SECE 3031 Data Warehouse & Data Mining

Page No.
Date

3 Marks Question

1. How to handle missing values

Every dataset we come across will have missing values. The following are the ways to handle missing values

1. Ignoring the tuple, usually we use this when class label is missing
2. Filling up the missing values manually, is a very tedious and infeasible task so we can use a global constant to fill in the missing values
3. Use the attribute mean to fill in the missing value or use the attribute mean for all sample of the same class to fill in the missing value
4. Use the most probable value to fill in the missing value, like inference based such as regression, baysian formula, decision tree.

Q.3 How to handle noisy data?

Noisy data can be handled by

1. Binning Method i.e.

First sort data and partition into (equal depth) bins then one can smooth ~~by~~ by means, medians, or ~~smo~~ smooth by bin boundaries.

J8SE02IT047

2. Through clustering we can detect and remove outliers. similar values are organized as clusters and values that fall outside the set of clusters called outliers.

3 Regression, here data can be smoothed by fitting data to a function
Here we can use linear regression and multiple linear regression etc.
Through linear regression we can find the best line to fit two attributes, so that one attribute can be used to predict the other

2 Major issues in data mining

1. Handling noisy or incomplete data - The data cleaning is a tedious and major issue in data mining, as the larger the dataset the more noisy or incomplete data is difficult to handle The data cleaning methods are required that can handle noisy data, incomplete objects. If data cleaning methods are not present then the accuracy of the discovered patterns will be poor

2. Pattern Evaluation :- It refers to interestingness of the problem, sometimes the pattern discovered should be interesting because most of the time, they represent common knowledge or lack novelty

Page No.
Date.

3. Efficiency and scalability of data mining algorithms

In order to ~~get~~ effectively extract the information from huge datasets, data mining algorithms should be efficient and scalable

4. Presentation and visualization of results :- The ~~results~~ results should be expressed easily and in a understable manner. Once the patterns are discovered they should be expressed easily in high level language such that even a lay-man should understand it.

4. Star schema in data warehouse

Star schema is one of the multidimensional schema designed to address the unique needs of very large data base designed for analytical purposes

Star schema is a data warehouse, in which the centre of the star can have one fact table and a number of associated dimension table. It is also known as Star Join Schema

Some characteristic of star schema

Every dimension is represented with only one dimension table
The dimension table should contain the set of attributes
Fact table would contain key and attributes measure

18SE02IT047

- The dimension tables we not normalized
- The dimension table is joined to the fact table using a foriegn key
- The dimension table are not joined to each other

5. Associative classification

It is a classification method based on the association rules. An association rule with the class labels as its consequent provides a clue that a tuple satisfying its antecedent belongs to a specific class This can therefore be used as the bassis of classfication.

Let $I = \{i_1, i_2, i_3, i_4, \ldots i_n\}$ ⇒ items in itemset

The rule is $D = \{t_1, t_2 \ldots t_m\}$ ⇒ set of transactions called database

Each transaction in D has a unique transaction ID and contains a subset of items in I.

An example rule:- {butter, bread} ⇒ {milk}
ie that if butter and bread are bought customers also buy milk.

Confidence of rule
$= conf(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$

$Lift(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)}$

support (supp(x)) is defined as proportion of transactions in the dataset which contains itemset

1BSE02IT047

Page No.
Date

5 Marks Questions

4. Data Mining functionalities are used to define patterns or correlation in datamining tasks

1. Data Characterization

It refers to the summary of general characteristics or features of the class e.g. To study the characteristics of a software product whose sales increased by 5%.

2. Data discrimination:- It compares common features of class which is being studied. The output of this process can be represented in many forms like pie chart, bar graph, curves etc.

3. Association analysis:- The process involves uncovering relationship bw data & deciding the rules of association. It is a way of discovering the relationship bw various items e.g. it can be used to determine sales of item that are frequently purchased together.

4. Correlation Analysis:- It is a mathematical technique that can show whether & how strongly the pairs of attributes are related to each other. e.g. short people have less fat weight

Data mining activities can be divided in two ways predictive & descriptive. Aim is to understand data without previous knowledge about data E.g. count, avg.

18SE0257047

5. Explain linear & non-linear regression

Linear regression often tries to find the mathematical relationships between variables.
It is a simplest form of regression. It tries to find out model the relationship between two variables by fitting a linear equation to observe the data

If the outcome is a straight line, then it is considered as linear and if it is curved line, it is a non linear model
Linear regression always uses a linear equation, $Y = a+bx$, where Y is the dependent variable and x is explanatory variable

Non-linear regression

It can be modeled by adding polynomial terms to basic linear model
Polynomial regression is a special case of multiple regression. that is the addition of higher terms like x^2, x^3,
It uses two or more independent variables to predict an outcome and a single continuous dependent variable

eg. $Y = a_0 + a_1 x + a_2 x_2 + a_n x_{n-1}$

Linear regression relates with straight line
Non linear with curve
Eg of non-linear regression use $\Rightarrow$ prediction of population growth over time

18SE02ST042

Page No.
Date.

ed Scatterplot of changing population data over time shows that that there seems to be a relation bw time and population in a non-linear model

