7/1/2019

Efficient algorithm { Given T, generate all possible freq itemsets }

Apriori → too many repetitions → bottom up approach.
fp growth → 2 scans.

Last step of Apriori is $C_{i+1} = \phi$.

- Upward closure – if an itemset is infrequent then it's & immediate supersets are infrequent.
- Downward closure – exact opp. of upward closure.

No of scans in Apriori = $|C_2| + |C_3| + \cdots + |C_n|$
Time complexity of apriori –

- How is fp-growth better?
  2nd scan is to generate fp-tree.
- Sequence pattern mining in Apriori → Generalised Sequence pattern mine.

Apriori is adjustable in SPM domain.
fp growth is not suitable for GSP.

fpgrowth was developed by Han + Team (2010).
fpgrowth is not for sequence pattern mining.

- fpgrowth for sequence pattern → Prefix Span.
  SPM used in web patterns. Generic to specific websites.

Yes ← fpgrowth, does it require only 2 overall scans?
fTree is referred as the projection of the database.
Projection is to support efficient counting.

Apriori given by Rakesh Agarwal.

DIC was about parallel counting.

Where do more countings occur in fp tree?
I_1, I_2 might occur in many paths. So the subset merging requires smaller scans (not the entire DB). So it suffers from repeated scans.
Subset merging :- looking at sequences where I_1 is present.

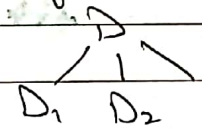
Both Apriori & fp-growth are time complex.
DIC is space complex.
fp growth is a good balance b/n Apriori and DIC.

Pincer Search (Maximal FI); A-close. (Closed FI).

Apriori is good for sparse dataset. That means when there are many infrequent itemsets.

- Can we try improving the efficiently of Apriori?
Apriori is a balance b/n space & time complexity.

- Partitioning approach of Apriori.



if $|D| = 15$. (minsup = 3)
$|D_1| = 5 = |D_2| = |D_3|$.

local mining approach is done. (local Apriori).
To arrive at the global count we need merging of D_1, D_2 & P_3. (Union all three).

4/1/2019

In partitioning, something that is locally infrequent might be globally frequent.
Do union.

Hash pruning approach → in early levels for counting. Ist in the C_2 level.

→ Why not in higher levels?
Occurance of the itemset with higher length will come down.

referred in ECLAT

1, 2, 5
2, 4
2, 3.
1, 2, 4.
1, 3.
2, 3.
1, 3.
1, 2, 3, 5
1, 2, 3

Minsup = 2

if a b c d are there then

element	order
a	1
b	2
c	3
d	4

1 1
2 2
3 3
4 4
5 5

1 2 23 34 45
1 3 24 35
1 4 25
1 5

$L_1 = \{1, 2, 3, 4, 5\}$

$C_2 = L_1 \bowtie L_1$

C_2 → L_2 freq, C_2 infreq

instead of going thru the whole dataset, we look at the hash table. In this way Apriori with hashing is better.

C_2 is for 2 itemsets

$H(x, y) = (\text{order of } x \times 10 + \text{order of } y) \% 7$

Buckets →	0	1	2	3	4	5	6
	(1,4)	(1,5)	(2,3)	(2,4)	(2,5)	(1,2)	(1,3)
Contents →	(3,5)	(1,5)	(2,3)	(2,4)		(1,2)	(1,3)
			(2,3)			(1,2)	(1,3)
						(1,2)	(1,3)
	2	2	4	2	2	4	4

for (1,4)
$H(1,4) = 10 \times 1 + 4$
$= 14 \% 7$
$= 0$

Repeat for all. Collision happens only for infrequent.
C_2 is better.
In case of collision, revise the hash function

- **Transaction Reduction :-**

If L_2 (1,2) is infrequent, then remove the transaction from the Database.

1, 3, 7	(MS=3)	1 (5)
2, 3, 7		2 (7)
1, 2, 3.		3 (9)
2, 3.		4 (3)
2, 3, 4, 5		
2, 3.		$L_1 \neq C_1$
1, 2, 3, 4, 6.		
2, 3, 4, 6.		
1, 3		
1		

If an itemset is infrequent in the earlier levels remove it from the later levels.

5, 7, 6 are infrequent, Remove them from the Dataset.

So now dB is,

1, 3	
2, 3.	
1, 2, 3.	Atleast one character is being
2, 3.	thrown out.
2, 3, 4.	(Time complexity)
2, 3.	The order might remain same.
1, 2, 3, 4.	
2, 3, 4.	Apriori's complexity is exponential.
1, 3	
1	

9/1/2019

Transaction reduction is also level based.
Transaction reduction is not efficient as an algo, but implementation wise it is efficient.

ECLAT → fi's
AClose → cfi
Pincer → mfi

* ECLAT → Equivalence Class Lattice Traversal.
Breadth first
Apriori is BF traversal.
ECLAT → vertical representation of DB → focus on item
Horizontal focuses on Transactions.

Where 'i' item's 'i' transactions diff
Same for everything

$I_1 \rightarrow \{T_1, T_4, T_5, T_7, T_8, T_9\}$
$I_2 \rightarrow \{T_1, T_2, T_3, T_4, T_6, T_8, T_9\}$ MS = 2
$I_3 \rightarrow \{T_3, T_5, T_6, T_7, T_8, T_9\}$
$I_4 \rightarrow \{T_2, T_4\}$
$I_5 \rightarrow \{T_1, T_8\}$

'n' in order (intersection in order)

min sup

✓ $I_1 I_2 = \{T_1, T_4, T_8, T_9\}$.
✓ $I_1 I_3 = \{T_5, T_7, T_8, T_9\}$.
× $I_1 I_4 = \{T_4\}$.
✓ $I_1 I_5 = \{T_1, T_8\}$.
✓ $I_2 I_3 = \{T_3 T_6 T_8 T_9\}$.
✓ $I_2 I_4 = \{T_2 T_4\}$.
✓ $I_2 I_5 = \{T_1 T_8\}$.
× $I_3 I_4 = 0$.
× $I_3 I_5 = \{T_8\}$.
× $I_4 I_5 = 0$

→ C_2 interms of Apriori.

