International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-5)

Research Article | **May 2015**

Discovering Disease Pattern in Hospital Data Analysis

Monika Devi
Research Scholar, Department of CSE,
Doon Valley Institute of Engineering &Technology,
Karnal, Haryana, India

Dinesh Kumar
Assistant Professor, Department of CSE,
Doon Valley Institute of Engineering & Technology
Karnal, Haryana, India

Abstract:

This paper presents an experiment to discover disease patterns by using a statistical approach on hospital database. This study, help to know about the number of patient admitted every year suffered from various diseases. By using Data mining system, it helps to discovers patterns and relationships hidden in data, the system is actually is a part of a larger process called "knowledge discovery" which describes the steps that must be taken to ensure meaningful results. The presented work focus on implementing different data mining approaches on hospital database which is collected from PGIMR annual report. This paper combines two major approaches to provide profiling of patients and discovering disease pattern via clustering. According to defined approach, clustering is implemented to profile patients according to their month of admitted in hospital. It tells that how many patients are admitted in which month segment. In this work, a number of clusters are formed on the basis of type of disease acquired by patient. Patients are grouped into different clusters according to their disease.

Key Words: Patients, Disease,Disease pattern, Data mining, Clustering, Segments, Weka tool.

I. INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.[1].The primary goals of Data Mining in practice tend to be Prediction and

Description [2] and [3]. Data Mining is an analytic process designed to explore data in search of consistentpatterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns [4].

A **hospital** is a health care institution providing patient treatment with specialized staff and equipments. The best known type of hospital is the general hospital, which has an emergency department.

A **disease** or **medical condition** is an unhealthy thing happening to the body or mind. Diseases can cause pain, feeling bad, or death to the person who has the disease. The word *disease* is sometimes used to include:

- parts of the body being hurt,
- not having the usual abilities,
- medical problems or syndromes,
- infections by microorganisms,
- feeling unhealthy, such as having pain or feeling hot (called 'symptoms'),
- unusual shapes of body parts

1.1 Significance Of The Problem

This research work can provide the solution of following questions:

1. How many patients admit in hospital in particular month segment and in which month segment maximum patients are admitted?
2. What are different type of diseases thataffect the people most?
3. When maximum number of people becomes ill and admits in hospital. They suffered from which dominant disease?

Today most of peoples suffered from various type of diseases. Sometimes especially the seasonal diseases like dengue, typhoid etc. become viral and every second person found ill. Many patients lost their life due to lack of treatment. The facilities of hospital collapsed very rapidly because no. of patients suddenly increases and hospital have arrangements to provide treatment to patients. This research mainly focuses on these types of problems faced by both patients and hospital administration. This work finds out the dominant disease i.e. by which disease max. number of patients suffered and in which month so that hospital management can increase or decrease its facilities like no. of beds, medicines, doctors, specialists etc. In time and all the patients can get the treatment easily and no one have to lost one's life.

II. RESEARCH BACKGROUND

In year 2010, Ingmar Schäfer, Eike-Christin von Leitneret al, a study was done to get information about interactions of chronic diseases that can help to facilitate diagnosis, amend prevention and enhance the patients' quality of life. The aim

of this study was to increase the knowledge of specific processes of multimorbidity in an unselected elderly population by identifying patterns of statistically significantly associated comorbidity. [6].
In year 2013,ShamsherBahadur Patelet al., provide a method to predict the diagnosis of heart disease with reduced number of attributes. Here fourteen attributes involved in predicting heart disease. But fourteen attributes are reduced to six attributes by using Genetic algorithm. Subsequently three classifiers like Naive Bayes, Classification by Clustering and Decision Tree are used to predict the diagnosis of heart disease after the reduction of number of attributes. [7].

III. RESEARCH METHODOLOGY

For solving the above two problems some research techniques and methodologies are used for obtaining the desired result. Some tools and algorithms are required for obtaining the result. Main steps under the research methodologies are:-
Review literature or research papers – first of all literatures and research papers were reviewed for getting more information about the problem and knowing which type of work was done by others on this topic and by which method.
Identify tools – then tools required for solving the problem were identified and the best tool – "WEKA" was selected from all.
Study database attributes and data structure – attributes and structure of the database was thoroughly studied for finding out useful attributes from the annual report of PGIMER For critical attributes used in the database discussion with some specialist is made.
Determine nature and definition of research problem and work flow of the problem for getting accurate and desired result.
Organize the database [8] with useful attributes and populate it then perform data analysis using suitable tool e.g., WEKA in order to generate the result.

