ASSESMENT WORKSHOP PAPER -2022-23

Name: PRANCHAL

Class: AIML-6-A

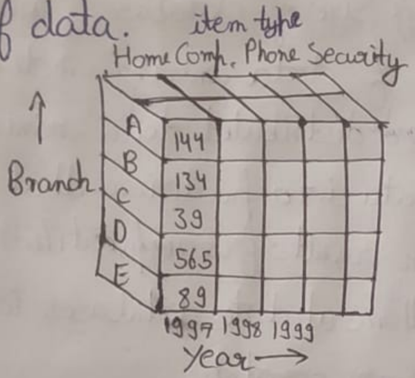
UID: 20BCS6858

Subject: Data Mining and Warehouse

Subject Code: 20CSF-333

Q1 Outline the major steps of the data cube –based implementation of class characterization.

Ans... Data class characterization in data mining is summarization of general characteristics or features of target class of data. The data corresponding to user specified class are typically collected by query. Data cubes are multidimensional matrix for grouping of data, these are used in class characterization as data cubes also helps in summarization of view of data.



This is pictorial representation of data cube having attibuts like

->branch (A, B, C, D, E)

->item type (Home, Comp., Phone, security)

->Year (1997, 1998, 1999)

Data cube has various steps for class characterization :->

① Roll up -> To aggregate certain similar data attributs having same dimensions, roll up is used. It reduces dimension.

② Drill down -> Reverse of roll up operation. In this we take a particular information and the divide it for further analysis. It increases dimension for betterement.

③ Slicing & dicing -> Slicing filters unneccessary operations. Only particular attribute asked by user is displayed. Dicing dols a multidimensional cutting. It selects dimensions on some criteria.

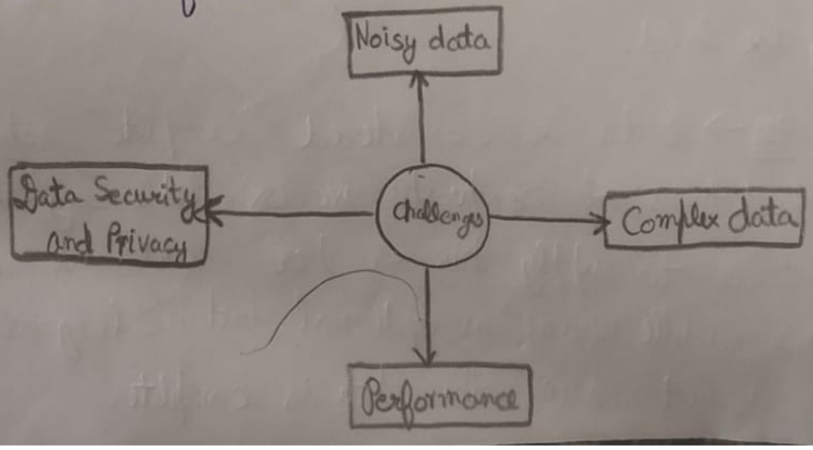
④ Pivot -> It is for viewing point of view. It basically tranforms the data by rotating the axes about a particular attribute.

Q2 What are the major challenges of mining a huge amount of data in comparison with small amount of data.

Ans.. <i> Security and Social Challenges :-> Large datasets contain more amount of information. So, if data is not extracted properly it may lead to missing of important or private information Illegal access to information and confidential nature of information may be an important issue for the data which is large as compared to small data.