ECLAT is better than Apriori and fp growth on counting.
No need to go to the dataset at all.

→ $I_1 I_2 \cap I_1 I_3$.

→ paper. Bart Goethals. → father of ARM

$I_1 I_2 I_3 = \{T_8, T_9\}$.
$I_1 I_2 I_5 = \{T_1, T_8\}$.

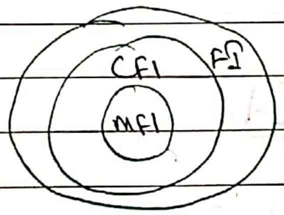
CHARM, MAFIA — Read about them

- What was the need for CFI, MFI and others?
  Rare is not infrequent entirely. Sensitive & high utility itemsets are other types of itemsets.
  Hiding sensitive itemset → Privacy preservation data mining.

14/1/2019

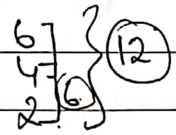
A-close → Apriori-close.

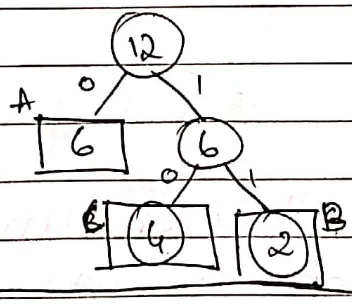
- Why do we need CFI, MFI & LFI?



The application is not happy with so many FI's. So shrink it down.

Closed is from closure. The result should belong to the universe of Discourse. Here FI is the universe of discourse. CFI is lossless, MFI is lossy.
Lossless → from CFI can we reconstruct FI? Yes!.
gzip, pkzip, RLE (Run Length Encoding)
Huffman trace was taught.
~~AAAAAAB~~ AABAAAABCCCC.
A → 6 B → 2 C → 4.
descending order of frequencies.





D → text.

for A → 0 ; B = 11 ; C = 10 → code table

The letter with more frequency gets the ~~more~~ lesser code length.

Biology + Data Mining = Bioinformatics.

Communique
Page No.:
Date: / /

Decoding should be unambiguous. (Prefix free encoding)

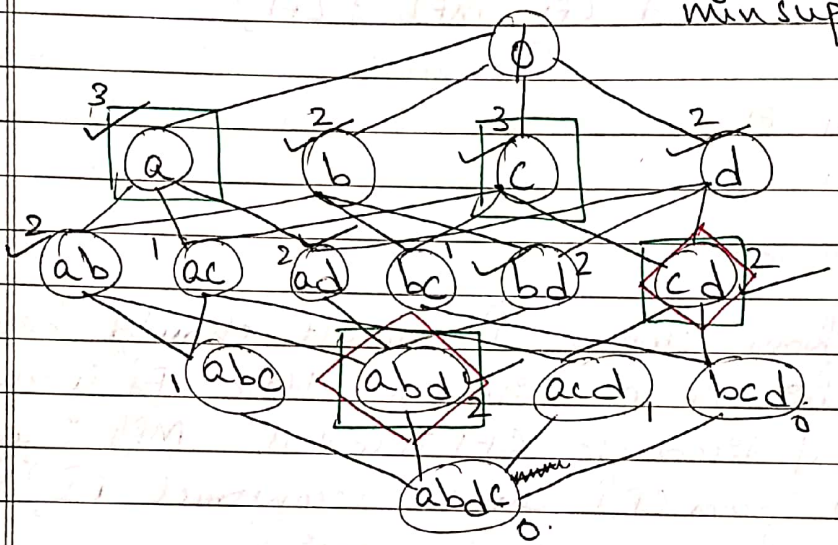
CFI, MFI find applications in compression.

Video and image compression can be lossy.

CFI = A(S) AC(S).

17/1/2019

min support = 2.



✓ → fi's. (9)
□ → cfi's (4)
◇ → mfi (2).

Cfi is lossless, mfi is lossy.

↓

are we getting the entire fi from cfi's.

MFI is more about |FI| (cardinality)
CFI is a tighter constraint on MFI.
MFI also comments on the subsets.
We will not be able to say the subset count from MFI. Confidence calculation won't be possible.

With CFI we can tell the support count of the subsets.

x → if there is no superset of x. with the same support count as x. (or less support count)
→ CFI.

$$\text{superset}(x) < sc(x)$$

x → ∄ any superset of x which is frequent.
→ MFI.

- Difference between closed itemset and closed frequent itemsets.

* A-"close" Algorithm :-

S items (A, C, T, D, W).
→ all the transactions that contain.
t(x) itemset x. $\exists x^3 \subseteq T$

i(y). $y \subseteq T$

1	A C T W.
2	C D W.
3	A C T W.
4	A C D W.
5	A C D T W.
6	C D T.

t(C) = {1, 2, 3, 4, 5, 6}.
i(C) = undefined.
i(T1, T2) = intersection of 1 & 2. = C, W.
i(t(C)) = C.
t(C, D) = {2, 4, 5, 6}.
i(t(C, D)) =
t(A) = {1, 3, 4, 5}
i(t(A)) = {A, C, W}
i(t(A)) ≠ A. unlike the examples before. this. it is equal to a greater set.
i(t(ACW)) = {A, C, W}.
Do, i(t(A, C, W)) = {A, C, W}.
Do t(x) then i(t(x)). If i(t(x)) ≠ x but is equal to y, Do i(t(y)). there it has to stop.
1 level we have elements gaining.

(i) $i(+(x)) = x$.

(ii) $i(+(x)) = y$; $x \subset y$ → elements gained.
$i(+(y)) = y$.

- **Generators :-**
Identify those elements that act as generators.

x is gen(y) if $x^+ = y$.
generator of y ; closure of x yields y.

In the example above, A is a generator of ACW

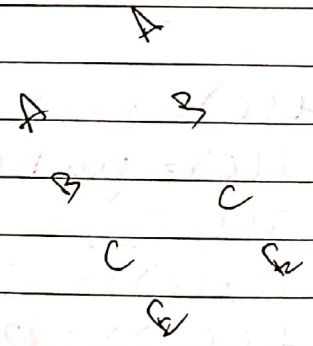
In CFI we are interested about minimal generators.
There is no proper subset of x generates y.

