08/06/2021.

CS 402.

Data Mining and Warehousing

Christy Varghese
CSE- A
Roll No. 34.

1 (b)

Backpropagation learns by iteratively processing a data set of training tuples, comparing the n/w's prediction for each tuple with the actual known target value. The target value may be the known class label of the training tuple or a continous value. For each training tuple, the weights are modified so as to minimize the mean-squared error b/w the network's prediction and the actual target value. The modifications are made in the 'backwards' direction.

7 (b) Apriori based frequent subgraph mining.

Apriori-based frequent subgraph mining algorithms share similar characteristics with Apriori based frequent itemset mining algorithms. The search for frequent graphs start with graphs of small size and proceeds in a bottom up approach by generating candidates having extra vertex, edge or path. It adopts a level-wise mining methodology. At each iteration the size of newly discovered frequent subgraph is increased by one. These new subgraphs that were discovered in the previous call to Apriori Graph.

ALGORITHM :

Input : D, a graph dataset.
min_sup, the minimum support threshold.

Page No :
Date :

Output: • S_k, the frequent subgraph set.

Method

$S_1 \leftarrow$ frequent single elements in the data set.
Call AprioriGraph (D, min_sup, S_1);

procedure AprioriGraph (D, min_sup, S_k)
step 1: $S_{k+1} \leftarrow \phi$
step 2: for each frequent $g_i \in S_k$ do
step 3: for each frequent $g_j \in S_k$ do
step 4: for each size (k+1) graph g formed by the merge of g_i & g_j do
step 5: if g is frequent in D and $g \notin S_{k+1}$ then
step 6: insert g into S_{k+1};
step 7: if $S_{k+1} \neq \phi$ then
step 8: AprioriGraph (D, min_sup, S_{k+1});
step 9: return;

(5) (a) The 3 clusters are A_1, B_1, and C_1 so calculating the Euclidean distance of each point from all the 3 clusters.

1st

	Centroid A_1	B_1	C_1
A_1	0	3.6	8.06
A_2	5	4.24	3.16
A_3	8.48	5	7.28
B_1	3.6	0	7.21
B_2	7.07	3.6	6.7
B_3	7.21	4.2	5.38
C_1	8.06	7.21	0
C_2	2.23	1.41	7.61

cluster 1 = {A1 (2,10)}.
center 1 = (2,10)
cluster 2 = {A3 (8,4), B1 (5,8), B2 (7,5), B3 (6,4), C2 (4,9)}
$$\text{center} = \left\{\frac{(5+8+7+6+4)}{5}, \frac{(8+4+5+4+9)}{5}\right\} = \{6,6\}$$

cluster 3 = {A2 (2,5), C1 (4,9)}
center = (1.5, 3.5)

(b) 2nd

	A1	B1	C1
A1	0	4.12	6.51
A2	5	4.12	1.58
A3	8.48	2.82	6.51
B1	3.6	2.23	5.7
B2	7.07	1.41	5.7
B3	7.21	2	4.52
C1	8.06	6.4	1.58
C2	2.23	3.6	6.04.

cluster 1 {(A1, C2, B1)}
cluster 2 {(A3, B2, B3)}
cluster 3 {(A2, C1)}

(5)

(a) Lazy classification uses richer hypothesis space, which can improve classification accuracy. It requires less time for training than eager classification. A disadvantages of lazy classification is that all training tuples need to be stored, which leads to expensive storage costs and requires efficient indexing techniques. Another disadv. is that it is slower at classification because classifiers are not built until new tuples need to be classified.

Eager classification is faster at classification than lazy, because it constructs a generalization model before receiving any new tuples to classify. weights can be assigned to attributes, which can improve classification accuracy. Disadvantage of eager classification are that it must commit to a single hypothesis that covers the entire instance space, which can decrease classification and more time is needed for training.

④ ⓐ characteristics of Social N/w.

Social networks are rarely static. Their graph representations evolve as nodes and edges are added or deleted over time.

ⓘ Densification power law:

Earlier, it was believed that as a n/w evolves the number of degrees grow linearly in the number of nodes. However extensive experiments have shown that networks become increasingly dense over time with the average degree increasing.

ⓘⓘ shrinking diameter: It has been experimentally shown that the effective diameter slowly ~~increases~~ decrease as the n/w grows. This contradicts an earlier belief that the diameter slowly increase as a func. of n/w size.

④ ⓑ No. of points = 4.

cluster (3,5) (2,3) (4,3) & (1,5)

$CF = (N, \vec{LS}, SS)$.

