- & calculat average dissiomilarity of the cluster obtained from step 3 If this value is less than current minium & retain the k medoids found in step 2 as the best set of medoids obstained found in step 2 as the best set of medoide obtained so far
- & Return to ~~the~~ Step 1 to stard the Next Iterati

C) ENHANCED CLARANS:-

This Method is different From PAM CLARA AND CLARANS Thus method is produced to improve the accuracy of outlier CLARANS is a Neo partitioning alogorithm which is an Nodes Instead of selecting as Random Searching oprations It is similar to CLARANS but theese selected arbitrary Nodes Reduce the Nor of Iteration of CLARANS procedure

- & Input parameters Num locals & Max Neighbour Initialize i to 1 & mincost to a large Number
- & calculating distance between each data points
- & If s has a slower cost set current + s & go to step
- & other wise when j > max neighbour compare the lost of anrent with min cost If the former is less than min cost set min Cost of scurent & set best Node to current Increment i by 1 If i > Num local but put best Node & halt other wise go to step 2

A) PAM (Partition around Medoid)

PAM uses a K Method medaid Method for clustering It is very robust when compared with K-means in the presence of Noise & outlier Mainly in the presence of Noise & outliers & swap phase

1) Build phase: This step is sequentially select K objects which is centrally located This K-objects to be used as Kmedoide.

2) swap phase: clculat the total cost for each poir of selected & Non-selected object

PAM Procedure:

- Input the dataset D
- Randonly select K objects froms the dataset G
- calculate the Total cost T For each pair of selected Si & Non selected object Sh
- For each poir if T Si<0 then it is Repleced Sh
- Then find similar medoid for each non-selected object
- Repeat step 2,3 & 4 until find the medoids

B) CLARA (clustering Large Applications)

- For i=1 to 5 Repeat steps 2 to 5
- Draw a sample of 40+2K objects Randomly From the entire data set & call PAM algorithm to find K medoids of the sample
- For each object o in the entire data set determine K-medoids which is most similar to

along with simplicity Naive Bayes is known as out perform even highly sophisticated classification methods

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) & P(x|c) Look at the eqn below

Likelihood ← P(x|c) P(c) → class prior probability

P(c|x) = P(x|c) P(c) / P(x)

P(c|x) → posterior probability

P(x) ← predictor prior probability

$(Cx) = P(x_1|C) \times P(x_2|C) \times \cdots \times P(x_n|C) \times P(x)$

8) Discuss clustering based approaches to detect outliers

⇒ considered outliers The advantage of the clustring based In these approaches small cluster are considered outliers The advantage of the clustering based approaches is that they do Not have to be supervised moreover clustering-based techniques are capable of being used in an incremental mode

There are various kinds of clustering based outlier detection approach have been proposed which are following.

Name	age	income	loan	risk
sandy	young	low	risky	
Bill	"	"	low risky	
Rick	M-a	high	safe	
Jack	"	low	risky	
susan	shiva	"	safe	
...	...	...	...	...

John Mican aged low income loan decision risky

f) Explain the process of Predicting a class label using Naive Bayesian classifier

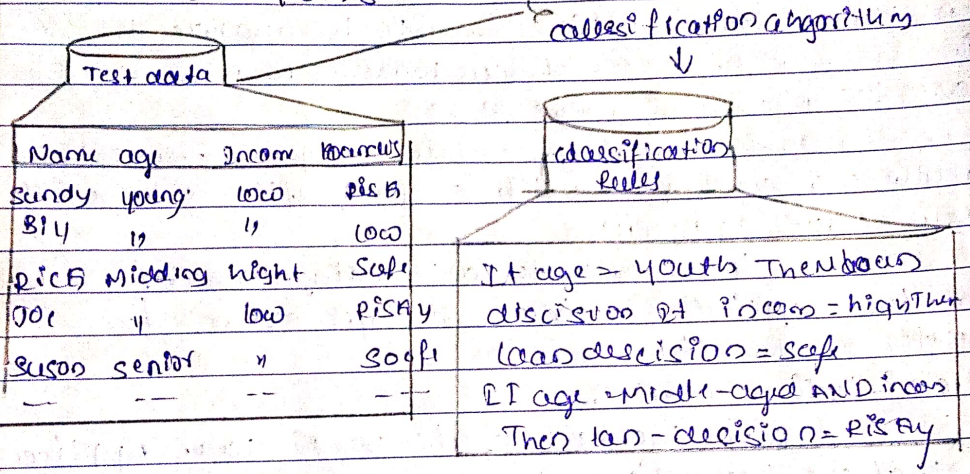
It is a classification based on Bayes Theorem with an assumption of independence among predictors In simple terms a naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature

