Printed Pages– 5 Roll No. .....

15.5
3.5
17.8

AS-2361

M. C. A. (Fifth Semester) Examination, 2013

DATA MINING

Paper : Third

Time Allowed : Three hours

Maximum Marks : 60

Note : *Attempt five questions in all. Q. No. 1 is compulsory. Answer any four from rest.*

1. (i) What are the steps involved in data mining when viewed as a process of knowledge discovery? 2

   (ii) What are the other names of data mining? 2

(iii) What are the different operations that can be performed on data cube? 2

(iv) Write an application of outlier analysis. 2

(v) What is evolution analysis? Write an applications of it. 2

(vi) What are the types of data mining systems? 2

(vii) How you will identify positively skewed and negatively skewed data? 2

(viii) Define Q1, Q3 and IQR. 2

(ix) How a q-plot is different from Box plots? 2

(x) Draw a 2 : 3 : 1 neural network structure. 2

2. (a) Briefly describe the following advanced database systems and applications : object-relational databases, spatial databases, text databases, multimedia databases, the www. 5

(b) Define each of the data mining functionalities briefly. 5

3. Suppose that a data warehouse consists of the three dimensions time, doctor and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. 4+4+2

   (a) Draw a star schema diagram for the above data warehouse.

   (b) Starting with the base cuboid (day, doctor, patient), what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004.

   (c) To obtain the same list, write the SQL query assuming the data is stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge).

4. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 5×2=10

   13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

   (a) What is the mean of the data? What is the median?

   (b) What is the mode of the data? Comment on the data's modality.

(c) What is the midrange of the data?

(d) Find the first quartile (Q1) and third quartile (Q3) of the data.

(e) Show a boxplot of the data.

5. Suppose a hospital tested the age and body for data for 18 randomly selected adults with the following result. 10

Age	23	23	27	27	39	41	47	49	50
%fat	9·5	26·5	7·8	17·8	31·4	25·9	27·4	27·2	31·2
Age	52	54	54	56	57	58	58	60	61
% fat	34·6	42·5	28·8	33·4	30·2	34·1	32·9	41·2	35·7

Normalize the two variables based on z-score normalization.

6. (a) What is data transformation? Why it is required? What are the different strategy use for data transformation? 7

(b) Suppose the minimum and maximum values for the attribute income are \$12000 and \$98000 respectively. We would like to map income to the range [0·0 1·0]. What is the normalized value of \$73000 using min-max method? 3

7. (a) What is Discrete Wavelet Transform (DWT)? Write the steps of dimensionality reduction using Discrete Wavelet transform. 7

(b) How it is better than DFT? 3

8. (a) Write the step-by-step training procedure of an ANN-GA model for classification. 7

(b) How we can validate the ANN-GA model for classification? 3

1(i) The steps involved in data mining when viewed as a process of knowledge discovery are (a) Data cleaning
(b) Data Integration
(c) Data selection
(d) Data transformation
(e) Data mining
(f) pattern evaluation
(g) Knowledge presentation.

(ii) The other names of data mining are →
Knowledge discovery
knowledge extraction
pattern analysis
Data archaeology

(iii) The operations that can be performed on data cube are
Roll up
Drill down
slice and dice
pivot (rotate)

(iv) outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account.

(v) Evolution analysis describes and models regularities or trends for objects whose behavior changes over time.

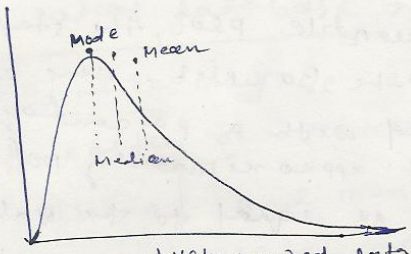
Application → Data mining study of stock exchange data may identify stock evolution regularities for stocks. Such regularities may help predict future trends in stock market prices, contributing to decision making regarding stock investments.

(vi) Data mining systems can be categorized according to various criteria, as follows.

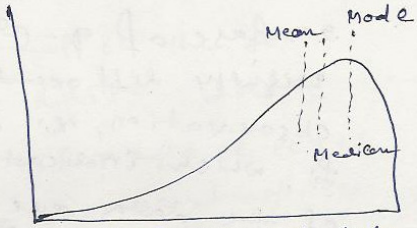
Database technology.
Information science.
Statistics.
Machine learning
Visualization
other disciplines.

(vii) In case of positively skewed data, the mode occurs at a value that is smaller than the median.

In case of negatively skewed, the mode occurs at a value greater than the median.



+vely skewed data



-vely skewed data.

(viii) Q_1 denotes first quartile. It is 25th percentile. The Kth percentile of a set of data in numerical order is the value xi having the property that K percent of the data entries lie are below xi.

Q_3 denotes the third quartile.

