Data Mining

By

[NIKUNJ K. DESAI]
[16BCE033]

DEPARTMENT OF COMPUETR ENGINEERING
Ahmedabad 382481

[Data Mining]

Seminar

Submitted in fulfillment of the requirements

For the degree of

Bachelor of Technology in Computer Engineering

By

[NIKUNJ K. DESAI]
[16BCE033]

Guided By
[PROF. VISHAL PARIKH]

[DEPARTMENT OF COMPUTER ENGINEERING]

DEPARTMENT OF COMPUETR ENGINEERING
Ahmedabad 382481

CERTIFICATE

This is to certify that the Seminar entitled "**Data Mining**" submitted by **Nikunj K. Desai (16BCE033)**, towards the partial fulfillment of the requirements for the degree of Bachelor of Technology in Computer Engineering of Nirma University is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination.

Prof. Vishal Parikh
Department of Computer Engineering,
Institute of Technology,
Nirma University,
Ahmedabad

Dr. Sanjay Garg
Department of Computer Engineering,
Institute of Technology,
Nirma University,
Ahmedabad

ACKNOWLEDGEMENT

This data mining study would not have been possible without the kind support and help of many individuals. I would like to say thanks to all of them.

I am highly thankful to Prof. Vishal Parikh for his guidance and continuous supervision as well as for providing essential information regarding the seminar topic and also for his support in completing data mining study.

I would like to express my thankfulness towards my parents and member of Department of Computer Science & Engineering for their kind co-operation and encouragement which help me in completion of this seminar study.

I would like to express my special gratitude and thanks to my friends who helped me in exploring my seminar study.

ABSTRACT

This chapter introduces basic concepts of the data mining which is generally known as knowledge discovery from data (KDD). In the knowledge discovery process, basically involves pattern recognition, data selection, data integration etc. This chapter basically focuses on the pattern recognition in various types of data collected from various sources in which data mining techniques used also this chapter focuses on the various types of data mining techniques which are used to find pattern in the bunch of data collection. Also I will introduce to reader about what is data warehouse and the difference between database and data warehouse. Finally my study on data mining is outlined.

CONTENTS

- **List of Figures**

- **List of Table**

Chapter 1
Introduction

1.1 General:

Before you start reading this study paper the first question is "What is Data mining?" so data mining refers to Extracting data from large amount of data and also referred as knowledge discovery in database (KDD). It is a process of discovering interesting knowledge from large amount of data stored either in database or other data repositories.

2.2 Objective of data mining:

We know that there is a large amount of data falling over computers day to day. There is a large amount of data daily produce by different agencies, different institutes, many other business companies and industries collect large amount of data daily, but in fact we know that we need only small amount of data from large amount of data to figured out which data is useful for that we need particular data structure or techniques. Manually extract meaningful data is next to impossible so that we required some algorithms and techniques that can be implemented using computer so that abstraction of data can be done efficiently and faster so that this process of abstraction of data is known as data mining.

1.3 Categories of data mining:

There are two primary categories of data mining.

(i) Prediction: - Predictive data mining creates the model from given data set and using that model using any data mining techniques and using that model it only predicts the value or result for the unknown variables.

(ii) Description: - Descriptive data mining creates information using the available data set and finds patterns from the data set and that pattern can be very useful for human to interpret.

1.4 Applications:-

There is infinite number of application of data mining but here is the list of industries in which data mining is widely used.

- Marketing
- Education
- Agriculture
- Healthcare
- DBMS
- Software Engineering
- Finance
- Research analysis

Chapter 2
DBMS and Data warehouse

2.1 What is Database?

Before learning what is database first understand that what is data? Data is a relative fact that has future reference value. Many business companies, government and scientific institutes etc. produce large amount of data daily so that they need to store and manage that data, so managing and storing of data is generally known as database. There are various types of database management techniques on the bases of how we store the data but nowadays there is most popular and most efficient database management system or technique is Relational Database Management System (RDBMS). In RDBMS we store data in the form of rows and columns. MySQL, Oracle servers are well known DBMS.

2.2 What is Data Warehouse?

Data warehouse is just like a storage room in the house. Data warehouse stores historical data. Data warehouse also known as the central database which consist data of many databases. Data warehouse also known as the server for many databases which are connected to one data warehouse. There are basically three stages of data warehouse.

(i) First stage for store raw data which is used by developers for data analysis.

(ii) Second stage is basically an integration stage it provides data abstraction to the user.