$\vec{LS} = \sum_{i=1}^{N} \vec{x_i}$

$SS = \sum_{i=1}^{N} x_i^2$

$9+$
$4+$
16
29

50
$+18$
68

$CF_1 = \{4, (3+2, 4+1, 5+3+3+5), (3^2+2^2+4^2+1^2, 5^2+3^2+3^2+5^2)\}$

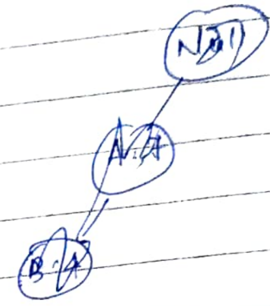
$= \{4, (10, 16), (30, 68)\}$

Q7

(b) ~~Itemset~~ min sup = 2

$C_1 \Rightarrow$

Itemset	Support count
A	7
C	2
B	6
T	2
S	6



$L_1 \Rightarrow$

A	7	cluster 1
C	2	
B	6	
T	2	
S	6	

$C_2 \Rightarrow$

(A,C)	2
(A,B)	4
(A,T)	2
(A,S)	4
(C,B)	0
(C,T)	0
(C,S)	0
(B,T)	2
(B,S)	4
(T,S)	1

$L_2 \Rightarrow$

(A,C)	2	cluster 2
(A,B)	4	
(A,T)	2	
(A,S)	4	
(B,S)	4	
(B,T)	2	

$C_3 \Rightarrow$

(A,B,C)	1
(A,B,T)	2
(A,T,S)	1
(B,A,S)	2

$L_3 \Rightarrow$

(A,B,T)	2	cluster 3.
(B,A,S)	2	

$C_4 =$ (A,B,S,T) 1

Consider $L_3 = \{(A,B,T), (A,B,S)\}$

min confidence = 70%

~~A~~ Association rules formed from (A, B, T)

$\{A, B\} \Rightarrow T$ confidence = $\frac{2}{4}$ = 50%

$\{A, T\} \Rightarrow B$ confidence = $2/2$ = 100% ✓

$\{B, T\} \Rightarrow A$ ~~confidence ⇒ 2/1 = 2~~ $2/2$ = 100% ✓

$A \Rightarrow \{B, T\}$. $\frac{2}{7}$ = 28.5%

$B \Rightarrow \{A, T\}$ $2/6$ = 33.3%

$T \Rightarrow \{A, B\}$ $2/2$ = 100% ✓

Association rules formed from (A, B, S)

$\{A, B\} \Rightarrow S$ conf = $2/4$ = 50%

$\{A, S\} \Rightarrow B$ $2/4$ = 50%

$\{B, S\} \Rightarrow A$ $2/4$ = 50%

$A \Rightarrow \{B, S\}$ $2/7$ = 28.5%

$B \Rightarrow \{A, S\}$ $2/6$ = 33.3%

$S \Rightarrow \{A, B\}$ $2/6$ = 33.3%

min^m conf = 70%

∴ Strong association ⇒

$\{A, T\} \Rightarrow B$ = 100%

$\{B, T\} \Rightarrow A$ = 100%

$T \Rightarrow \{A, B\}$ = 100%

②

ⓐ $$\beta = \frac{\sum_{i=1}^{8} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{8} (x_i - \bar{x})^2}$$

$$\alpha = \bar{y} - \beta\bar{x}$$

$$\bar{x} = \frac{866}{12} = \underline{\underline{72.167}}$$

71
~~72~~ . 72 .
~~71~~ . 79

$$\bar{y} = \frac{888}{12} = \underline{\underline{74}}$$

$$\beta = \frac{2004}{3445.668} = \underline{\underline{0.5816}}$$

$$\alpha = 72 - 0.5816 \times 72.167$$
$$= \underline{\underline{32.028}}$$

2 ⓑ $y = \alpha + \beta x$

$x = 86$

$y = 32.028 + 0.5816 \times 86$

$y = \underline{\underline{82.0456}}$

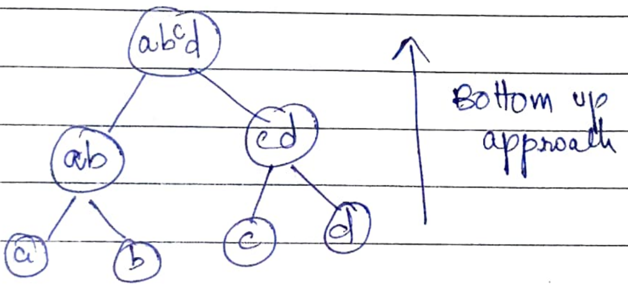
(7)

(a) ~~Aglo~~ Agglomerative clustering also known as bottom up approach. A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to prespecify the no. of clusters. Bottom up algorithms treat each ~~other~~ data as a singleton cluster at the outset and then successively agglomerate pairs of clusters until all clusters have been merged into a single cluster that contains all data.

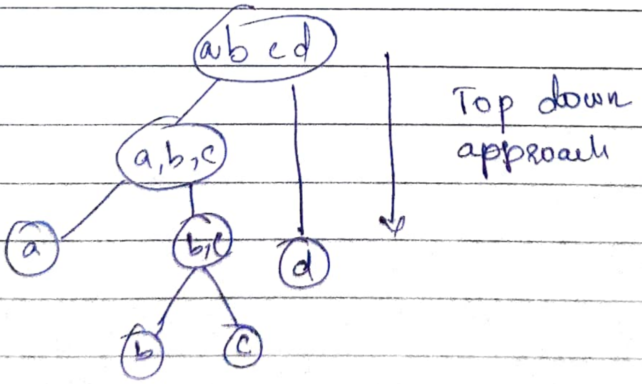
Divisive clustering also known as top-down approach. This algorithm also does not require to prespecify until individual data have been splitted into singleton clusters.

Divisive clustering is more efficient, complex and accurate than agglomerative clustering.

eg:



agglomerative.



divisive