For Example a fruit may be considered to be an apple if it red Round about 3 inches in diameter Even if these features depend on each other or upon the existence of the other features all of these properties independently contribute to the probability that this fruit is and apple that is why it is known as Naive

Naive Bayes model is easy to build & particular use full for very large data sets

of data base tuples & their associated class labels

* Each tuples that constitutes the training set is referred to as a category or class these tuples can also be refered to as sample object or data points



2) Using classifer for classification:-

In this step the classifer is used for classification Here the test data is used to estimate the accuracy of classification Rules can be applied to the New data tuples if the accuracy is considered acceptable

6) Briefly explain the data classification process

classification models predict categorical class labels

For Ex: we can build a classification model to categori - zi bank loan applications either safe or Risky

following are the examples of cases where the data in order to analysis a is classification

1) A bank loan officer wants to analyze the data in order to know which customer (loan applicant) are Risky or which are safe.

2) A marketing manager at a company needs to analyze a customer with a given profile who will buy a New computer

These examples are Risky or Safe for loan application data and yes or no for marketing data

Classification work :-

With the help of the bank loan Applica - tion that we have discussed above let us understand the working of classification the data classificatio - n process includes the two steps

1) Building the classifier or model
2) Using classifier for classifier

=> 1) Building classifier or model :-

* This step is the learning step or the learning set or the learning phase
* In this step the classification algorithm build the classifier
* The classifier is built from the training set made

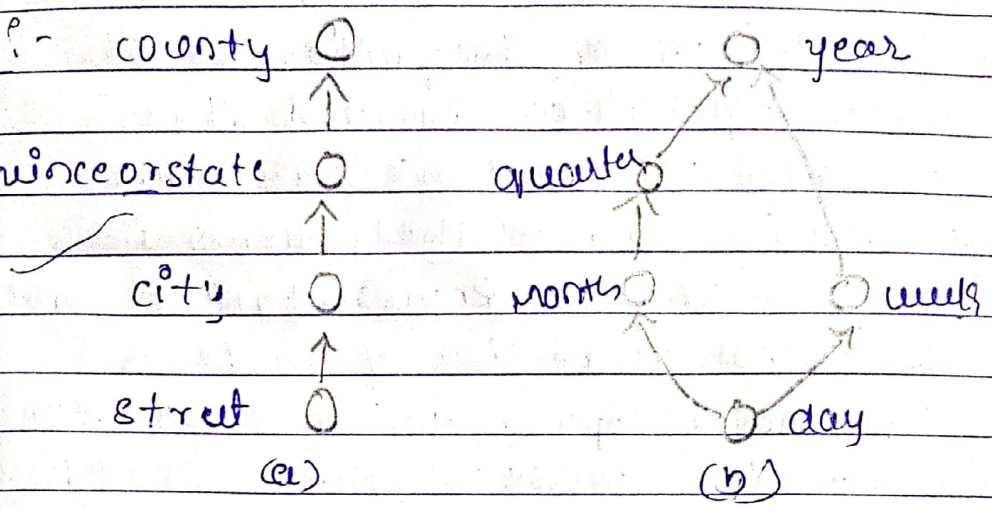
Given a transaction data base let = O {o1, o2, ---- od} be the set of items that apper in the set of items in transaction ti for For any itemes set P. let O(P) be the corresponding set of items & T(P) = {t ∈ T | O(P) ≥ min supponding set of items we say O(P) is the expression of P. & |T(P)| is the support of P An item set P is frequent if |T(P)| ≥ min sup where min sup is a user specified threshold The task of frequent item sets mining is to find all the frequent itemsets.

There have been many scable methods developed for frequent pattern mining [3.13.9] However the Real bottleneck of the problem is not at the efficiency but at the usability.