IV. CONCEPTUAL FRAMEWORK

Clustering can be considered the most important unsupervised learning problem so, as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data.
A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". [9]. Data are grouped into clusters in such a way that data of the same group are similar and those in other groups are dissimilar [10]. The greater the similarity within a group, the greater will be the difference between groups and the better the clustering [11].Clustering has many applications, including part family formation for group technology, image segmentation, information retrieval, web pages grouping, market segmentation, and scientific and engineering analysis [12].
K-Means[13] is iterative expectation-maximization type approach, which attempts to address the following objective: given a set of points in a Euclidean space and a positive integer k (the number of clusters), split the points into k clusters so that the total sum of the (squared Euclidean) distances of each point to its nearest cluster center is minimized
This paper presents solution to two main problems related to discovering disease patterns in hospital data analysis. These problems are:

4.1 Profiling Of Patients

In this part of problem, the patients admitted in hospital are segregated into different groups to get information about the no. of patients admitted in hospital. The annual report of PGIMER is refined and create a new database after extracting desired attributes and then feed to Weka tool for clustering. Then patients are segregated into different groups according to their month segment of admission in hospital.
In fig. 1M1, M2, M(n) are various clusters refers different segments of month. Here month segment means number of patients admitted in every two months.

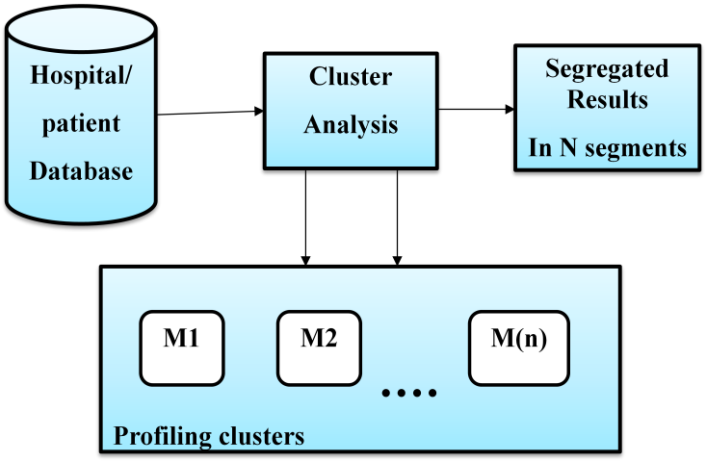

Fig. 1 Profiling of Patients

From its result, the dominant month and the dominant month segment are known to us i.e. in which month and in which month segment maximum no. of patients were admitted.

4.2 Discovering Travel Pattern Via Clustering

Segment Wise Clustering: Dataset of different disease segments (which are found by profiling), will collect and feed to Weka to find out the dominant month in which number of patients of particular disease is maximum. From here a link between month and disease segments will found. In other words, in a month there are different type of patients are admitted but we have to find out the dominant disease i.e. the name of disease having maximum patients in that particular month.

The same database is feed to Weka and form six clusters. These cluster contain information about month segments as well as the dominant disease in each segments. Dominant can be defined as maximum number of patients suffered by one disease.

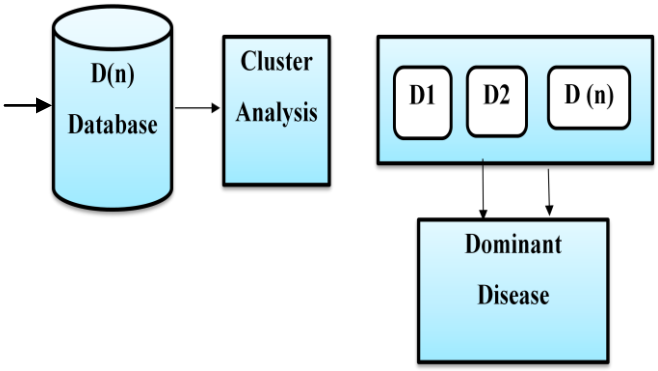

Fig. 2 Disease Pattern by Clustering

In fig 2, the disease having maximum entries of same month segment is dominant disease for that segment.