(iii) Third stage is the access stage which provides functionality to user to access the data and use that data from data warehouse.

2.3 Difference between DBMS and Data warehouse:

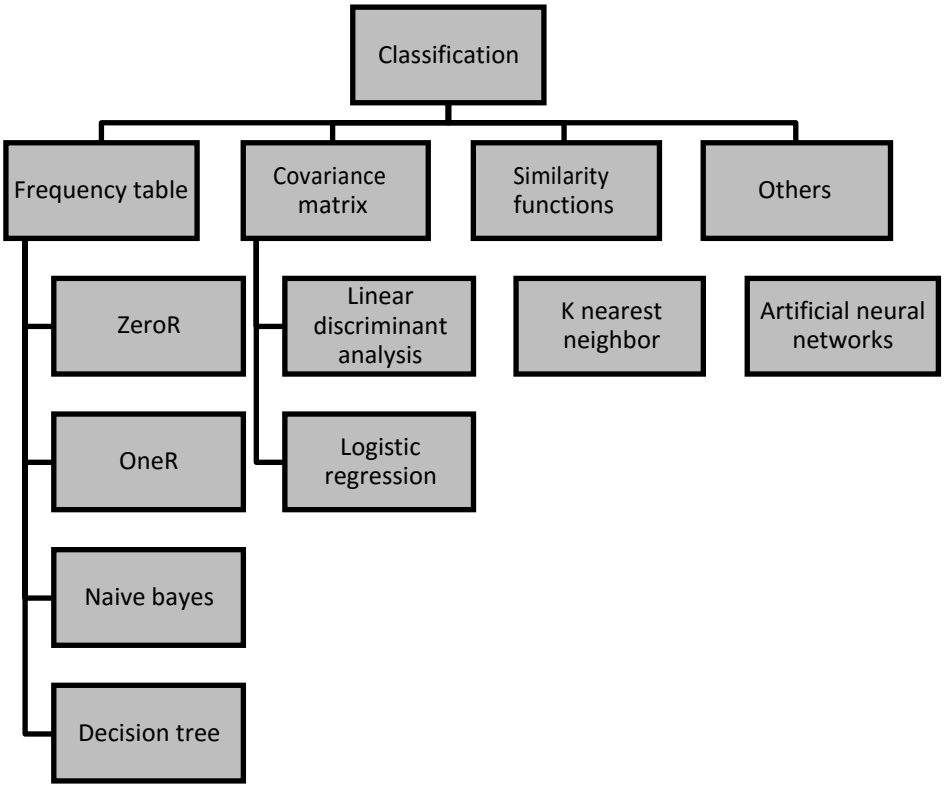
DBMS	Data Warehouse
Stores current data.	Stores historical data.
Often changes due to frequency update done on it so cannot be used for analysis.	Extract data and reports them to analyze and reach decisions.
Uses for online transactional processing.	Uses for online analytical processing.
Contains highly detailed data.	Contains summarized data.
Provides a detailed relational view.	Provides a summarized multidimensional view.

Chapter 3
Data Mining Techniques

Basically there are only four techniques of data mining, **classification, association rule discovery, clustering** and **regression.** Let us discuss all the techniques one by one.

3.1 Classification

Classification is a data mining task of predicting a certain outcomes based on given input. Classification algorithm tries to discover relationships between the attributes that would make it possible to predicting the output. Classification algorithm chart is given below.



3.1.1 Frequency Table:

- **ZeroR**

ZeroR stands for Zero-Rule. ZeroR is a very basic method to predict the output and it is not very useful. As a name suggest ZeroR data mining techniques has no rule to consider for predicting the future value of unknown variable.

- **Working:-** Construct frequency table and choose the most frequent value or the value of the final result for the given dataset and ignore all the predictors or attributes given in dataset only consider the final output.

Example:-

Day	Outlook	Temperature	Humidity	Wind	Play ball
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Table 3.1

- From the above frequency table we have decide that play or not play.
- In ZeroR method ignore all the parameters just consider only final prediction value for all the results.
- In the above frequency table we have 9 Yes values while 5 No values so the positive predictive value will be 9/14 = 0.64 while negative predictive value will be 5/14 = 0.35 < 0.64 so we conclude that given ZeroR model for given dataset is “play ball = Yes” with an accuracy of 0.64.

- **OneR**

OneR stands for One-Rule. OneR is also a simple classification algorithm but yet it is also accurate algorithm than ZeroR classification algorithm.