What all generates ACW.
ACW and A.
We cannot have more than 1 minimal generator.

21/1/2019

A-Close Trace.

1	A C D	
2	B E F	MS = 3.
3	A B C E	
4	B E	
5	A B C E	



$C_1 = \{A\ B\ C\ D\ E\}$.
$L_1 = \{A, B, C, E\}$.
$C_2 = \{\underset{2}{AB}, \underset{3}{AC}, \underset{2}{AE}, \underset{3}{BC}, \underset{4}{BE}, \underset{3}{CE}\}$.
(4, ×)

With CFI we reconstruct the FI's but with Closed Itemset we can traceback to the transactions.

$L_2 = \{AC_3, BC_3, BE_4, CF_3\}$.

Is any item having the same support as it's subset. Here AC ⊂ A and BE ⊂ B. Throw them away.

$L_2 = \{BC_3, CF_3\}$.

$C_3 = \emptyset$

Since we left out AC & BE we must prove that AC can be reached from some other minimal set.

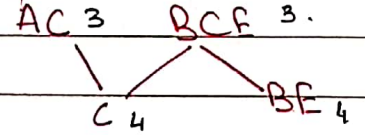
(x) Generators	t(x)	i(t(x))
A	1, 3, 5	AC →
B	2, 3, 4, 5	BE →
C	1, 2, 3, 5	C
E	2, 3, 4, 5	× BE
BC	2, 3, 5	BCE
CF	2, 3, 5	× BCE

We are able to reach them here. ↓

Whatever is left out is able to be reached from i(t(x)). Reaching out in mathematical terms is closure (x^+) $x^+ = y$

Remove the duplicates in order.

CFI's = {AC, C, BCE, BE}.



→ with this as a start pt we should be able to trace back all the FI's.

By not retaining the higher cardinality sets, we are saving space. The cardinality of the Generators should be minimum.

One CFI alone cannot give all the FI's. Go for the CFI with the larger count.

topdown and bottom up.

* Pincer Search. ↑ ↓ "Dataset is given"

support = 20%, relative support = 15.

∴ min sup = 3.

{1, 2, 3, 4, 5, 6, 7, 8, 9} 9 symbols.

Move in both bottom up and topdown.

maximal freq set ← MFS = ∅

maximal freq candidate set. MFCS = {1 2 3 4 5 6 7 8 9}. (Maximal frequent candidate set).

S_i → infrequent at i^{th} level.

$L_1 = \{1, 2, 3, 4, 5, 6, 7, 8\}$.

$S_1 = \{1, 9\}$.

We ~~don't~~ keep record of infrequent to see that they are not part of MFS.

Keep resizing MFCS.

Trimming down MFCS is top down approach.

MFCS and S_i are helping to bring in the top down flavour to the trace.

updating MFCS.

$S_1 = \{\{1\}, \{9\}\}$ MFCS = {1 2 3 4 5 6 7 8 9}.

for {1} in S_1 :- {2, 3, 4, 5, 6, 7, 8, 9} = MFCS

MFCS \ S_1 → minus in set theory.

for {9} in S_1 :- {2, 3, 4, 5, 6, 7, 8} = MFCS.

MFCS \ S_1

$C_2 = L_1 \bowtie L_1$

$L_2 = \{23, 24, 35, 37, 56, 57, 67\}$.

23/1/2019

$S_2 = \{25, 26, 27, 28, 34, 36, 38, 45, 46, 47, 48, 58, 68, 78\}$

Exhaust S_i

smt18@iiitdm.ac.in

2 & 5 are infrequent together.

(25)
$S_2 \rightarrow MFCS = \{345678, 234678\}$ — '26' is occurring together only here.

(26)
$S_2 \rightarrow MFCS = \{345678, 34678, 23478\}$ — trim down only if they are occurring together

34678 is already in

So do not consider '34678'

∴ updated $MFCS = \{23478, 345678\}$

(27)
$S_2 \rightarrow MFCS = \{3478, 2348, 345678\}$

3478 is a subset of 345678

$MFCS = \{2348, 345678\}$

(28)
$S_2 \rightarrow MFCS = \{348, 234, 345678\}$

348 is a subset of 345678

$MFCS = \{234, 345678\}$

MFS is a lossy reconstruction of FS. We do not have the sup count. So, we cannot have the confidence

(34)
$S_2 \rightarrow MFCS = \{23, 24, 45678, 35678\}$

Continue for the other S_2's

(78)
$(S_2) \rightarrow MFCS = \{23, 24, 357, 5678, 8\}$

$C_3 = \{234, 357, 567\}$

Top down will stop if S_i is nullset. There is no more trimming down of MFCS.

Here S_3 is a null set.

LAB-2.

Implement Aclose and Pincer Search.
Two more algos MAFIA & CHARM.

24/1/2019

- Difference b/n BigData and Data Mining
~~How well my~~ how well the storage systems can handle the distributed data. Incremental learning (with added data, the system must be intelligent enough to relearn the patterns). Storage issues according to scaling up of data must be addressed. Scaling up is done across networks.

Online DB ≡ Incremental learning.

Bucket Brigade → Genetic Algorithms
↓
evolutionary computation
↓
evolve for the good.

GA is one evolutionary computation strategy.
Tabu search, PSO, Ant colony are others.
↓
particle swarm optimisation.

Reinforcement learning thrives on penality & awards.
Most evolutionary computation algos work on Initial population.
All the evolutionary computation methods are basically for optimisation (on the computers front).

Bothers for optimum after seeing if the solution is feasible or not.

David Goldberg → GA book [1-3]
James Holland → inventor of GA.

- Enumerative search → point by point search. Also called try and error. (but it is random).
- Operations Research → mathematical optimization methods.
- Derivative based → Gradient ascent or descent. Hill climbing is also another word. A slope should always exist. But real life data is not always continuous. Also, the data can be multimodal, the hill climbing gets stuck in local optimum.

In the search for optimal value, the path towards optima is also important.

Where will GA fit in? → Classification.
If we are not able to give unique class labels in D-tree, it is non feasible.

Convergence → how quickly will I reach the leaf node.

