USN									
1	P	E			M	C	A		

PESIT Bangalore South Campus

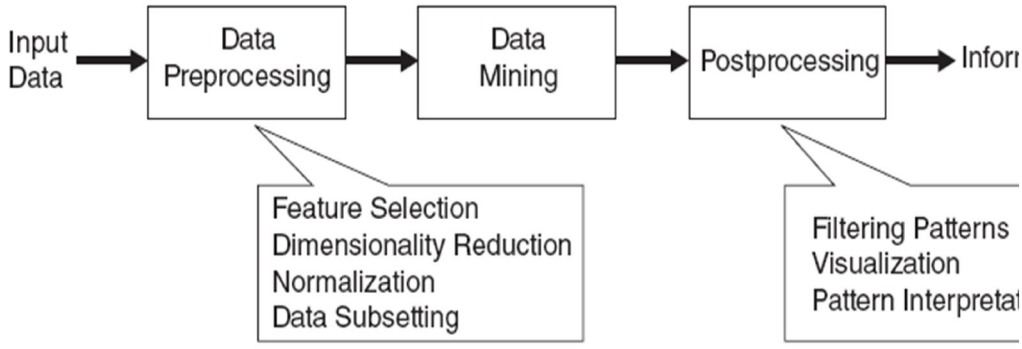
Hosur Road, 1km before Electronic City, Bengaluru -560100

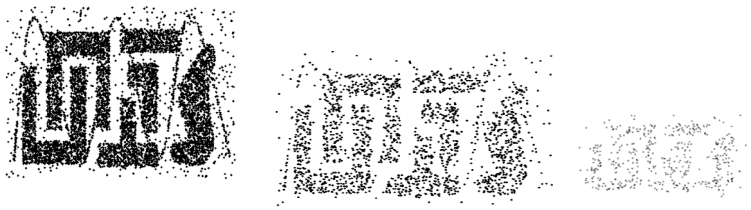
Department of Master of Computer Application

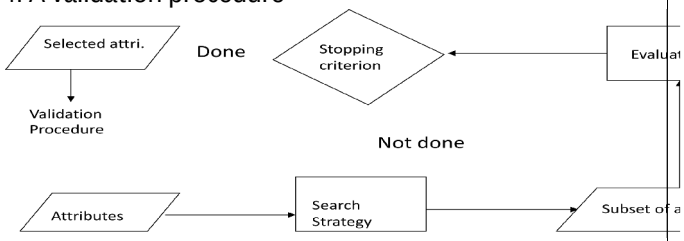
INTERNAL ASSESSMENT TEST I

Date : 19-02-2019	**Max Marks: 40**
Subject & Code: Data Warehousing and Data Mining (17MCA442)	**Sem & Sec: 4th Sem MCA**
Name of Faculty: Dr. Arti Arya	**Time: 11:30AM – 1:00PM**

Note: *Answer FIVE full questions. Select one question from each part.*

	Part I	
Q 1	i) Briefly explain any two motivating challenges in Data Mining Scalability, Heterogeneous and complex data, High dimensional data, Data ownership and distribution, Non-traditional analysis. Any two can be explained. ii)With a block diagram, explain the process of knowledge discovery and data mining.  **Figure 1.1.** The process of knowledge discovery in databases (KDD).	**4** **4**
	OR	
Q 2	What are the different methods of handling missing values in a dataset. Differentiate between noise and outlier. • Handling missing values – Eliminate Data Objects – Estimate Missing Values – Ignore the Missing Value During Analysis – Replace with all possible values (weighted by their probabilities) Noise::Can be wrong measurement. May be distortion of a value. May be addition of spurious objects. Outliers: Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set. These can be meaningful for user	**8**

	sometimes. e.g. fraud detection, Intrusion detection.	
	Part II	
Q 3	Write the formula for Minkowski's distance and show that d(A,B)=size(A-B)+ size(B-A) satisfies the properties of a metric. $dist = (\sum_{k=1}^{n} \| p_k - q_k \|^r)^{\frac{1}{r}}$ Refer classwork for the rest of the ans.	**8**
	OR	
Q 4	What are the various types of sampling? Explain with an example. Give an example to show that an inappropriate size of sample may result in loss pof information. • Sampling is the main technique employed for data selection. • Simple Random Sampling • Random Sampling without replacement • Random Sampling with replacement • Stratified sampling • Progressive sampling *8000 points* *2000 points* *500 points* 	**8**
	Part III	
Q 5	Compute L_1, L_2 and $L_{infinity}$ norm for the following data points: P1(5,0,2,1) and P2(3,4,0,7). Also, Compute similarity between two document vectors d1=(1,0,0,1,2,5) and d2=(4,7,0,0,1,2) using SMC and Jaccard's coefficient. Ans: Use Euclidean distance for L2: $? \sqrt{(5-3)^?} + (-4)^? + 2^? + (-6)^? = 7.74$ L1=14, $L_{infinity}$=6	**8**
	OR	
Q 6	What are the various applications of data mining. Write a short note on 'Curse of dimensionality'and discuss any one way of reducing dimensionality. Applications in the area of Financial Analysis, Fraud detection, Intrusion detection, Predicting stock prices, predicting health concerns, Market basket analysis etc. • When dimensionality increases, data becomes increasingly sparse in the space that it occupies. That is for classification, there are not enough data points for creation of a model that confidently assigns a class to all possible objects. Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful. • PCA is a linear algebra technique for cont. attributes that finds new attributes called as Principal components that are Are linear combination of original attr.Orthogonal to each other.Captures max. variation in the data.SVD- also a DR technique related to PCA	**8**

	Part IV	
Q 7	Give an architecture for feature subset selection with diagram. – 1. A search strategy that generates new subsets of feature. – 2. A measure for evaluating a subset – 3. A stopping crieteria – 4. A validation procedure 	**8**
	OR	
Q 8	Differentiate between i)Discretization and Binarization with an example ii) Feature extraction and feature construction. Binarization::Both *cont.* and discrete attr. may be transformed to binary attr. Discretization::Transformation of a continuous attribute into categorical attribute…. Feature Extraction::The creation of new set of features from the original raw data.The feature extraction process results in a much smaller and richer set of attributes. highly domain-specific. The techniques for FE, developed for one field are often not applicable to other fields. Construction:: Assuming there are n features A1, A2,..... An. After feature construction we may have additional m features An+1, An+2,.... An+m. All new constructed features are defined in terms of original features as such no inherently new information is added through feature construction	**4+ 4**
	Part V	
Q 9	Compute the correlation coeffecient between two variables x and y defined as x=(3,6,0,3,6) and (1,2,2,4,5) and interpret the result. $Corr(x,y)=\frac{\text{cov}(x,y)}{sd(x).sd(y)}=\frac{s_{xy}}{s_x s_y}$ $\text{cov}(x,y)=s_{xy}=\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})$ $sd_x=s_x=\sqrt{(\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2)}$ $sd_y=s_y=\sqrt{(\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\bar{y})^2}$ $\bar{x}=\frac{1}{n}\sum_{k=1}^{n}x_k$ $\bar{y}=\frac{1}{n}\sum_{k=1}^{n}y_k$	**8**

	OR	
Q 1 0	What are the different data mining tasks? Explain with brief examples. • Prediction Methods ◦ Use some variables to predict unknown or future values of other variables. ◦ The attribute to be predicted- target/dependent variable. ◦ Independent/explanatory variable- used for making predictions. • Description Methods ◦ Derive human-interpretable patterns that summarize the underlying relationships in data. ◦ These are exploratory in nature & requires post processing techniques	**8**