  - **Working:-**
    - Generate one rule for each attribute and select the attribute as a rule which has less error.
    - Using the frequency table create a rule for a predictor.

Example:-

In figure 3.1 find out which one is the best predictor?

Solution: -

Outlook	Sunny	Overcast	Rainy
Yes	2	4	3
No	3	0	2
Table 3.2			

OneR	Yes	No	Predictive value
Yes	7	2	Positive = 0.78
No	2	3	Negative = 0.60
Table 3.3			

Temperature	Hot	Mild	Cool
Yes	2	4	3
No	2	2	1
Table 3.4			

OneR	Yes	No	Predictive Value
Yes	9	0	Positive = 1
No	0	5	Negative = 1
Table 3.5			

Humidity	High	Normal
Yes	3	4
No	6	1
Table 3.6		

OneR	Yes	No	Predictive Value
Yes	6	3	Positive = 0.67
No	1	4	Negative = 0.80
Table 3.7			

Windy	False	True
Yes	6	3
No	2	3
Table 3.8		

OneR	Yes	No	Predictive Value
Yes	6	0	Positive = 1
No	3	3	Negative = 1
Table 3.9			

From the above analysis we found that Positive predictive value is greater than Negative Predictive Value happens in only outlook attribute so set outlook as one rule and according to that rule.

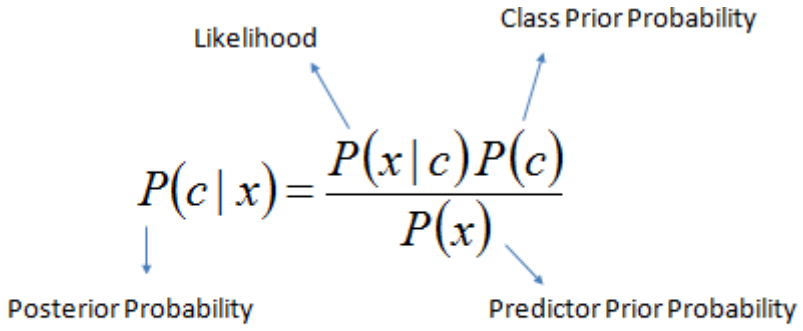
- If outlook is sunny than play : Yes
  If outlook is overcast than play : Yes
  If outlook is rainy than play : No

So using above analysis we can predict the value for any unknown variable.

- **Naïve Bayes:**

  This classification method basically based on the Bayes theorem with independence assumption between predictors. This classification method is not very complicated because it has no iterative parameters and we get very accurate result too so that this method used to predict the future value of unknown variable for very large datasets.

  - **Working:-**



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 3.1

  - Transform the given frequency table (figure 3.1) into likelihood tables and then using the naïve Bayes equation given in figure 3.2 calculate P(c/x) for each class. Than the class with higher P(c/x) will be the outcome of the prediction.

Example:-

Using the table given in figure 3.1 find the play ball value for sample <Sunny, Hot, Normal, Strong>.

Solution: - Assume that x_1 = Yes and x_2 = No.

So using table 3.1 assume that c_1 = sunny, c_2 = overcast, c_3 = rainy, c_4 = Hot, c_5 = Mild, c_6 = Cool, c_7 = High, c_8 = Normal, c_9 = Strong, c_{10} = Weak

I	J	$P(c_i/x_j) = P(x_j/c_i)*P(c_i)/P(x_j)$
1	1	3/9
	2	2/5
2	1	4/9
	2	0/5
3	1	3/9
	2	2/5
4	1	2/9
	2	2/5
5	1	4/9
	2	2/5
6	1	3/9
	2	1/5
7	1	3/9
	2	4/5
8	1	6/9
	2	1/5
9	1	3/9
	2	3/5
10	1	6/9
	2	2/5

Table 3.10

P(Yes1) = P(Sunny/Yes)*P(Hot/Yes)*P(Normal/Yes)*P(Strong/Yes)*P(Yes)

= (2/9)*(2/9)*(6/9)*(3/9)*(9/14)

= 0.007054

P(No1) = P(Sunny/No)*P(Hot/No)*P(Normal/No)*P(Strong/No)*P(No)

= (3/5)*(2/5)*(1/5)*(3/5)*(5/14)

= 0.010285

Here we can clearly see that P(No1) > P(Yes1), so according to naïve bayes theorem for given dataset <Sunny, Hot, Normal, Strong> our result play ball = NO.

- **Decision Tree:**

  Decision tree is widely used and efficient classification model nowadays. As a name suggest decision tree classification model is in the form of tree data structure.