IQR is called the interquartile range defined as

$$IQR = Q_3 - Q_1$$

(ix) A box plot incorporates the following.

(a) the ends of the box are at the quartiles so that the box length is the interquartile range.

(b) The median is marked by a line within the box.

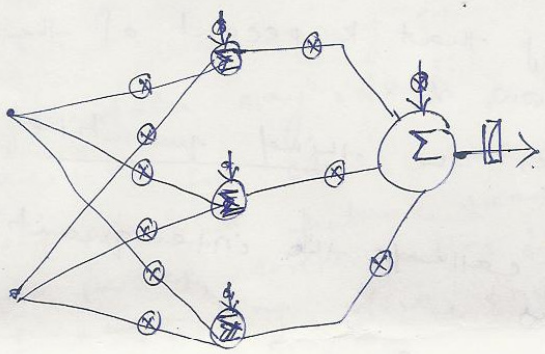
(c) Two lines outside the box extend to the smallest and largest observations.

on case of q-plot or quantile plot, the step is slightly different from the Boxplot - Here each observation, x_i is paired with a percentage f_i, which indicates that approximately $100 f_i\%$ of the data are below or equal to the value x_i. f_i is defined as -

$$f_i = \frac{i - 0.5}{N}, \quad N = \text{total no. of data}$$

on a quantile plot, x_i is graphed against f_i.

(X)



Q.2 (a)

An Objected-oriented database - is designed based on the object-oriented programming paradigm where data are a large number of objects organized into classes and class hierarchies. Each entity in the database is considered as an object. The object contains a set of variables that describe the project, a set of messages that the object can use to communicate with other objects or with the rest of the database system, and a set of methods where each method holds the code to implement a message.

A spatial database → contains spatial-related data, which may be represented in the form of raster or vector data. Raster data consists of n-dimensional bit maps or pixel maps and vector data are represented by lines, points, polygons or other kinds of processed primitives. Some examples of spatial databases include geographical databases, VLSI chip designs, and medical and satelite images databases.

A text database → is a database that contains text documents or other word descriptions in the form of long sentences or paragraphs, such as product specifications, error or bug reports, warning messages, summary reports, notes or other documents.

A multimedia database → stores images, audio and video data and is used in applications such as picture content-based retrieval, voice mail systems, video-on-demand systems, the WWW and speech based user interfaces.

The World Wide Web → provides rich, world-wide, on-line information services, where data objects are linked together to facilitate interactive access. Some examples of distributed information services associated with the WWW include America online, Yahoo, Alta Vista and prodigy.

(b) (i) Characterization → is a summarization of the general characteristics or features of a target class of data.

(ii) Discrimination → is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.

(iii) Association – is the discovery of association rules showing attribute value conditions that occur frequently together in a given set of data.

(iv) classification differs from prediction in that the former constructs a set of models that describe and distinguish data classes or concepts, whereas the latter builds a model to predict some missing or unavailable and often numerical data values. Their similarity is that they are both tools for prediction. classification is used for predicting the class label of data objects and prediction is typically used for predicting missing numerical data values.

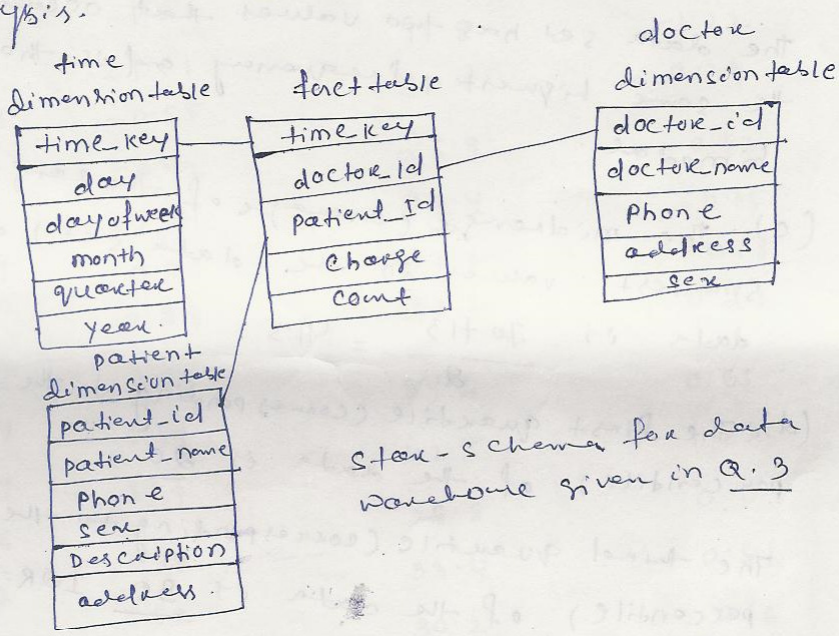
(v) clustering → analyzes data objects without consulting a known class label. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.