The output of this part of the problem contains information about the relation between disease and month segment as shown below:

Month segment (m) =>disease (d), symptoms(s)

V. RESULTS AND DISCUSSION

The data analysis is processed using WEKA data mining tool for exploratory data analysis. A database of 550 records/entries is collected from annual report of PGIMER available on internet. Primary data collected by random sampling, is used to solve the problems. After loaded the database in Weka all other attributes are removed and keep only three attributes named **Month segment, Monthof admit** &**Disease Name**to get the more accurate result.

5.1 Profiling of Travellers

To find out how many patient were admitted in every month segment (two months in one month segment)and from these two months in which month no. of admitted patients are maximum.

To get the result, input the prepared database to Weka tool, then select Simple K-mean algorithm and perform clustering to form 6 clusters of month segments.

For profiling of patients, only **Month segment** and **Month of Admit** attributes are selected and all other are removed, to get the accurate and reliable result.

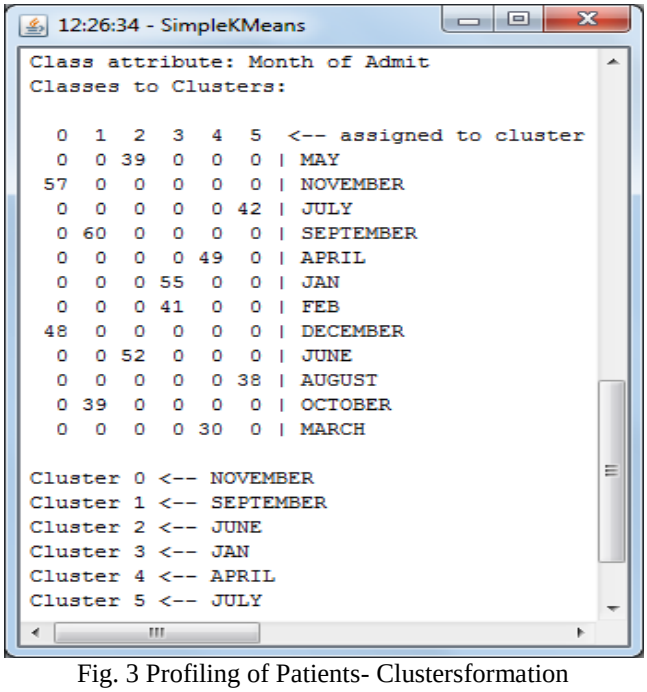

Fig. 3 Profiling of Patients- Clustersformation

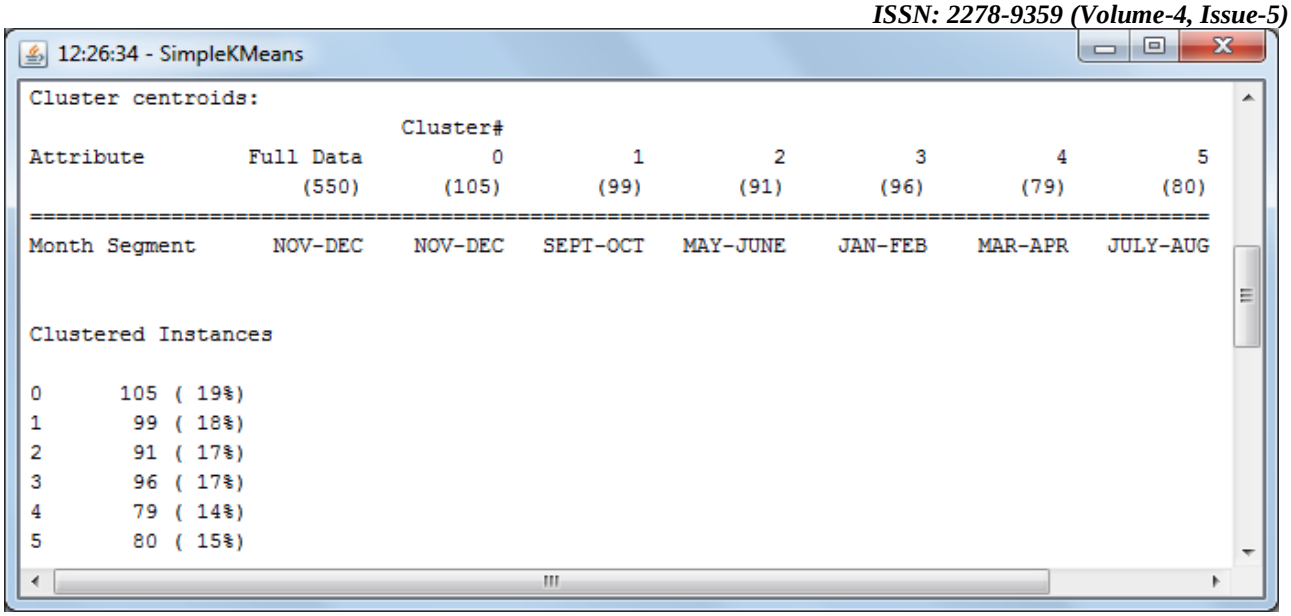

Fig. 4Clusters view

The information retrieved from Fig. 4, is as follows:
