P Sai Varun
17241A05A4
DWDM

MID-II

1. The Decision tree induction algorithm is called with three parameters : Data partition D, attribute_list and Attribute_selection_method.
- Initially, it is the complete set of training tuples and their associated class labels. The parameter attribute list is a list of attributes describing the tuples.
- Attributes_selection_method specifies a heuristic procedure for selecting the attribute that "best" descriminates the given tuples according to the class.
- This procedure employs an attribute selection measure such as information gain & the Gini index.
- whether the tree is strictly binary is generally driven by the attribute selection measure. Some attribute selection measures such as the Gini index enforce the resulting tree to be binary
- The tree starts as a single Node N, representing the training tuples in D
- If the tuples in D are all of the same class, then node N becomes a leaf and is labeled with that tal class

Note that steps 4 and 5 are terminating conditions. All terminating cond are explained at the end of algorithm.

- otherwise the algorithm calls attribute selection method to determine the splitting criterion to separate or partition the tuples in D into individual class (step 6)

- The node N is labeled with the splitting criterion which serves as a test at the node (S-7) A branch is grown from Node N for each of the outcomes of the spitting criterion. The tuples in D are partitioned accordingly (S-10 to 11)

- There are three possible scenarios (a) If A is discrete-valued then one branch is grown for each known value of A

b) If A is continous-valued then two branches are grown corresponding to A <= split point and A > split point

c) If A is discret valued and a binary tree must be produced then the test is of the form $A \in S_A$ where S_A is the splitting subset of A

- The Algo uses the same process recursively to form a decision tree for the tuples at each resulting partition, D_j of D (S-14)

- The recursive partitioning stops only when any one of the following terminating cond is satisfied.

1. All the tuples in partition D belong to the same class

2. There are no remaining attributes on which the tuples may be further partitioned.

3. There are no tuples for a given branch that is a partit D_j is empty. In this leaf is created with the majorityclass

3. i) Spatial Data mining:

A spatial Database stores a large amount of space related data such as maps preprocessed remote scanning or medical imagining data and VLSI chip layout data. spatial databases have many features distinguishing them from relational d.b They carry topological or distance information, usually organized by sophisticated, multidimensional spatial indexing structures that are accessed by spatial data acess methods and often require spatial reasoning, geometric computation and spatial knowledge representation techniques

spatial data mining refers to the extraction of knowledge spatial relationships or other intresting points patterns not explicitly stored in spatial databases such mining depends on integration of data mining with spatial db technologies It can be used in understanding spatial data, discovering spatial relationships and relationships between spatial and non spatial data, constructing spatial knowledge, optimizing queries

Text-mining

It is also referred as text data mining roughly equivalent to text analytics, is the process of deriving high quality information from text The definition strikes at the primary chord of text mining - to delve into unstructured data to extract meaningfull patterns and insights required for exploring textual data sources

Text mining incorporates and integrates the tool of info retrieval data mining, machine learning, statistics and computational linguistics and hence it is nothing short of a multidisciplinary field. Text mining deals with natural language texts either in semi structured or unstructured.

Text mining techniques

1. Information Extraction
2. Information retrieval
3. categorization
4. clustering
5. summarization

Part - B

1. B
2. D
3. C
4. A
5. B
6. C
7. D
8. D
9. A
10. A