Each cluster that is formed can be viewed as a class of objects.

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. The distinct features of such an analysis include time series data analysis, sequence or periodicity pattern matching and similarity based data analysis.

Q.3

(a)



Star-schema for data warehouse given in Q.3

(b) The operations to be performed are

(i) Roll up on time from day to year.

(ii) Slice for time equal.

(iii) Roll-up on patient from individual patient to all.

(c) fee(day, month, year, doctor, hospital, patient, count, charge)

Select doctor, SUM(charge)
from fee

where year 2004
group by doctor.

Q.4

(a) mean of the data = $\frac{809}{27} = 30$

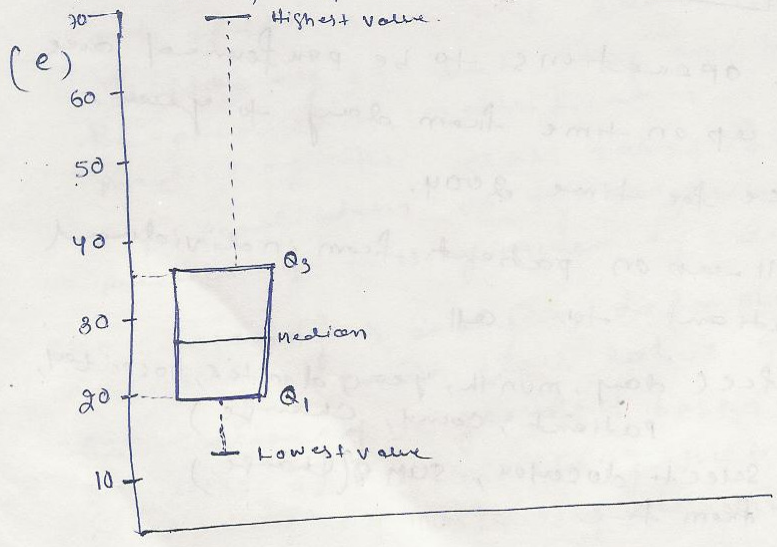
median of the data = 25

(b) The modes of the data are 25 and 35.
The data set has two values that occur with the same highest frequency and is therefore, bimodal.

(c) The midrange (average of the largest and smallest values in the data set) of the data is $\frac{70+13}{2} = 41.5$.

(d) The first quartile (corresponding to the 25th percentile) of the data is 20.
The third quartile (corresponding to the 75th percentile) of the data is 35. IQR = $Q_3 - Q_1$ = 35 − 20 = 15

(e)



Q.5

Z-score normalization, $v' = \frac{v - \bar{A}}{\sigma_A}$

where $\bar{A}$ and σ_A are the mean and standard deviation respectively.

age	z-score normalization	% fat	z-score normalization
23	-1.83	9.5	-2.14
23	-1.83	26.5	-0.25
27	-1.51	7.8	-2.33
27	-1.51	17.8.	-1.22
39	-0.58	31.4	0.29
41	-0.42	25.9	-0.32
47	0.04	27.4	-0.15
49	0.20	27.2	-0.18
50	0.28	31.2	0.27
52	0.43	34.6	0.65
54	0.59	42.5	1.53
54	0.59	28.8	0.0
56	0.74	33.4	0.51
57	0.82	30.2	0.16
58	0.90	34.1	0.59
58	0.90	32.9	0.46
60	1.06	41.2	1.38
61	1.13	35.7	0.77

$\bar{A}_{age.} = 46.44$ $\quad \bar{A}_{\%\,fat} = 28.78$

$\sigma_{age.} = 12.85$ $\quad \sigma_{\%\,fat} = 8.99$

Q.6 (b) ~~$\frac{73600 - 12000}{98000 - 12000}(1.0 - 0) + 0 = 0.716$~~

(b) $v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$

$= \frac{73600 - 12000}{98000 - 12000}(1.0 - 0) + 0 = 0.716$

(a) Data transformation → The data are transformed from original state to an appropriate new set of consistent state so that the old value can be identified by the one of the new values. Data transformations are used to correct data inconsistencies. If using the neural network backpropagation algorithm for classification mining, normalizing the input values for each attribute measured in the training tuples will help speed up the learning phase.

The different strategy used for data transformation are -

(1) Smoothing - Remove noise from data using techniques as Binning, Regression clustering -

(2) Aggregation - Summarization, or aggregation option is applied.

(3) Generalization - Hierarchies concept applied

(4) Normalization – Attribute values are scaled to fall within a small specified range as -1.0 to 1.0 or 0 to 1.0

(5) Attribute / feature construction — New attributes are constructed and added to existing attributes to help mining process.

7(a) Discrete wavelet transform is a linear signal processing technique, when applied to a data vector X, transforms it to a different vector X', wavelet coefficient.

