfrac{TP}{TP + FP}$

$\text{Accuracy} = \dfrac{TP + TN}{TP + TN + FP + FN}.$

Chaity

SJC17CS035

(i) $\text{Accuracy} = \dfrac{100+50}{100+50+5+10}$

	+ve	-ve
+ve	100 TP	5 FN
-ve	10 FP	50 TN

$\text{Accuracy} = \underline{\underline{0.909}}$

(ii) $\text{Recall} = \dfrac{TP}{TP+FN} = \dfrac{100}{\cancel{110}\ 105} = \underline{\underline{0.952}}$

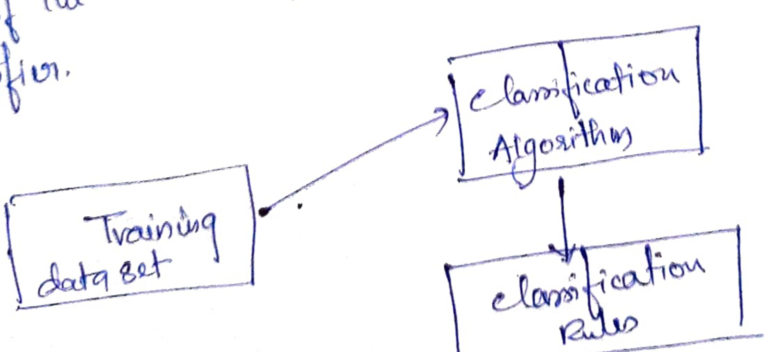
(4) (a) <u>Working of classification</u>

Data classification is a two-step process.

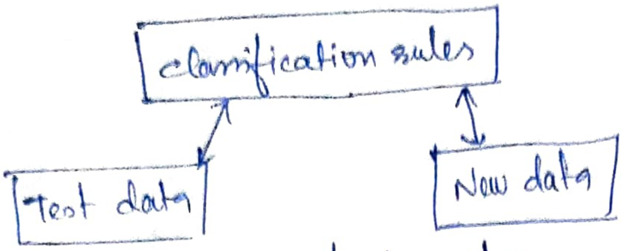
<u>step 1</u>: In the first step, a classifier is built describing a predetermined set of data classes or concept. This is the <u>learning step</u>, where a classification algorithm builds the classifier by analysing or "learning from a training set.

<u>step 2</u>: In the second step the model is used for classification. First, the predictive accuracy of the classifier is estimated. If we were to use the training set to measure the accuracy of the classifier, this estimate would likely be optimistic. It is also known as the <u>classification step</u>

The accuracy of a classifier on a given test set is the % of the test tuple that are correctly classified by the classifier.



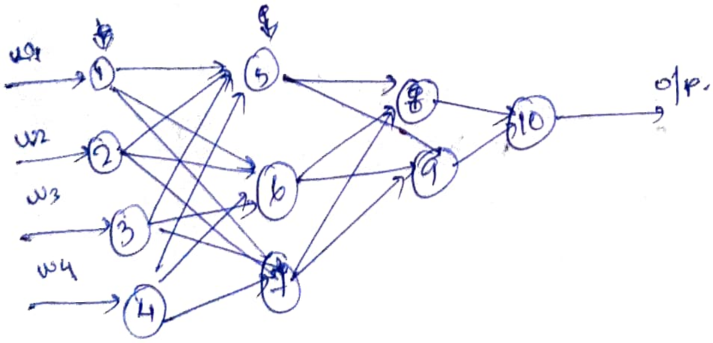
step1: learning step.



step 2: classification step.

5 (a) Backpropagation Algorithm

Backpropogation learns by iteratively processing a data set of training tuples, comparing the network's prediction for each tuple with the actual known target value. The target value may be known class label of training tuple. For each training tuple, the weights are modified so as to minimize the mean squared error between the network's prediction and the actual target value. These modifications are made in "backwards" direction. The ~~algo~~ steps involved are expressed in terms of i/ps, o/ps and errors and may seem awkward if this is your first look at neural network learning.



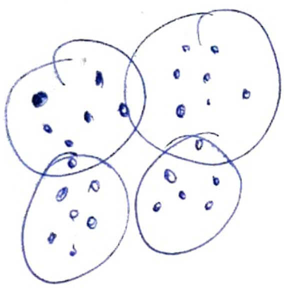
SJC17CS035

(8) (a)

* Agglomerative Method

Agglomerative methods work by starting with singleton sets and then merging them until S is covered. The agglomerative methods cannot be used directly as it scales quadratically with the number of data points. ~~However~~ Hierarchical methods usually generate spherical, ~~cubes~~ clusters and not of arbitrary shapes.

Each objects initially represents a cluster of its own. Integrating hierarchical clustering with other integrating hierarchical clustering with other techniques are BIRCH, CURE etc.

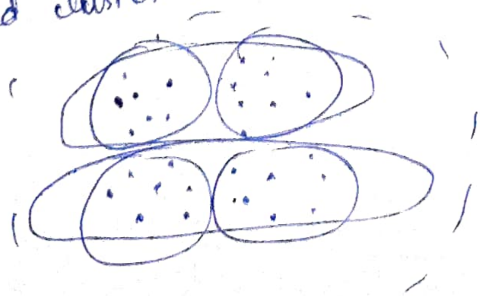


It is a bottom up approach

Diagram.

* Divisive Method : It works by recursively partitioning the set of data points S until singleton sets are obtained.

All objects initially belong to one cluster. Then the cluster is divided into sub-clusters which are successively divided into their own sub-clusters. This process continues until the desired cluster structure is obtained.



It is a top-down approach.

Diagram.

Signature

8 (b) Characteristics of social networks

Social networks are rarely static. Their graph representation evolve as nodes and edges are added or deleted over time. In general, social networks tend to establish the following performance.

(i) Densification power law : It was believed that as a n/w evolves, the number of degree grows linearly in the number of nodes. However, extensive experiments have shown the n/w become increasingly dense over time with the average degree increasing.

(ii) Heavy tailed out degree & indegree distributions :
The no. of out-degrees for a node tends to follow a heavy-tailed distribution by observing the power law, $1/n^a$, $a<2$. The smaller the value of heavier the tail. The in degree also follows a heavy tailed distribution although it tends to more skewed than the out degrees distribution.

(iii) Shrinking diameter : It has been experimentally shown that the effective diameter tends to decrease as the network grows. This contradicts an earlier belief than the diameter slowly increase as a function of n/w size.