  Working:-

  - Decision tree partitions the whole frequency table into smaller frequency table (subsets) using different values of predictors like entropy, gain and information gain (If you are using ID3 algorithm). After separating whole frequency table using one predictor and at that time build up a decision tree step by step.

  Algorithm:-

  - If you are using ID3 algorithm than we need entropy and information gain. Before starting of the algorithm first we need to know that what is entropy, what is information gain and what is gain for given dataset and how to calculate these parameters.
  - **Information Gain:-**

    Basically information gain is calculated for every segment of the frequency table. Using information gain we can find out that which attribute value will be the root of the decision tree because we are using the ID3 algorithm so that we have to create decision tree incrementally means step by step. Formula to find information gain is given below.

$$\mathbf{I(P,Q)} = \frac{-\boldsymbol{P}}{\boldsymbol{P+Q}}\boldsymbol{log2}\left(\frac{\boldsymbol{P}}{\boldsymbol{P+Q}}\right) - \frac{\boldsymbol{Q}}{(\boldsymbol{P+Q})}\boldsymbol{log2}\left(\frac{\boldsymbol{Q}}{\boldsymbol{P+Q}}\right) \ldots Eq[3.1]$$

    Where, P = number of Yes results from given dataset

    Q = number of No results form given dataset

  - **Entropy:-**

    Entropy is nothing but it is a parameter which finds the object from the given dataset with similar values (homogenous). Maximum value of the entropy will be 1 when the sample dataset is similarly separated while minimum value will be 0 when the sample dataset is completely homogeneous. So the value of entropy varies between 0 and 1.

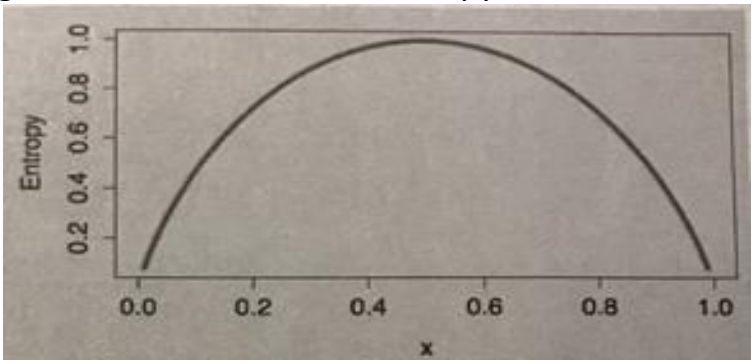


Figure 3.2

Figure 3.2 show that how entropy is changes with the value of x.

Formula to find entropy is given below.

$$\boldsymbol{Entropy\ E} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \frac{\boldsymbol{Pi + Qi}}{\boldsymbol{P + Q}} * \boldsymbol{I(P, Q)} \ldots Eq[3.2]$$

- **Gain:-**

  Gain of the information is nothing but the difference between information gain and entropy so the formula for the gain is given below.

$$\boldsymbol{G(Gain) = I(P, Q) - E} \ldots Eq[3.3]$$

- After learning all three parameters of the ID3 algorithm let's start the algorithm. First of all from the given dataset we need to find that which attribute will be root of decision tree classification model for that first of all find the information gain of the given dataset using equation 3.1 after that find the entropy for every attributes using equation 3.2 and finally calculate the gain for every attributes. From all the attributes one attribute which has highest gain will be the root of the decision tree classification model and continue this process recursively until every attributes individual parameters are added in to decision tree.

Example:-

Find the decision for set <sunny, high, normal, weak> for given dataset in figure 3.1.

Solution:-

For the whole table P = 9 and Q = 5 so using equation 3.1 we will get ( ) = 0.940

Outlook	Sunny	Overcast	Rainy	Entropy	Gain
Yes	2	4	3	0.692	0.248
No	3	0	2		
$I(P, Q)$	0.970	0	0.970		

Temperature	Hot	Mild	Cool	Entropy	Gain
Yes	2	4	3	0.911	0.029
No	2	2	1		
$I(P, Q)$	1	0.918	0.811		

Humidity	High	Normal	Entropy	Gain
Yes	3	4	0.789	0.151
No	6	1		
$I(P, Q)$	0.918	0.721		

Windy	False	True	Entropy	Gain
Yes	6	3	0.892	0.048
No	2	3		
$I(P,Q)$	0.811	1		

From the above 4 tables we say that G(Outlook)>G(Humidity) G(Wind)>G(Temperature) so that the root of the decision tree classification model will be Outlook and nodes of the root will be attributes of the Outlook like in our example nodes of the root will be Sunny, Hot and Overcast. Now assume that Sunny is root to continue build the decision tree so that make a separate frequency table in which all the outlook attribute will be only Sunny as shown in the below figure.