To salve this problem is natural to explore how to "compress" the patterns i.e find a concise & succint Representation that describs the whole collection of patterns

Two major approaches have been develo -ped in this direction lossless compression & lossy apprximation the former Represented by the closed frequent item set (16.18.19) emphasi too much an the supports of Patterns so that its compression power is quite limited the latter Represented by the maximal frequent itemsets [8,10,11) O(p)

* Many concept hierarchies are implicit within the database schema
* A concept hirarchy that is a total or partial among attributes in a database schema is called schema hierarchy
* conapt hierarchies that are common to many applications may be predefined in the data mining system
* Data mining should provide users with the flexibility to tailor predefined hirachys according to their particular needs



5) Explain how pattern compression can be achieved by pattern clustering

=> Frequent-pattern (or itemsets) mining has been a focused research theme in a data mining due to its broad applications of mining association (2,3) correlation (6) causalications (7) sequential pattern (4), episodes (5) partial periodicity (12), emerging pattern (8) & many other important data mining task

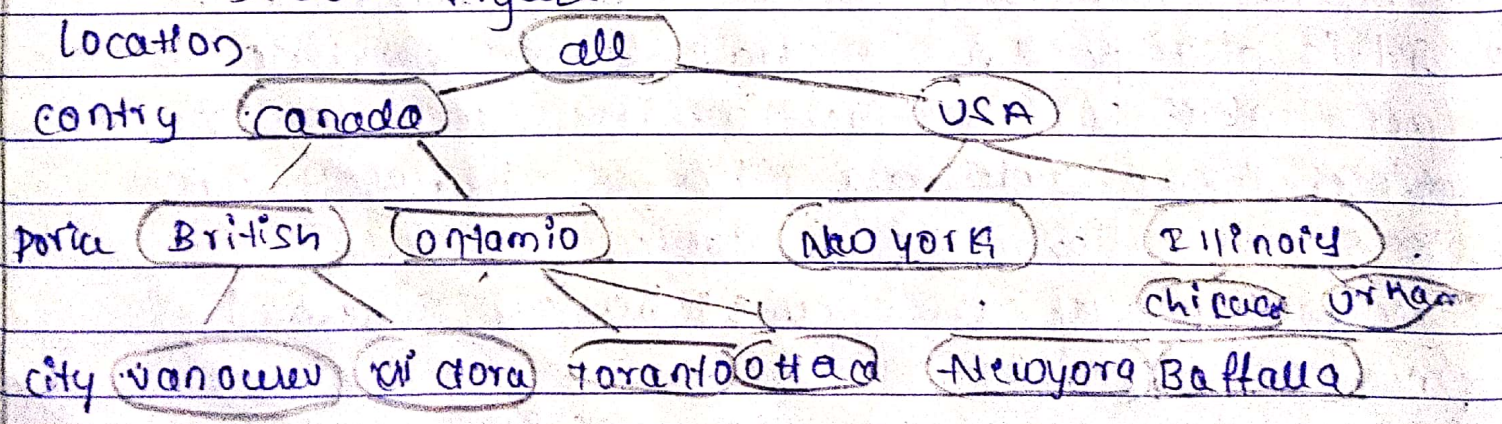
to the dimensions

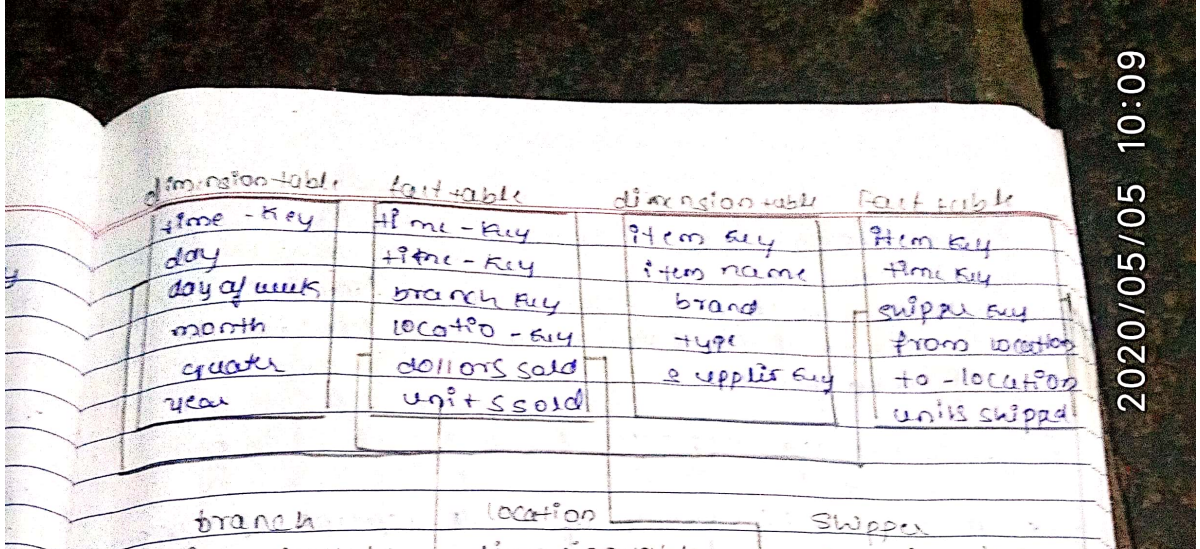
=> Disadvantages:-

* Fact constellation solution is difficult to Maint
* complexity of the schema involved due to the No of aggregations