Cluster 0 contain 105patients, they all are admitted in NOV- DEC month segment. Cluster 1 contain 99patients, they all are admitted in SEP- OCT month segment. Cluster 2 contain 91patients, they all are admitted in MAY- JUNE month segment. Cluster 3 contain 96patients, they all are admitted in JAN- FEB month segment. Cluster 4 contain 79 patients, they all are admitted in MAR- APR month segment. Cluster 5 contain 80 patients, they all are admitted in JULY- AUG month segment.
Here the above said results represent in tabular form in Table 1.

Table 1 Profiling of Patients - Clusters formation

CLUSTER NO.	**MONTH SEGMENT**	**NO. OF PATIENTS**	**NO. OF PATIENT IN EVERY MONTH**	**DOMINANT MONTH**
CLUSTER 0	NOV-DEC	105	**NOV= 57** DEC= 48	**NOV**
CLUSTER 1	SEP-OCT	99	**SEPT= 60** OCT = 39	**SEPT**
CLUSTER 2	MAY-JUNE	91	MAY = 39 **JUNE= 52**	**JUNE**
CLUSTER 3	JAN-FEB	96	**JAN = 55** FEB = 41	**JAN**
CLUSTER 4	MARCH-APRIL	79	MARCH = 30 **APRIL = 49**	**APRIL**
CLUSTER 5	JULY-AUG	80	**JULY = 42** AUG= 38	**JULY**
DOMINANT	**NOV-DEC**	**105**	**60**	**SEPT**

5.2 Discovering Disease Pattern Via Clustering
To find out dominant disease (max. no. of patients were admitted of a particular disease) among all type of disease in particular month segment. It shows the relation between month segment and disease name i.e. within two months maximum number of patients admitted were suffered from which disease.
To get the result, input the prepared database to Weka tool, then select Simple K-mean algorithm and perform clustering to form 6 clusters of month segments and classes to cluster evaluation with disease name.

For Discovering Disease patterns, only **Month segment** and **Disease Name**attributes are selected and all other are removed, to get the accurate and reliable result.