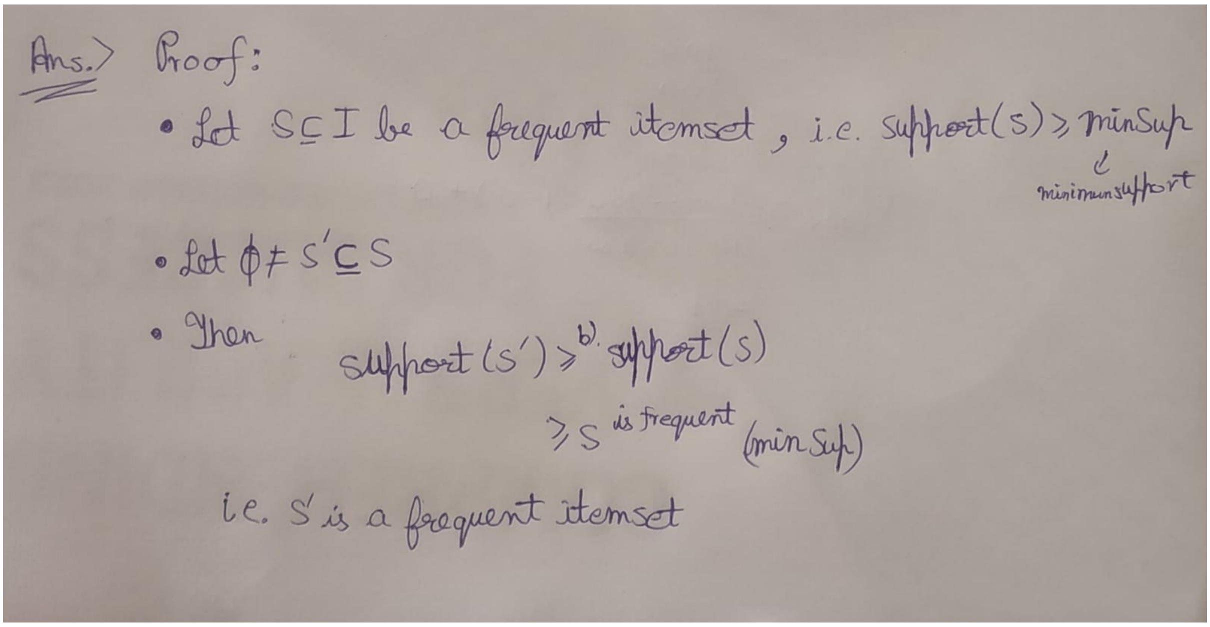
<ii> Complex data :-> Real world is heterogeneous and it could be multimedia data containing images, audio and video, complex data, temporal data, spatial data, time series etc. Large dataset may be not suitable for extraction of information in such complex data whereas distributing in small datasets may help to extract information easily.

<iii> Performance :-> The performance of data mining system depends on the (on × the) efficiency of algorithms and techniques we are using. The techniques that are used may take a huge amount of time on large datasets whereas it may be suitable for small datasets. Performance may be a hardcore issue for a device which has less specifications.



Q3 **The Apriori algorithm makes use of prior knowledge of subset support properties**

a) **Prove that all non- empty subset of a frequent item set must also be frequent .**



Ans.> Proof:

- Let $S \subseteq I$ be a frequent itemset, i.e. $\text{support}(S) \geq \underset{\substack{\downarrow \\ \text{minimum support}}}{\text{minSup}}$
- Let $\phi \neq S' \subseteq S$
- Then

$$\text{support}(S') \overset{b).}{\geq} \text{support}(S) \overset{S \text{ is frequent}}{\geq} (\text{min Sup})$$

i.e. S' is a frequent itemset

Q4 **Use the method below to normalize the following group of data**

200, 300, 400, 600, 1000

a) **Min-max normalization.**

b) **Z- score normalization .**

Ans: 200, 300, 400, 600, 1000

<i> Using min-max normalization;

$$v' = \left\{\frac{v - max}{max - min}\right\}(new\,max_A - new\,min_A) + new\_min_A$$