Day	Outlook	Temperature	Humidity	Wind	Play ball
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Table 3.11

Now using above table as a frequency table and recursively do the process which we have done for the whole dataset so that after performing recursive process we will get the following values.

- $G(T_{Sunny}$, Temperature) = 0.57
- $G(T_{Sunny}$, Humidity) = 0.970
- $G(T_{Sunny}$, Wind) = 0.02

So from the above values $G(T_{Sunny}$, Humidity)> $G(T_{Sunny}$, Temperature)> $G(T_{Sunny}$, Wind)
So that Humidity will be the node of the Sunny node and the nodes of the Humidity root will be High and Normal. From the given frequency table of Sunny we can clearly say that node of the High humidity will be No while node of the low humidity will be Yes. When we create the separate table for the overcast attribute of the outlook than result is always Yes (from figure 3.1) so the node for the overcast will be Yes. Now separate the dataset using rain as the outlook as shown in the figure below.

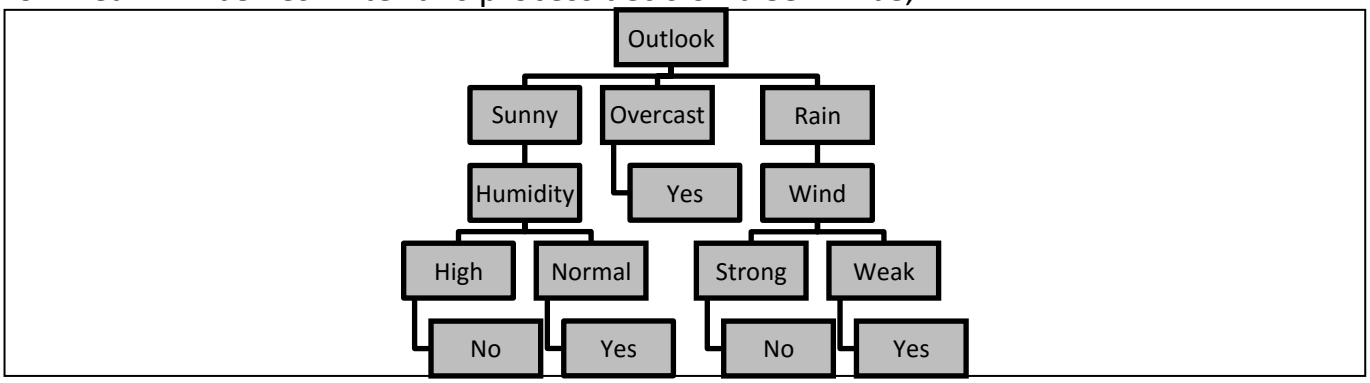
Day	Outlook	Temperature	Humidity	Wind	Play ball
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Table 3.12

Now using above table as a frequency table and recursively do the process which we have done for the whole dataset and ignore the humidity parameter for above figure because we have already used it in our model so that after performing recursive process we will get the following values.

- $G(T_{Rain}, Wind) = 0.970$
- $G(T_{Rain}, Temperature) = 0.019$

So from the above values $G(T_{Rain}, Wind) > G(T_{Rain}, Temperature)$ so that the node of the rain root will be Wind and the children node for Wind will be Strong and Weak from figure 3. We can clearly say that children node of Strong will be No while children node for Weak will be Yes. After this process decision tree will be,



So from the above decision tree we can make the decisions for the dataset which are not given initially like for our question we have find the decision for <sunny, high, normal, weak> dataset to from the tree it is clear that the result will be Yes.

3.1.2 Covariance Matrix:-

- **Linear Discriminant Analysis:**

  As a name suggests linear discriminant analysis (LDA) method uses the concept of linear equations. LDA creates a linear equation from given attributes or predictors and find the linear combinations between the variables.

- **Logistic Regression:**

Logistic regression model also predicts the result or probability of the result using 2 values. Logistic regression basically a curve and the values of this curve are basically between 0 and 1. Generally LDA and logistic regression both are similar.