④ With Example illustrate how concept hierachies are useful in OLAP

=> * A concept hierarchy difines a sequence of mappings from a set of low-level concepts to highers level more general concepts
* consider a concepts for the dimension location
* city values for location include van couver Toronto, New york and chicago
* Each city, however can be mapped of the provin or state to which it belongs
* The provinus and states can is turn be mapped to the county to which they belong
* These mappings a concept hierarchy for the dimension location mapping a set of low level concepts to hight-level more general concepts
* This concept hierarchy is illustrated in bellow Figure.



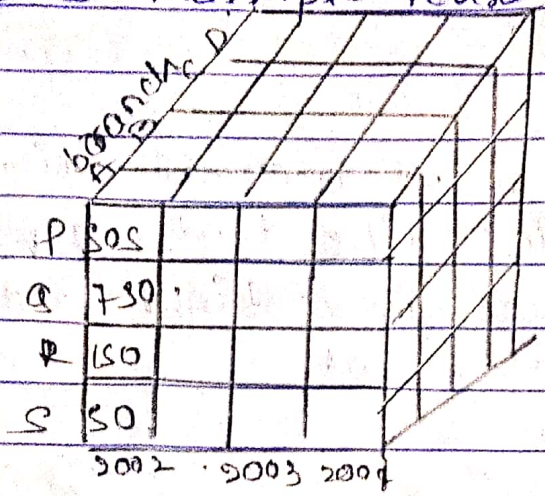


* The Sales fact table is same as that in the star schema
* The shipping fact table table has the five dimensions. Namely item key, time. key, shipper key from. location to. location
* The shipping fact table also contains two measures, namely dallars sold & units sold
* It is also posible to share dimension tables between fact tables for example, time, item and loction dimension tables are shared between the sales & shipping fact table

=> Advantages of Fact constellation Schema:

* provides a flexible schema
* Different fact tables are explicitly assigned to

Valu . corresponding to the data point in multidimensional space concept hierachies may exists for each attribute, allowing the analysis of data at multiple levels of abstraction



The cube create at the lowest level of consideration is Referred to as the base cuboid The base cuboid should Resemble an individual entity of interest, such as sales or customer.

The above diagram is that you have collected the data for your analysis. These data consist of sales per quarter for the years 2002 to sales rather than total per quarter

6) sequential patterns

This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period

7) Predictions:-

Predictions are used a combination of the other data mining techniques like trend sequences past events are instances in a right sequence for predicting of future event

2) Discuss the need for data Reduction with an illustrative example explain data cube aggregation technique

=> Need for data Reduction :-

* A database or data warehouse may store terabytes of data so it may take very long to perform data analysis and mining on such huge amounts of data
* data Reduction techniques can be applied to obtain a Reduced Representation of the data set that is much smaller in volume but still contain critical information
* complex data analysis or mining may take a very long time to run on the complete data set

Data cube aggregation :-

Aggregation operations are applied to the data is constructmultidimensional aggregate cube information Each cell holds an aggregate data

2) clustering :-

clustering analysis a data mining mining technique to identify data that are like each other This proces helps to understand the differences and similarities between the data

3) Regression :

This analysis the data mining method of identifying & analyzing the relationship between variables It is used to identify the likli hood of a specific variable given the presence of other varable

4) Association Rules :

This data Mining technique helps to to find the association between two or more Items It discovers a hidden pattern in the data set

5) outer detection :-

This type of data mining technique refers to observation of data items in the data set which do not match an Expected Pattern or expected behaviour This technique can be used in a varity of domains such as intrusion detection trand of fault detection etc outer delterns also called outer detection

Appayya S Badiger

CS181005

data mining

2017

1) What is data mining ? Discuss the techniques that influence the development of data mining methods

Data Mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics and data base systems. Data mining also known as Knowledge Discovery in databases refers to the Nontrivial extration of implicit, previously unknown and potentionly usefull information from data stored in databases

Data mining Techniques :-

1) classification 2) clustering 3) Regression 4) outlier 5) sequential patterns. 6) prediction 7) Associal Rule

1) classification :-

This analysis used to Retrieve important and Relvant information about data and method This data mining method helps to classify data in different classes

2) clustering :-

clustering analysis is a data mining technique to identify data that are like each