Efficiency vs efficacy.

28/1/2019

- Why EVOLUTIONARY?
Can my overall objective keep bettering. GA, Tabu Search etc are search algos that search for optimal.

Fitness of a solution :- how good is a solution.

iteration ≡ Generation.
As we do more generations, the fitness should improve.

Darwin's theory of evolution forms the crux of the GA's.
GA - search algo based on natural selection + Genetics.

The power of ~~th~~ evolutionary algos comes in if the search base is really large and the function is too ~~to~~ complex.

- Objective function :- it is to be identified. This is the first step.
- Initial population :- random collection of solutions

(2, 6, 8) → 3 random search points from search space.
↳ "Roll of a die"
Can be one approach of constructing the initial population.

Using the initial population can we do some Genetics (biological opns).

$$f(x) = x^2.$$
$$[0, 31].$$

initial popn = 2, 6, 8 → max $x^2 = 64$.
using genetics = 5, 9, 11 → max $x^2 = 121$.
we have to reach (31). 31^2.

Human "flair" of a flexible search procedure.
Biological system solutions are far more robust than the existing search procedures.

So, can the innovative search brought into GA?

GA
- Selection
- Crossover
- mutation.

know atleast 1 way of each.

Application specific operators have improved GA over time.

Search space = 10k.
Initial popn = 20 pts. (random) — larger the search space, larger can be the initial popn
Selection opn :- out of 20 pick only those which are fitter ones.

Allele → attributes → Chromosomes.
Initial popn :- is a collection of chromosomes.
In the probability of fitter solution. Cross parents who have fitter characteristics.

After selection of the fit ones, cross over.
Mutation is hardly recommended. It is done when stuck in the local optima.

Roulette wheel → Simple selection (random)
↳ minimizes the owner's loss.

Initial popn = 13, 24, 8, 19. $f(x) = x^2$

169 576 64 361.
1 2 3 4 ← Chromosomes.
better fit

Sum of fitness = 1170. (wheel dimension) (sample space)

Probability of selection = $f_i / \Sigma f_i$.

$= \frac{13}{1170} \quad \frac{24}{1170} \quad \frac{8}{1170} \quad \frac{19}{1170}$ → Sum is 1
1 2 3 4.

$\frac{13}{1170}$ → percent of the wheel is occupied by 13 on the wheel.

0.14	0.49	0.05	0.30
1	2	3	4
14%	49%	5%	30%

Assume that the population size is 4 (fixed).
So cross over a population of size 4.
We don't want any copy of 8!

29/1/2019

- How does GA gell well with classification?
Information gain using entropy, gini index etc. Choosing the optimal path in Dec. Tree. There might more than one DecTree for one classification. We must be able to choose the optimal path from so many trees.

Expert system forms the base for Bucket Brigade classifier.
Expert System :- control is given by domain experts
ELIZA, ALEXA.
Expert Systems = Training model.
Expert systems are basically classifiers. Paths in DecTree is rules in expert systems.

The fitness of the path is important.
GA strikes the balance between efficiency and efficacy.

Efficacy + effective → closer to feasibility.
The solution must be ROBUST too.

Select from existing → generate new fitter points.

(Select → crossover → mutate) → cumulative in nature

Efficacy in searching is Relevance. It is qualitative measure.

One solution is a chromosome in GA terminology.

Bucket Brigade is a multiclassifier approach.

Look at the past fit solutions. Discourage unfit solutions, encourage fit solutions.

Index	x.	Chromosome (16 8 4 2 1)	f(x) (fitness)
1	13	0 1 1 0 1	169
2	24	1 1 0 0 0	576
3	8	0 1 0 0 0	64.
4	19.	1 0 0 1 1	361

$\max(x^2)$ $[0, 31]$.

Result of 20 coin tosses. 1170 → population fitness.

(Pp = 4)

1st frame the FP.

x can be precision, recall, Fiscore, anything

$\bar{f}$ = 293 → avg fitness of popn

Keep track of current maximum → here 576 (flip of the wheel)

↑ prob of selection on Roulette wheel.

Index	$f_i / \Sigma f_i$		Expected Count ($f_i / \bar{f}$)	Round
1	.14	draw a wheel.	.58.	1
2	.49.		1.97	2
3	.06		0.22	0
4	.31		1.23.	1

Flipping the wheel is selection.

If Round = 1 then we got it once.

2 then we got it 2ce. (2 copies)

Next generation is the output of selection.

Population size is the same.

01101

11000

11000

10011

The essence of selection is duplicate fitter ones, Throw away the unfit ones.

1 (single point) crossover. Identify the crossover site.

crossover site

01101 Completely random crossover
11000 is done.
11000 Identify a crossover site at
10011 random

01100
11001
11011
10000

1-1 crossover site possible.
Portion after crossover site gets swapped.

Jumped from	Index	x	Chromosome	f(x)	$\frac{f(x)}{\sum f}$	$\frac{f(x)}{\bar{f}}$	Round
	1	12	01100	144			
24 to	2	25	11001	625			
27	3	(27)	11011	729			
	4	16	10000	256			
				1754			

$\bar{f} = 438.5$

You will get the fitter solutions no matter where you cross and whomever you cross.

31/1/2019

Bigdata is not only about the volume.

Collection of databases or datawarehouses. How is the data organised? → Bigdata.

Structured data contains a schema.
Unstructured, Semi structured, Quasi-Structured.
location. weblinks.
Bigdata deals with all the above types of data.

Because of IOT, the amount data generated is increasing.
We have ~~te~~ petabytes of data.

The V's.

Volume → how huge the data is
Velocity → streaming data (weather, tweets)
Variety → structured, unstructured, etc
Business insights:
- Veracity → business.
- Value → historic data's importance.
- Variability → error in data, how can it be avoided

* Challenges of Bigdata.

Capturing
Curation
Storage
:

Hadoop is a framework.
Scale up and scale down (out)
SPOF → single point of failure.
how many machines can be added.
Architectures:- MapReduce, HDFS.

4/2/2019

Recap

- What is the main aim of Roulette wheel?
- Disadvantages of Roulette wheel:-
  - → Monopoly : fitter solutions occupy major part of the wheel.
  - → we may sometimes loose the optimum solution.