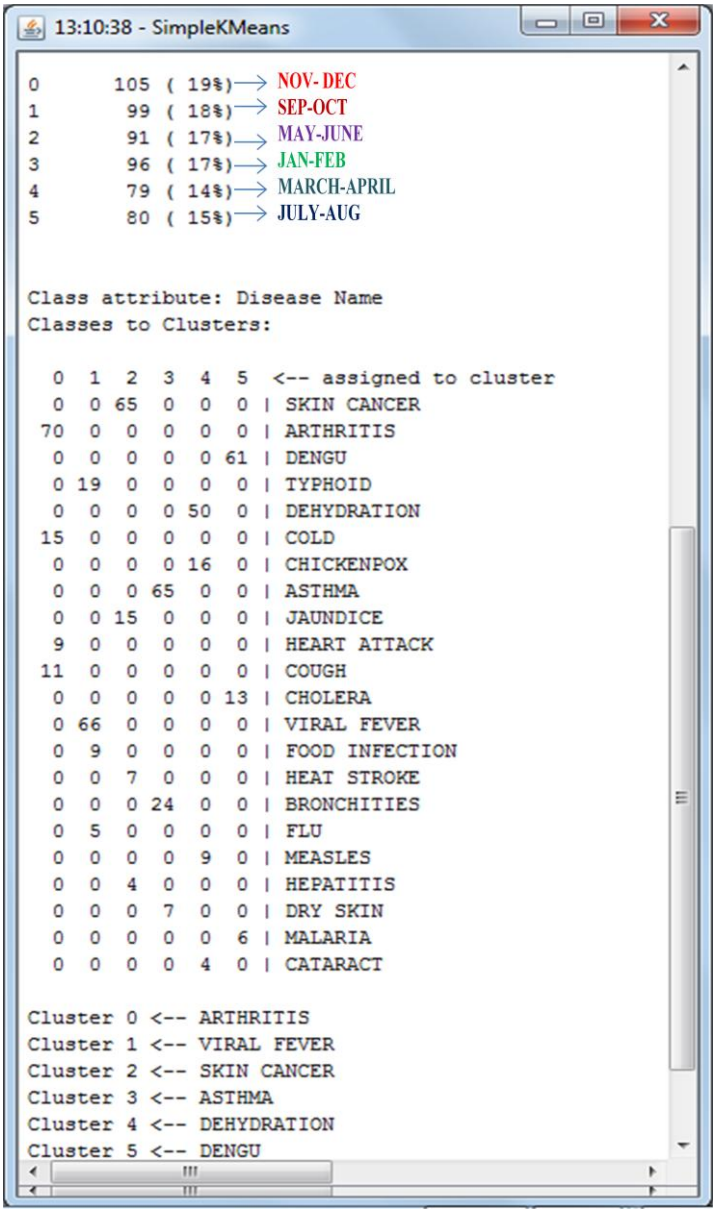

Fig.5Disease wise distribution of Month Segments

Fig. 5 show the relation between Month Segments and Disease name. Clusters are formed on basis of 6 different Month segments and the dominant Disease in all clusters or we can say segments.
In this fig. all the remaining destination are also shown. Except from the dominant disease in each segment, all other disease are also determined.
The information retrieved from here is as follows:

- **Cluster 0** represent patients admitted in Nov- Dec month segment and most of the patients were suffered from **Arthritis.**
- **Cluster 1**representpatients admitted in Sep-Oct month segment and most of the patients were suffered from **Viral Fever**.
- **Cluster 2**represent patients admitted in May-June month segment and most of the patients were suffered from **Skin Cancer**.
- **Cluster 3**representpatients admitted in Jan-Feb month segment and most of the patients were suffered from **Asthma.**
- **Cluster 4**represent patients admitted in March-April month segment and most of the patients were suffered from **Dehydration.**
- **Cluster 5**representpatients admitted in July-Aug month segment and most of the patients were suffered from **Dengue.**

The above description is shown in Table 2, in a nutshell.

Table 2 Disease Pattern via Clustering

CLUSTER NO.	MONTH SEGMENT	DOMINANT DISEASE	NO. OF PATIENTS OF DOMINANT DISEASE
CLUSTER 0	NOV-DEC	ARTHRITIS	70
CLUSTER 1	SEP-OCT	VIRAL FEVER	66
CLUSTER 2	MAY-JUNE	SKIN CANCER	65
CLUSTER 3	JAN-FEB	ASTHMA	65
CLUSTER 4	MARCH-APRIL	DEHYDRATION	50
CLUSTER 5	JULY-AUG	DENGU	61
DOMINANT	**NOV- DEC**	**ARTHRITIS**	**70**

VI. CONCLUSION

The purpose of this thesis is to discover disease patterns and then validate these patterns by matching the results. In our country there are so many diseases which are impossible to count on fingers even, and we have population in millions. So it is really a hectic or a major problem for our government and private hospitals to provide the treatment to all in such a huge amount. Usually it is seen that there is always a shortage of medicine and other resources in hospitals for provide treatment to all type of patients. But it will be quite easier to provide treatment to all, if the hospital will already aware about most spreading disease in a particular season or months. Then the hospital can already arrange all resources in advance and then no patient suffers due to the lack of resources.

This work provides information about the patient that how many patients are admit in hospital, their type of disease and in which month(s) that particular diseased patient admitted the most. Such that the hospitals come to know the requirements of number of resources for patients of particular disease in that particular month and will able to fulfil all requirement in advance.

VII. FUTURE WORK

This research can discover the disease patterns by using clustering. These disease patterns can also be find out by applying a different approach. Furthermore their results can also be compared to validate disease patterns discovered by two or more different approaches, concepts, algorithms etc.

ACKNOWLEDGEMENT

Author would like to thanks to their head Dr. Dinesh Garg, Assistant Professor in CSE & I.T department, DIET, Karnal, for his valuable support and help.