Steps of dimensionality reduction

(1) The length L, of the input data vector must be an integer power of 2. This condition can be met by padding the data vector with zeros as necessary.

(2) Each transform involves applying two functions. The first applies some data smoothing, such as sum or weighted average. The second performs a weighted difference which acts to bring out the detailed features of the data.

(3) The two functions are applied to pairs of data points in X, i.e. to all pairs of measurements (x_{2i}, x_{2i+1}). This results in two sets of data of length L/2. In general, these represent a low frequency version of the input data and the high frequency content of it, respectively.

(4) The two functions are recursively applied to the sets of data obtained in the previous loop, until the resulting data sets obtained are of length 2.

(5) selected values from the data sets obtained in the above iterations are designated the wavelet coefficients of the transformed data.

(b) Advantages of DWT over DFT

(1) DWT achieves better lossy compression.

(2) For an equivalent approximation, the DWT requires less space than the DFT.

(3) There is only one DFT, but there are several families of DWT.

(4) Popular wavelet transforms include Haar-2, Daubechies-4, Daubechies-6.

(5) DWT uses a hierarchical pyramid algorithm that halves the data at each iteration resulting in fast computational speed.

Q. 8 (a) Let us take the example of IRIS data classification using neural network. The weights of the neural network model is updated using Genetic algorithm instead of the back propagation algorithm. The step-by-step procedure for training is as follows :—

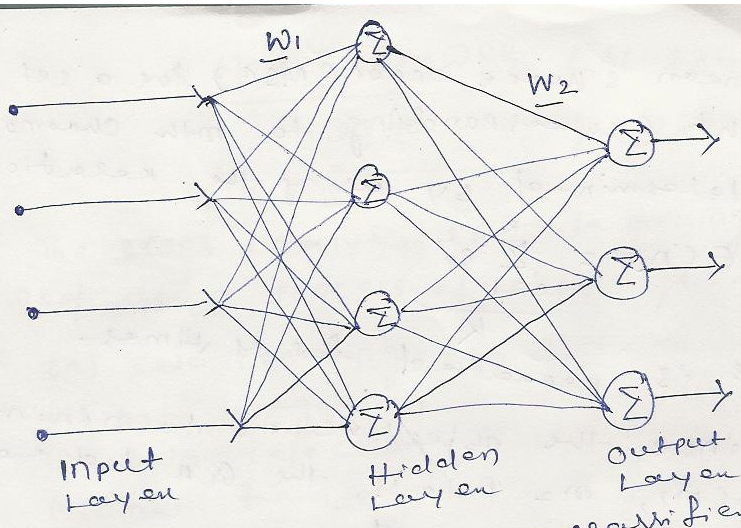


Fig. Neural network classifier for IRIS dataset.

(1) The weights W_1 and W_2 of the classifier are initially chosen from a population of M chromosomes. Each chromosome constitutes NL number of random binary bits, each sequential group of L-bits represent one weight of the classifier, where N is the number of weights of the model.

(2) Since the IRIS data is having 150 tuples, so we have K = 150 input patterns. The number of class is three and each class is having 50 tuples each.

(3) Each of the input pattern is fed to the classifier ~~and the output is obtained~~ using each chromosome as classifier weights and M sets of K outputs are obtained.

(4) Each of the output is compared with corresponding target value and K errors are produced.

The mean square error (MSE) for a set of weight (corresponding to mth chromosome) is determined by using the relation.

$$MSE(m) = \frac{\sum_{i=1}^{K} e_i^2}{K}$$

This is repeated for M times.

(5) since the objective is to minimize MSE(m), m = 1 to M, the GA based optimization is used.

(6) Then crossover and mutation are carried out respectively.

(7) 'selection operator is finally used to select the best M chromosomes.

(8) In each generation the minimum MSE (MMSE) is obtained and plotted against generation to show the learning characteristics.

(9) The learning process is stopped when MMSE reaches the minimum level.

(10) the adaptation is stopped when MMSE level is reached. Each chromosome represents the optimized weights of the classifier.

(b) once the training of the ANN-GA classifier was over, we will get the optimized weights of the classifier. Each chromosome represents the optimized weight vector for the ANN-GA classifier.

(2) The number of tuples, which we kept aside for validation (30 numbers) purpose we'll be used as the input patterns.

(3) First we apply one input pattern to the classifier and calculate the output of the model. In this way we will apply all the 30 patterns to the classifier and calculate the outputs.

(4) Then we will calculate the number of mismatch. For class-1 we have 10 input patterns and same for class-2 and class-3.

When we are giving a input pattern to the classifier, its class label is known to us. So, we will compare the class label given by the ANN-GA classifier and actual class label for that particular input tuple and calculate the no. of mismatch.

(5) Finally we will calculate the percentage of accuracy given by the classifier after calculating the total number of mismatches obtained.