$\frac{1}{3}$ of the population should be increasing to get the single solution.
GA leads to the path of convergence on every iteration.

Based on increase in avg, we will eventually come up with better results. - Mathematical function application.

Schemata is a pattern of chromosomes.

1 0 0 0
1 0 0 1
1 1 0 0.

The pattern in the above data is MSBs are 1.
Pattern is 1***
where * can be either 0 or 1.

Need for schemata - we can reach the optimum solution faster.

Choosing the pattern after two or three results we need to choose the best pattern.

Fundamental Theorem of Schemata :- (single pattern)

Choosing the better pattern may result to the optimum solution. - Mathematical proof.

Fitness of schema is average of population.

We are concerned about the increase in fitness from previous generation. (Betterment).

Schema → H

Length of the schema $\delta(H)$:- Difference b/n 1st fixed and the last fixed bits of schema.

Example :-

$H_1 = *1****0$ (positions 2 and 7)

$\delta(H_1) = 5$

$H_2 = ***10**$

$\delta(H_2) = 1$

Selection :- Selecting the pattern based on schema pattern.

Crossover is a better solution than Mutation.

GA is more probabilistic because of randomness.

Schema are fundamental theorem of GA.

→ Short defined length $\delta(H)$

→ low $O(H)$

→ Above Average.

Selection. Schema Theorem :-

$$m(H, t+1) \geq m(H, T) \frac{f(H)}{f} \left\{ 1 - \frac{P_c \delta(H)}{L-1} - P_m O(H) \right\}$$

(survival)

$m(H, t+1)$ → count function.

$f(H)$ → fitness of schema i.e. Avg.

f → popln.

$\frac{P_c \delta(H)}{L-1}$ → Destruction probability of crossover.

$P_m O(H)$ → Destruction of mutation

Schema :-

$H_1 = 1*****$

$H_2 = *10***$

$H_3 = 1***0$

Initial population :-

13 : 0 1 1 0 1 → String 1
24 : 1 1 0 0 0 → String 2
8 : 0 1 0 0 0 → String 3
19 : 1 0 0 1 1 → String 4

H_1 schema is the string representation of 2, 4.
H_2 schema is the string representation of 2, 3.
H_3 schema is the string representation of 2.

Why are we keen on string representations.

$$M(H_1, t+1) = \frac{2 \times 668.5}{293}$$

$$= 3.19 \{1 - 0\}.$$
$$= 3.19.$$

for H_2

$$M(H_2, t+1) \rightarrow \frac{2 \times 320}{293} = 2.18.$$

∴ There are 2 copies of H_2 in pop.

$$2.18 * \left\{1 - \frac{1}{4}\right\}.$$

$$= 1.63.$$

Communique
Page No.:
Date: / /

Benefit of schemabased GA:-
- It is the parallel processing capability of GA.
Schema based processing is parallel.
∴ processing is fast.

Main objective of schema theorem:-
In selection we are concerned with betterment

fixed positions survival is exponentially increasing.
are near:

fixed positions exponential decay or distinctness
are apart:

11/2/2019

GA classifier : entire concept is ~~bored~~ bucket brigade

efficient path which leads to the o/p faster is
done by GA.

Rule + Msg system - in the domain of expert system
Bidding results in
① Penalty.
② Reward.

Credit Apportionment System
good classifier = +ve credit.
bad classifier = -ve credit.

What is encoding in GA context?
Soln:- It is used to extract patterns like MSBs are
1 etc.
Using Decimal numbers extracting patterns becomes
difficult.

In travelling salesman problem.
The feasible solution is cost of visiting each node.

Each solution is called chromosome in GA.

1 2 3 4 5 6 } swap.
6 3 4 5 1 2.

Not feasible

Fitness of classifier means?
Strength of classifier which gives better fitness.

Five alarm example

In GA O(n3) schema is effectively useful where n is the population

GA is inherently parallel.

Among the different paths, the poor paths are not neglected straight away but they are exponentially decreasing in GA.

When m classlabels are there, it is less time consuming than binary class label because getting patterns becomes difficult in binary tree.

18/2/2019

3 aspects of Bucket Brigade :-

1. Rule - Msg system.
2. Apportionment of credit
3. Genetic Algo.

Rule :- <classifier> = <condn> : : <message>

<condn> → attributes → $\{0, 1, \#\}^*$

<message> → class labels → $\{0, 1\}^*$

Fitness value = strength.

Increasing or decreasing the strength of classifier is Apportionment of credit.

Table to simulate credit sharing.

~~Rule no. Classifier Str. Msg. Mtch Bid S Msg Bid. S Msg M B~~

Sn	Classifier	t=0				t=1				t=2				t=3			
		S	Msg	Mtch	Bid	S	Msg	Mtch	Bid	S	Msg	Mtch	Bid	S	Msg	M	Bid
1	01## : 0000	200		E	20	180	0000			220				220			
2	00#0 : 1100	200				200		1	20	180	1100			218			
3	11#1 : 1000	200				200				200		2	20	180	1000		
4	##00 : 0001	200				200		1	20	180	0001	2	18	162	0001	3	16.2

Env = 20.

Bid = 10%.

E = 0111 → environment variable

First strength value is random.

Action or Payoff scheme → submit bits.

One classifier reduces it's strength so that the others can get triggered in the subsequent generations.

The Bid is added onto the other classifiers. The same path should not keep getting the priority in each generation.

GA kicks in after one run.

New classifier paths are not explored by bucket brigade.

Bucket Brigade is not a contribution of GA.

$t = 4$.

S	Msg	Match Bid.
220		
218		
196		
146	0001	

terminal.

Crossover b/w ① & ④.

```
01 | # #
# # | 0 0
```

= 01 0 0.

#.

21/2/2019

TUTORIAL ①

① A decision tree is being error pruned by collapsing of nodes. For the particular node, on the left branch there are 3 training points [5.7, 9.6] and for the right branch there are 4 training point [8.7, 9.8, 10.5, 11]

(i) What were the original responses along left, right branches?

(ii) what is the new response after collapsing?

(2) Say True / False - Reason also.