3.1.3 Similarity Functions:-

- **K Nearest Neighbor:**

  K nearest neighbor algorithm or classification method contains the whole data and analyzing pattern from that stored data it gives the appropriate output. This algorithm is basically depends on the votes of the neighbor of the predictor. KNN is measured using different 3 types of distance functions as given below.

  - Euclidean distance
  - Manhattan distance
  - Minkowski distance

  Above 3 types of distance functions are used while we have continuous values of the variables in the dataset.

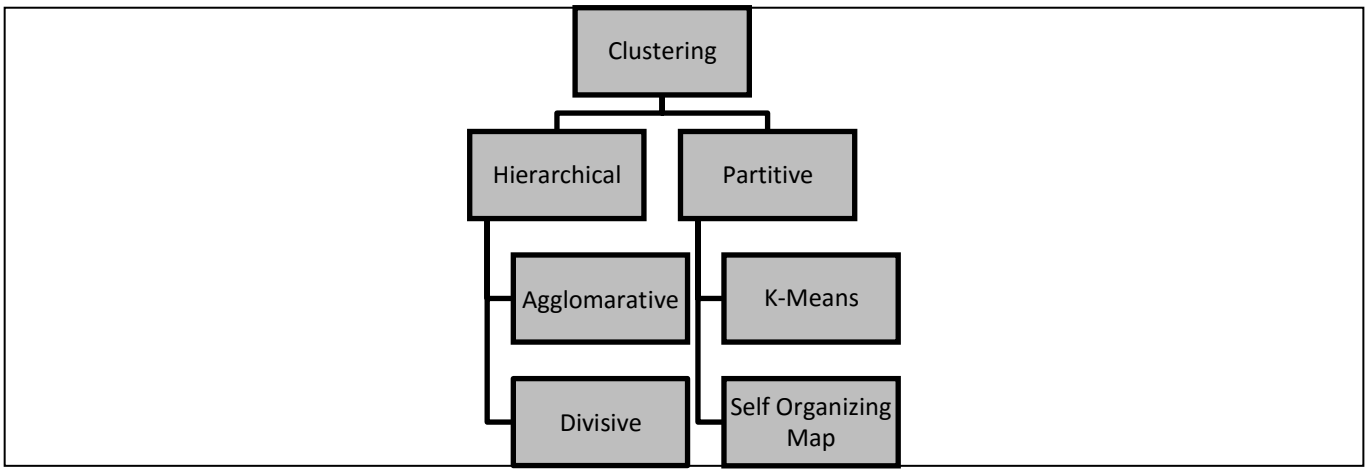
3.1.4 Others:-

- **Artificial Neural Network:**

  Nowadays we know that the technology which will completely change the world is the Artificial Intelligence. So using the basic concept of the artificial intelligence we can predict the value for the unknown variable and predicts the future values. ANN uses its previous results and uses that results analyze those results and finally gives use the output so for ANN we do not need to maintain an proper algorithm because artificial intelligence predicts the values using its own artificial brain.

3.2 Clustering

Cluster is nothing but the group. One cluster contains all the variables which have same patterns or nearly equal values. The process of finding clusters in the given dataset is known as the clustering. There are basically two type of the clustering as given below.



3.3 Association Rules:

In data mining association rules are used for finding the patterns from the frequent behavior of certain things. Basically association rule statements are of **if…than** type statements for example after observing the frequent behavior of many people that if a person by bananas than out of those people 85% of people likes to buy milk so this will become a pattern that "**if** a person by bananas **than** it will buy milk". From this example we can simply say that association rule is used for finding frequent patterns and analyzing the behavior of the customer. So association rule plays a major role in marketing business.

Chapter 4
Summary and Conclusion

4.1 Summary

Data mining is helping every organization to find the hidden pattern in the large amount of data which is collected by organization from various sources and that can be useful for the prediction of the customer's behavior, products and processes.

Using different data mining techniques we can only build different models of the data set to predict the future value but building of the model is just a one step towards KDD. We need to find the most efficient and faster models for that we have to create model using different techniques and algorithms and from that we can find the perfect model for such type of data.

4.2 Conclusion

The goal of data mining as seminar study is that students can know that how data mining models are useful and developed and capable of predicting the chances of the employment for him/her to choose a particular branch for his/her engineering studies. After studying all the data mining techniques I can clearly say that some classification models will be considered for study like decision tree, Naïve Bayes classifier and neural networks.

References

1. **“**Introduction to data mining**”** by Tan, Steinbach and Kumar (2006)
2. “Data Mining: Concepts and Techniques”, Third Edition by Han, Kamber and Pei (2013)