$$v'_{200} = \left\{\frac{200-200}{1000-200}\right\}(1-0) + 0$$

$$= \underline{\underline{0}}$$

$$v'_{300} = \frac{300-200}{1000-200}(1-0)+0$$

$$= \frac{100}{800}$$

$$= \underline{\underline{0.125}}$$

$$v'_{400} = \frac{400-200}{1000-200}(1-0)+0$$

$$= \frac{200}{800}$$

$$= \underline{\underline{0.25}}$$

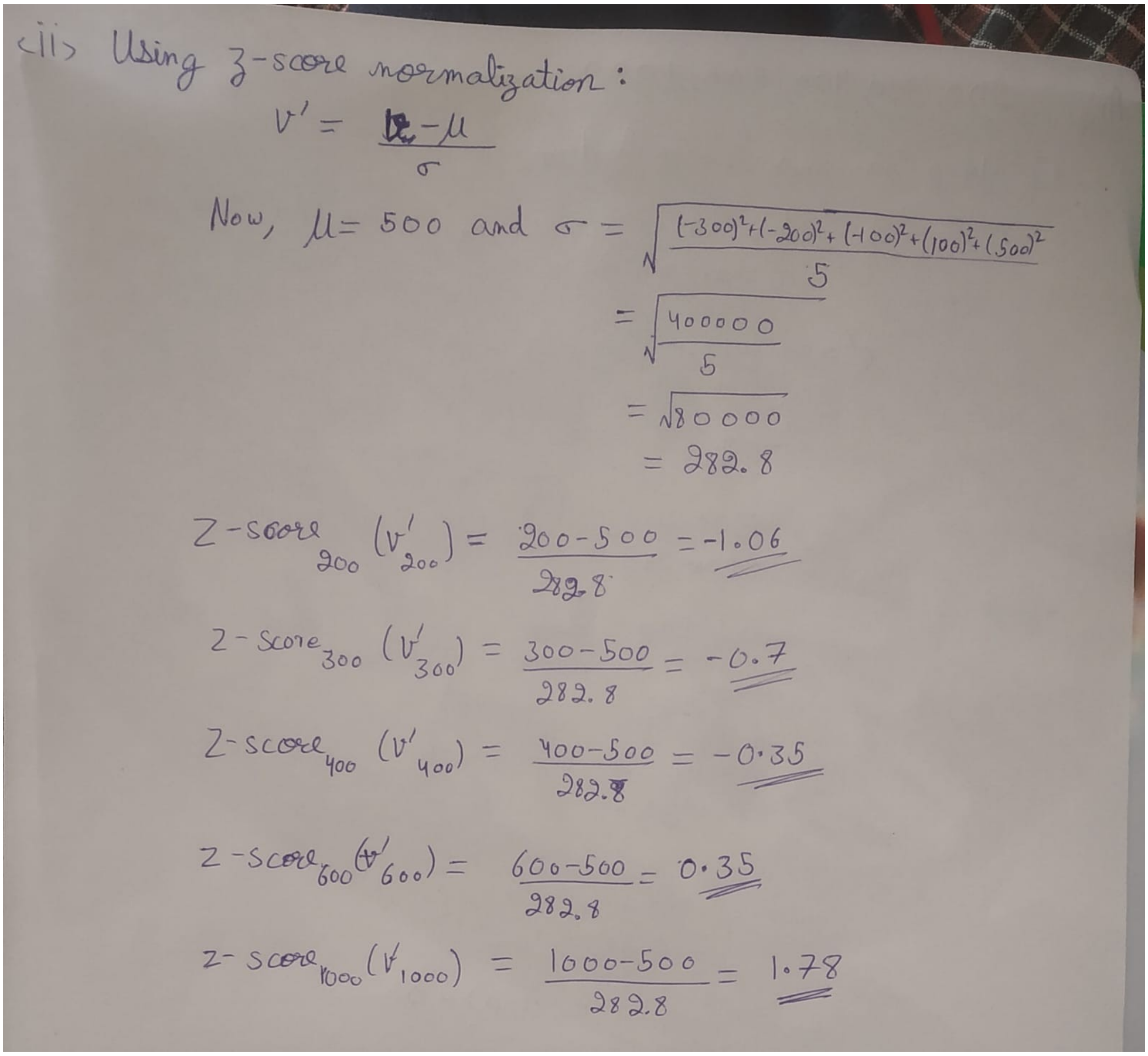
$$v'_{600} = \frac{600-200}{1000-200}(1-0)+0$$

$$= \frac{400}{800}$$

$$= \underline{\underline{0.5}}$$

$$v'_{1000} = \frac{1000-200}{1000-200}(1-0)+0$$

$$= \underline{\underline{1}}$$



<ii> Using z-score normalization :

$$v' = \frac{v - \mu}{\sigma}$$

Now, $\mu = 500$ and $\sigma = \sqrt{\frac{(-300)^2+(-200)^2+(-100)^2+(100)^2+(500)^2}{5}}$

$$= \sqrt{\frac{400000}{5}}$$

$$= \sqrt{80000}$$

$$= 282.8$$

$$Z\text{-score}_{200}\ (v'_{200}) = \frac{200-500}{282.8} = \underline{\underline{-1.06}}$$

$$Z\text{-Score}_{300}\ (v'_{300}) = \frac{300-500}{282.8} = \underline{\underline{-0.7}}$$

$$Z\text{-score}_{400}\ (v'_{400}) = \frac{400-500}{282.8} = \underline{\underline{-0.35}}$$

$$Z\text{-score}_{600}(v'_{600}) = \frac{600-500}{282.8} = \underline{\underline{0.35}}$$

$$Z\text{-score}_{1000}(v'_{1000}) = \frac{1000-500}{282.8} = \underline{\underline{1.78}}$$