(a) U can list all fi's & their support if u know the MFI — ~~F~~ (F)

(b) to perform hierarchical clustering, we do not need to know the coordinates of elements, only their mutual distances are enough (F)

(c) If one used cosine measure to measure interesti-ngness of A → B. its value doesn't change if we add a new transaction. that doesn't contain A or B. (T)

$$\frac{P(X \cdot Y)}{\sqrt{P(X) \cdot P(Y)}} \qquad \frac{|X \cdot Y|}{\sqrt{|X| \cdot |Y|}}$$

(3) Given T[A .... J]

A B C D E F G H I J

1 → 1110000111 2 → 1010110111

3 → 1001001111 4 → 0110101100.

5 → 0111101000 6 → 1100111010.

7 → 1100000010. 8 → 1110110010.

9 → 1100000011 10 → 0011110000.

(i) Find all fi's with support ≥ 5. {A B C E I}.

(ii) which in (i) are closed?

(4) Given T(A, B, C, D) = {1110, 1110, 0111, 0110, 0111, 1100, 1100 (10 transactions). 1110, 1110, 0011}

(a) What is the confidence of (A, B) → ∅

(b) Given itemset X = {A, B, C}, List all rules of form Y → Z ; Y ≠ ∅ ; Y ⊆ X, Z = X \ Y. with support ≥ 4 and confidence = 2/3.

(5)

Classes.	Buys_comp = Yes.	Buys_comp = NO.
Buys_comp = Yes.	6954	46.
Buys_comp = NO	412	2588.

Given above, calculate. Precision, Recall, sensitivity, Specificity, Accuracy.

(6) Trace A-close, pincer search for the data labelled X.

X :- T_1 {M, O, N, K, E, Y}.
T_2 {D, O, N, K, E, Y}.
T_3 {M, A, K, E}.
T_4 {M, U, C, Y}.
T_5 {C, O, O, K, I, E}

Support = 2

4/3/2019

Neural network is a composition of layers.

Input variables are nodes in network topology.

Back propagation of errors.
Neural network is supervised learning. It has training and testing.
What is error in neural network?
How much prediction is deviated from expectation

Inputs are fed forward. If the error is greater than error threshold, they are fed back.
"Adjust weights".

Neural network is non linear.
One neuron = node in network.
firing of neuron

If we are not happy with the o/p, the add hidden layers.
Activation function: the o/p from hidden layer.

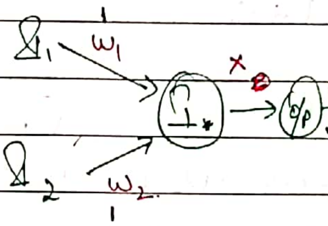
$I_1 w_1 + I_2 w_2 <= 0.5$

$<= 0.5 \Rightarrow 0$
$>= 0.5 \Rightarrow 1$

OR Gate.

I_1	I_2	Output	x	O		or output
0	0	0	0	0.5	$\frac{1}{1+e^{-x}}$ →	0
0	1	1	1	0.7	→	1
1	0	1	1	0.7	→	1
1	1	1	2	0.88	→	1

OR is non linear combination of inputs due to activation function.



Depending on the weights, the inputs are taken

$x = I_1 w_1 + I_2 w_2$

if $w_1 = w_2 = 1$.

then $x = I_1 + I_2$.

ACTIVATION FUNCTION.

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

$x - T$ → error threshold.

The quality of NN depends also on the activation function.

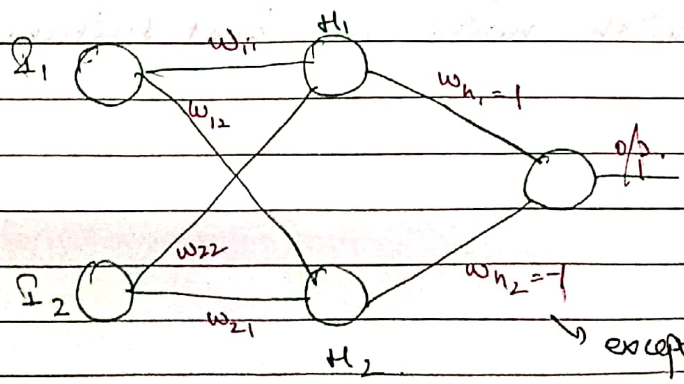
In this network activation function is applied in the output layer.

Where does exclusiveness come in XOR?

XOR needs atleast one hidden layer.

11/3/2019

For XOR ① Hidden Layer ② I/p neuron.



except for this rest of the weights are 1

overall input to the hidden layer H_1

$$\text{Sig } H_1(x) = \frac{1}{1 + e^{-(x - 0.5)}}$$

These numbers come by trial and error.

$$\text{Sig } H_2(x) = \frac{1}{1 + e^{-(x - 1.5)}}$$

$$\text{Sig o/p}(x) = \frac{1}{1 + e^{-(x - 0.2)}}$$

put these in

H_1, H_2 → 0.1951, 0.2450, 0.19

4.

I_1	I_2	x	H_1	H_2	O	out.
0	0	0	.3775	.1824	.4998	0
0	1	1	.6225	.3775	.5712	1
1	0	1	.6225	.3775	.5712	1
1	1	2	.8176	.6225	.4998	0

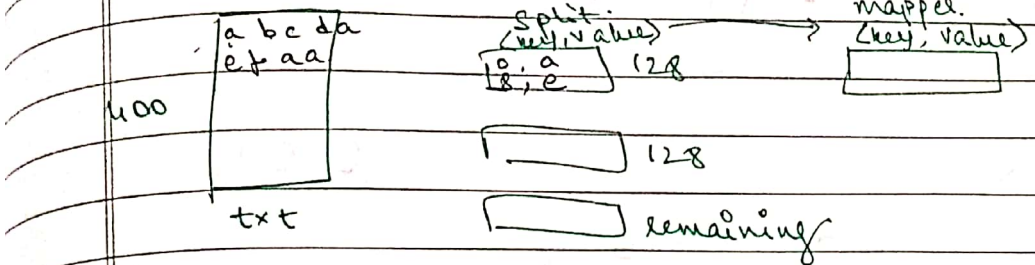
→ <0.5 then 0, >0.5 then 1

You can't have a neural network learning correctly without a hidden layer.