10 Marks Question

1 K means algorithm

Data :- 2, 4, 8, 10, 12, 3, 20, 30, 11, 13, 25

Sort data :- 2, 3, 4, 8, 10, 11, 12, 13, 20, 25, 30

let $M_1 = 4$ & $M_2 = 20$

iteration 1

	2	3	4	8	10	11	12	13	20	25	30
D_1	2	1	0	4	6	7	8	9	16	21	26
D_2	18	17	16	12	10	9	8	7	0	5	10

So $C_1 = \{2, 3, 4, 8, 10, 11\}$

$C_2 = \{12, 13, 20, 25, 30\}$

$\therefore M_1 = (2+4+3+8+11+10)/6 = 6.33$

$M_2 = (12+13+20+25+30)/5 = 20$

iteration 2 :-

	2	3	4	8	10	11	12	13	20	25	30
D_1	4.33	3.33	2.33	2.33	4.33	6.33	7.33	14.33	19.33	24.33	
D_2	18	17	16	12	10	9	8	7	0	5	10

$D_1 = 4.33$

$D_2 = 10$

$C_1 = \{2, 3, 4, 8, 10, 11, 12\}$ M1 = 14

$C_2 = \{13, 20, 25, 30\}$ M2 = 22

Iteration 3

	2	3	4	8	10	11	12	13	20	25
D_1	5.14	4.14	3.14	0.86	2.86	3.86	4.86	5.86	6.12	17.86
D_2	20	19	18	14	12	11	10	9	2	3

	D
D_1	22.30
D_2	22

$\therefore C_1 = \{2, 3, 4, 8, 10, 11, 12\}$

$C_2 = \{13, 20, 25, 30\}$

In iteration 2 & 3 we get same clusters

$C_1 = \{2, 3, 4, 8, 10, 11, 12\}$

$C_2 = \{13, 20, 25, 30\}$

2. Support = 22%. confidence = 70%.

Item	Frequency	Support
I_1	6	6/9 = 66%
I_2	7	7/9 = 77%
I_3	6	6/9 = 66%
I_4	2	2/9 = 22%
I_5	2	2/9 = 22%

All items support ≥ 22%.

Items	Frequency	Support
I_1, I_2	4	4/9 = 44%
I_1, I_3	4	4/9 = 44%
I_1, I_4	1	1/9 = 11%
I_1, I_5	2	2/9 = 22.2%
I_2, I_3	4	4/9 = 44%
I_2, I_4	2	2/9 = 22.2%
I_2, I_5	2	2/9 = 22.2%
I_3, I_4	0	0/9 = 0%
I_3, I_5	1	1/9 = 11.1%
I_4, I_5	0	0/9 = 0%

C_2 =

Item	S
I_1, I_2	4/9
I_1, I_3	4/9
I_1, I_5	2/9
I_2, I_3	4/9
I_2, I_4	2/9
I_2, I_5	2/9

C_3 =

item	F	Support
I_1, I_2, I_3	2	22.2%
I_1, I_2, I_4	1	11.1%
I_1, I_2, I_5	2	22.2%

Page No.
Date.

Association Rules

For (I_1, I_2, I_3) confidence

$I_1, I_2 \rightarrow I_3 = 2/4 = 50\%$.
$I_1, I_3 \rightarrow I_2 = 2/4 = 50\%$.
$I_2, I_3 \rightarrow I_1 = 2/4 = 50\%$.
$I_1 \rightarrow (I_2, I_3) = 2/6 = 33\%$.
$I_2 \rightarrow (I_1, I_3) = 2/7 = 29\%$.
$(I_3) \rightarrow (I_1, I_2) = 2/6 = 33\%$.

For (I_1, I_2, I_5)

confidence

$(I_1, I_2) \rightarrow I_5 = 2/4 = 50\%$.
$(I_1, I_5) \rightarrow I_2 = 2/2 = 100\%$.
$(I_2, I_5) \; I_1 = 2/2 = 100\%$.
$I_1 \rightarrow I_2, I_5 = 2/6 = 33\%$.
$I_2 \rightarrow I_1, I_5 = 2/7 = 23\%$.
$(I_3 \rightarrow (I_1, I_2) = 2/2 = 100\%$.

ISSE021T047

Page No.
Date

5 Marks Question

2. Rule based classifiers

Rule based classifier another type of classifier based which makes the class decision depending by using various if else rules. This classifier is generally used to make discriptive models

The condition with if is called antecedent

and the predicted class of each rule is called consequent

Properties of rule based classifier

Coverage:- The percentage of records which satisfy the antecedent conditions of a particular rule.

The rules generated by the rule based classifiers are generally not mutually exclusive i.e. many rules can cover the same record

The rules generated may not be exhaustive i.e there may be some records which are not covered by any rule

The decision boundaries created by them is linear

How to generate rule

Rules should be generated using general to specific approach or specific to general. Start with a rule with no antecedent and keep on adding conditions till we see major improvements in our evaluation metrics

E.g. Class Cap Shape Cap Surface Bruises Odour

edible flat scaly yes anise

poisonous convex scaly yes pungent

edible convex smooth yes almond

E.g. of rule

Odour = pungent and habitat = urban → class = poisonous

Bruises = yes → class = edible

3 Classification by backpropogation

It is a neural network algorithm. It is the method of fine tuning the weights of a neural network based on the error rate obtained in the previous epoch (iteration) Proper tuning the weights allows you to reduce error rates and make the model model more reliable by increasing its generalization. This method helps calculates the gradient of a loss of function w.r.t all the weights in the network

B.

1. DBSCAN clustering in detail

It is a base algorithm for density based clustering. It helps to discover the clusters of different shapes and sizes from a large amount of data containing noise and outlier.

It uses two parameters and eps (ϵ)

Minpts :- It is the minimum no. of points clustered together for a region to be considered dense

eps(ϵ) => It is a distance measure that is used to locate the points in the neighbourhood of any point

There are 3 points after DBSCAN clustering is complete

Core => point with least m points within distance n from itself

border => point with atleast one core point at a distance n

noise => point with neither a core nor a border

Algorithm

- Picking up a point in the dataset attributes [until all the points are visited]
- If there are at least minpoint' points within a radius of 'E' to the point then we consider all these points to be part of same cluster
- The clusters are then expanded by recursively repeating the ~~neigh~~ neighbourhood calculation for each neighbouring point