REFERENCES

[1] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
[2] PankajSaxena,Vineeta Singh and SushmaLehri," Evolving Efficient Clustering Patterns in Liver Patient Data through Data Mining Techniques", International Journal of Computer Applications (0975 – 8887), Volume 66– No.16, March 2013.
[3] Usama M. Fayyad. "Data mining and knowledge discovery: Making sense out of data", IEEE Expert: Intelligent Systems and Their Applications, 11(5):20–25, 1996.
[4] http://www.statsoft.com/textbook/data-mining-techniques.
[5] http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/
[6] Ingmar Schäfer,Eike-Christin von Leitneret al,Multimorbidity Patterns in the Elderly: A New Approach of Disease Clustering Identifies Complex Interrelations between Chronic Conditions,DOI: 10.1371/journal.pone.0015941,Published: December 29, 2010.
[7] ShamsherBahadur Patelet al, "Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques", IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), Volume 4, Issue 2 (Jul. - Aug. 2013), PP 61-64.
[8] Database - "Disease Pattern".
[9] Journal Of The Eastern Asia Society For Transportation Studies, Vol. 6, Pp. 4333 - 4348, 2005.

[10] Dr. SankarRajagopal, "Customer Data Clustering Using Data Mining Technique" International Journal of Database Management Systems (IJDMS) Vol.3, No.4, November 2011.
[11] Er. Arpit Gupta, Er.AnkitGupta,Er. Amit Mishra," Research Paper on Cluster Techniques of Data Variations", International Journal of Advance Technology & Engineering Research (IJATER).
[12] Pham, D.T. and Afify, A.A. (2006) "Clustering techniques and their applications in engineering". Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science.
[13] S.P. Lloyd. Least squares quantization in PCM. Unpublished Bell Lab. Tech. Note, portions presented at the Institute of Mathematical Statistics Meeting Atlantic City, NJ, September 1957. Also, IEEE Trans Inform Theory (Special Issue on Quantization), vol IT-28, pages 129-137, March 1982.
[14] Hand, D., Mannila, H., Smyth, P., Principles of Data Mining, Prentice Hall of India 2001.
[15] Praveen Rani et al, "Discovering travel pattern in passport data analysis", International Journal of Computer Science & Management Studies) Vol. 14, Issue 07, month-July 2014.
[16] Lefait, G. and Kechadi, T, (2010) "Customer Segmentation Architecture Based on Clustering Techniques" Digital Society, ICDS'10, Fourth International Conference, 10-02-2010.
[17] Fraley, Andrew, and Thearting, Kurt (1999). Increasing customer value by integrating data mining and campaign management software. Data Management, 49–53.
[18] I.Krishna Murthy, "Data Mining- Statistics Applications: A Key to Managerial Decision Making", Article/Report indiastat.com, April-May 2010.
[19] Er. Arpit Gupta, Er.AnkitGupta,Er. Amit Mishra," Research Paper on Cluster Techniques of Data Variations", International Journal of Advance Technology &Engineering Research (IJATER).
[20] Weka online documentation: *http://www.cs.waikato.ac.nz/ml/weka/index_documentation.html*
[21] Han, J. and M. Kamber, 2000: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.
[22] YAN-TAO ZHENG, Mining Travel Patterns from Geotagged Photos, ACM Transactions on Intelligent Systems andTechnology, Vol. V, No. N.
[23] Pham, D.T. and Afify, A.A. (2006) "Clustering techniques and their applications in engineering". Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science.
[24] Er. Arpit Gupta, Er.AnkitGupta,Er. Amit Mishra," Research Paper on Cluster Techniques of Data Variations", International Journal of Advance Technology & Engineering Research (IJATER).
[25] Mitchell TM. Machine learning. Boston, MA: McGraw-Hill, 1997.
[26] Hartigan, J., A. and Wong, M., A. 1979, "A K-Means Clustering Algorithm", Applied Statistics, Vol. 28, No. 1, pp. 100-108.
[27] Selim, S., Z. and Ismail, M., A. 1984, "KMeans Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality", IEEE Trans. PatternAnal. Mach. Intel., Vol. 6, No. 1, pp. 81–87.
[28] Krishna Murthy, "Data Mining- StatisticsApplications: A Key to Managerial Decision Making", Article/Report indiastat.com,April-May 2010.
[29] Nils J. Nilsson (1999) Introduction to machine Learning, California, United Stated of America.