Neural network → supervised.

- Disadvantage:-
  For live inputs there are too many wt adjustments. That means long training times. Interpretability is less than the previous algos.
- ~~Advantages~~
  Non linearity of model.
  Face recognition, etc where human like learning is required. in terms of precision, recall, etc.
  Domains with noise are best suitable for NN.

Aim of Hadoop is parallelism.



Before going to mapper, it will go to split.
Done in terms of lines.

<key, value>
↓ ↓
pos string.

Each mapper works on a single block.

mapper.
<a, 1>
<b, 1>
<c, 1>.
<d, 1>
<a, 1>

18/03/2018.

Back propagation possible only if error is there

Target o/p & Predicted o/p.

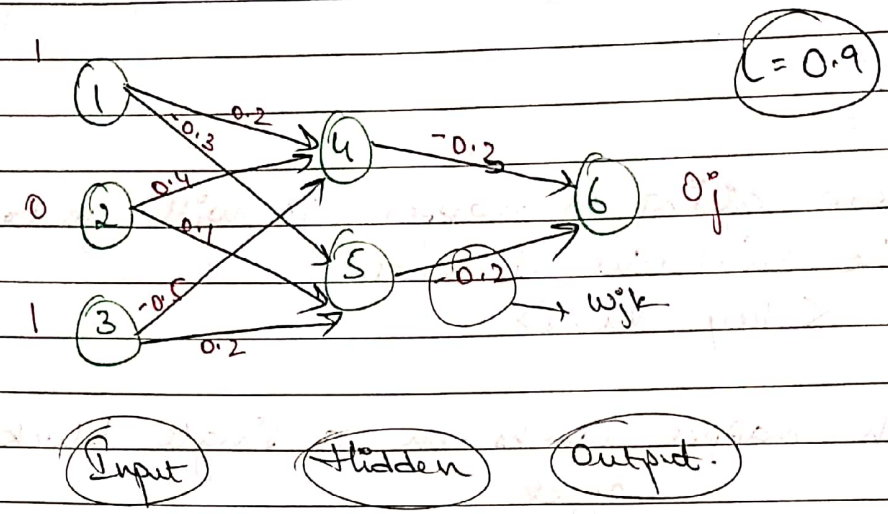
Error = $T_j - O_j$ > Accepted Error Threshold. then back prop.

possible with training Data.

Increase or decrease weights depends on the updated error.

$\langle a_0, a_1, a_2 \rangle$ → binary (1/0) in nature.

$\langle 1\ 0\ 1 \rangle = 1$, T_j



Network topology ≡ design of network

BPN was deemed fit for non linearly separable data.

SVM is for linearly separable.

There is no loop in the network. So, back propagation happens as weights update.

θ - Bias → (O/p nodes)

l - learning rate → (how quickly O_j converges to T_j).

Here $l = 1$

(t) → at any given point of time, the no of iterations done is 't'.

$$I_j = \sum_i w_{ij} O_i + \theta_j$$

i → input

j → output.

$$O_j = \frac{1}{1 + e^{-I_j}}$$

$$Err(O/P) = O_j(1 - O_j)(T_j - O_j)$$

→ comparision with overall output and target output

$$Err(H) = O_j(1-O_j) \sum_k Err_k w_{jk}$$

this is where back propagation happens.

$\theta_4 = -0.4$.
$\theta_5 = 0.2$
$\theta_6 = 0.1$
} also gets updated.

$0.2 + 0 - 0.5 - 0.4$.

Unit	I/p	O/p $(1/1-e^{-I})$
4	-0.7	0.332.
5	0.1	0.525.
6.	$-0.3(0.332)$. $-0.2(0.525)$. $+0.1$. $= -0.105$	0.474

Similar

~~% error threshold.~~

ERROR.

$6. = .474(1-.474)(1-.474)$.
$= 0.1311$

$5 = 0.525(1-0.525) \times (-0.2 \times 0.1311)$.
$= -0.0065$.

Similarly

$4 = -0.0087$.
$0.332(1-0.332) \times (-0.3 \times 0.1311)$

output of 'i' scaled by error of 'j'.

New w_{ij} = Old $w_{ij} + l \times Err_j O_i$

New $\theta_j = l \times Err_j$

DEAP.

$w_{46} = -0.3 + 0.9(0.1311 * 0.332)$.

$\theta_6 = 0.1 + 0.9(0.1311)$.

25/3/2019

Distance computation in clustering is good for numerical data.
But what do we do for categorical data?

TUTORIAL (As the population is thin)

26/3/2019

Matrix multiplication in Hadoop.

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \qquad B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$$

Input split

A, 0, 0, 1
A, 0, 0, 2
A, 1, 0, 3.
A, 1, 1, 4.
B, 0, 0, 5
B, 0, 1, 6.
B, 1, 0, 7
B, 1, 1, 8.

Mapper ($<i, (m, j, v)>$)

PHASE-1

i	value.
0	(A, 0, 1).
0	(A, 1, 2)
1	(A, 0, 3)
1	(A, 1, 4)

i	value.
0	(B, 0, 5)
0	(B, 1, 6).
1	(B, 0, 7).
1	(B, 1, 8)

B matrix is transposed for some reason in the mapper. (Vasanti knows!).
To do the multiplication in Row major fashion

Donot shuffle 'i' values.

PHASE - II.

key	value
0	(A, 0, 1)
0	(A, 1, 2)
1	(A, 0, 3)
1	(A, 1, 4).
0	(B, 0, 5).
0	(B, 0, 7).
1	(B, 1, 6)
1	(B, 1, 8).

key of phase-2. corresponds to the output indices.

28/3/2018

New matrices are:-

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}_{2\times4} \quad \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 1 & 2 & 1 \end{bmatrix}_{4\times3} \quad \begin{bmatrix} \quad & \quad & \quad \\ \quad & \quad & \quad \end{bmatrix}_{2\times3}$$

Do the input split.
Then mapper phase - 2

0	A	0	1	0	B 0 1
0	A	1	2	0	B 1 2
0	A	2	3	0	B 2 3.
0	A	3	4.	1	B 0 4
1	A	0	5	1	B 1 5
1	A	1	6	1	B 2 6
1	A	2	7.	2	B 0 7
1	A	3	8	2	B 1 8
				2	B. 2 9
				3	B 0 1
				3	B 1 2
				3	B 3 1

Mapper phase 2 :-

i	m j v
0	A 0 1
0	A 1 2
0	A 2 3
0	A 3 4.
1	A 0 5
1	A 1 6
1	A 2 7
1	A 3 8
0 ←	B 0 1
0 ←	B 0 4
0 ←	B 0 2.
0 ←	B 0 7.
1 ←	B 1 2.
1 ←	B 1 5.
1 ←	B 1 8.
1 ←	B 1 2.
2 ←	B 2 3
2 ←	B 2 6.
2 ←	B 2 9
2 ←	B 2 1.

from phase 1 only.

copied from 'j'

shuffled

Reduce

00

(A 0 1) (A 1 2) (A 2 3) (A 3 4)
(B 0 1) (B 0 4) (B 0 7) (B 0 1).

01

(A 0 1) (A 1 2) (A 2 3) (A 3 4).
(B 1 2) (B 1 5) (B 1 8) (B 1 2).

02

(A 0 1) (A 1 2) (A 2 3) (A 3 4).
(B 2 3) (B 2 6) (B 2 9) (B 2 1).

10

(A 3 4) (A 0 5) (A 1 6) (A 2 7)
(B 0 1) (B 0 4) (B 0 7) (B 0 1).

11

(A34) (A05) (A16) (A27)
(B12) (B15) (B18) (B12)

12

(A34) (A05) (A16) (A37)
(B23) (B26) (B29) (B21)

Now,

00

$(1 \times 1) + (2 \times 4) + (3 \times 7) + (4 \times 1)$
$\Rightarrow 1 + 8 + 21 + 4 = 36$

for all the combinations

1/4/2018

Euclidian distance → Numeric Attributes (Manhattan distance too)
(Binary Variables, Category Variables, Text Docs).

Similarity Measures for Text Docs :- Hamming code

Kmeans evolved as K-mediods.

Interval Scaled Data :-

One centre differs from the other centres in terms of units (kg, g).
Getting data from multiple sources.

mg to kg (or) kg to mg ?
Impacts magnitude of data.
You can't live with existing similarity measures unless there is some amount of standardization.

In JEE it's Normalization.

① Mean Absolute Deviation :-

$x_1 \dots x_n$.

$$m_f = \frac{x_{1f} + \dots + x_{nf}}{n}$$

frequency.

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

normalized.

$$s_f = \frac{|x_{1f} - m_f| + \dots + |x_{nf} - m_f|}{n}$$

Mean Abs Dev deals with outliers very well.

dissimilarity $[0, 1]$

↓

max. dis.

Some clustering algos work with $[-1, 1]$. Negative correlation.

Targeted Advertisements in mail is Clustering. Grouping of related products together.

Near Duplicate detection :- Old & new webpages. Identify duplicates & eliminate near duplicates. Don't get the two links to the same document.

Image Segmentation today is a clustering problem

Name a famous partitioning algo:-

Cosine Similarity Measure :- if docs are orthogonal then least similar

8/4/2019

Jaccard's Similarity
Hamming } distance
Edit
Taro's Similarity.
Levenstein
Winkler.

→ Categorical.
Strings.
Euclidean.
Minoks . . .
Manhattan.
} Num.

+How does clustering help in searching?
Cluster related customers together. Generate interesting rules.

↳ Clustering followed by ARM
Instead of searching 1,00,000 URLs search 100 URLs.

Document Vectors
D_1 → apple releases new ipod.
D_2 → apple releases new ipad.
D_3 → New apple pie recipe.
D_4 → phi releases new book with apple pie recipie

* Distance → scalar
What order are the words in., etc :- vector.

$$\text{Jacard's similarity} = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|}$$

similarity.

D_2, D_3
D_1, D_3
} 2

quantum of deviation.

3 out of 5 words

$J(D_1, D_2) = 3/5$. are common

Similar not just in terms of words but also ordering.

Other clustering options are DBSCAN & BIRCH.
↓ Density based.
↓ Balance iterative Reductive Clustering Hierarchy

Choosing the right similarity measure, will lead to better clustering

VLDB

***

TREC conference.

Term document frequency matrix.

	Apple	M.S.	Trump	Politics
D_1	10	20	0	0
D_2	30.	60.	0	0
D_3	0	0	10	20.

Apple occurs in D_1 10 times.

Stop & Stem words help in identifying real duplicates. Get unique solutions.

Singular value decomposition (SVD).
↳ What is LSI? Latent Symantic Indexing

"P - Tdf is transpose of Tdf"

Jaccards needs what are the 10 words

D_1 D_2 look ~~~~ closer to each other and are far from D_3.

$D_3 \perp (D_1, D_2)$.

D_1, D_2 growing in the same direction, D_3 is growing orthogonally.

* Similarity score is between 0 & 1.

** **H - distance :-**

no. of bit of difference.

d_1 = Peter ↑↑↓↓↓ d_2 = Pedro ↑↑↓↓↓

How many bits are different in the two docs.

$Hd(d_1, d_2) = 1+1+1 = 3$.

Do XOR

$P_1 = 10101$.

$P_2 = 10011$

$Hd(P_1, P_2) = 1+1 = 2$

No of one's gotten is the H-d.

* **Edit distance :-**

Another view of H-distance.

Jacardi is better here.

② Paul. Peter ①

No Jacard here. Pual. Pedro

Weird wierd. typo ✗ error.

Edit - Insert Delete Modify

What are the no of edits required to transpose / transform one string onto the other

Peter
Peteo.

Peter.
Pedro.

Paul. → req only 1 transpose opn

Pual.

↓

No new replacement
Only transposition (swapping).
Less costly operation.

Cost here is how many edits to transform along with swapping.

Swap is only for a pair.

Choose the one that minimizes the edit distance.

Some characters have more priority.

EXAMPLE:-

D_1 Jones
D_2 Johnson

D_1 to D_2 → 3 deletes & 5 adds & edits

This is feasible, but is it optimal.