Data Clustering Method for Discovering Clusters in Spatial Cancer Databases

Ritu Chauhan
Jamia Hamdard
Hamdard University
New Delhi

Harleen Kaur
Jamia Hamdard
Hamdard University
New Delhi

M.Afshar Alam
Jamia Hamdard
Hamdard University
New Delhi

ABSTRACT
The vast amount of hidden data in huge databases has created tremendous interests in the field of data mining. This paper discusses the data analytical tools and data mining techniques to analyze the medical data as well as spatial data. Spatial data mining includes discovery of interesting and useful patterns from spatial databases by grouping the objects into clusters. This study focuses on discrete and continuous spatial medical databases on which clustering techniques are applied and the efficient clusters were formed. The clusters of arbitrary shapes are formed if the data is continuous in nature. Furthermore, this application investigated data mining techniques such as classical clustering and hierarchical clustering on the spatial data set to generate the efficient clusters. The experimental results showed that there are certain facts that are evolved and can not be superficially retrieved from raw data.

General Terms
Data mining, k-means, Clustering Algorithms

Keywords
Data Mining, Clustering, K-means, Hierarchical agglomerative clustering (HAC), SEER.

1. INTRODUCTION
Recently many commercial data mining clustering techniques have been developed and their usage is increasing tremendously to achieve desired goal. Researchers are putting their best efforts to achieve the fast and efficient algorithm for the abstraction of spatial medical data sets.

Cancer has become one of the leading causes of deaths in India. An analysis of most recent data has shown that over 7 lakh new cases of cancer and 3 lakh deaths occur annually due to cancer in India [1]. Furthermore, cancer is a preventable disease if it is analyzed at an early stage. There are various sites of cancer such as oral, stomach, liver, lungs, kidney, cervix, prostate testis, bladder and many others. There has been enormous growth in the clinical data from past decades, so we require proper data analysis techniques for more sophisticated methods of data exploration. In this study, we are using different data mining technique for effective implementation of clinical data.

The objective of this paper is to explore several data mining techniques on clinical and spatial data sets. Data mining is also known as knowledge discovery from large data base; it is the process to extract hidden relevant patterns, information and regularities from large databases. It is an emerging field which is currently used in marketing, Surveillance fraud detection, human factor related issue, medical pattern detection and scientific discovery. Several data mining techniques are pattern recognition, clustering, association, classification and clustering. The proposed work will focus on challenges related to clustering on medical spatial datasets. Clustering is the unsupervised classification of patterns into clusters [2]. There are recently developed fast algorithms for clustering large data sets such as DBSCAN, CLARANS, BIRCH, STING[3], [4], [5] [6] . They are several series of facts have been gathered during the series of experiments.

This chapter is organized as follows: Section 2, we discuss the related works of clustering algorithms. Section 3 the Experimental analysis on spatial medical datasets has been discussed. Conclusions are presented in the last section.

2. Clustering Algorithms

The community of users has played lot emphasis on developing fast algorithms for clustering large data sets [13]. Clustering is a technique by which similar objects are grouped together. Clustering algorithms can be classified into several categories such partitioning-based clustering, hierarchical algorithms, density based clustering and grid based clustering.

Now a day's huge amount of data is gathered from remote sensing, medical data, geographic information system, environment etc. So everyday we are left with enormous amount of data that requires proper analysis. Data mining is one of the emerging fields that are used for proper decision-making and utilizing these resources for analysis of data. They are several researches focused on medical decision making [14] [15]. Data clustering techniques have been extensively used are:

2.1 Partitioning Based Clustering
The K-means algorithm is a classical clustering method which is used to group large datasets into clusters [8][16]. It is the unsupervised classification to find optimal clusters. The algorithm is often considered to be a partitioning clustering method, and it works as follows. It arbitrarily chooses the cluster center then the objects are assigned to the similar cluster, which are more similar. The cluster means are updated for each cluster until there is no change. The disadvantage of using K-means method is the number of cluster should be specified in the beginning and it is not able to generate the cluster with different shapes. Given the above disadvantages, there is the silhouette value also known as silhouette width, gives a sort of compactness

of a cluster with respect to the other clusters, See Ref.[8] for more detailed discussions and analyses of these issues.

2.2 Hierarchical Algorithms

It is the clustering method by which the data are grouped together in form of trees. The hierarchical clustering is generally classified into two types of approach such as agglomerative approach and divisive approach [9].

- Agglomerative approach is the clustering technique in which bottom up strategy is used to cluster the objects. It merges the atomic clusters into larger and larger until all the objects are merged into single cluster.
- Divisive approach is the clustering technique in which top down strategy is used to cluster the objects. In this method the larger clusters are divided into smaller clusters until each object forms cluster of its own. Figure 1 shows simple example of hierarchical clustering

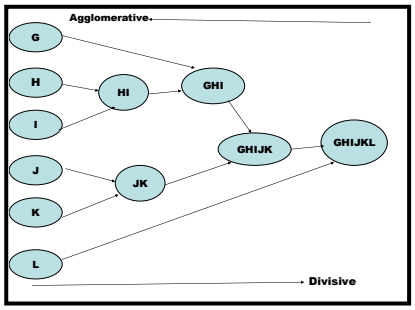


Figure 1 Hierarchical Clustering

2.3 Density Based Clustering

It is a clustering technique to develop clusters of arbitrary shapes. They are different types of density based clustering techniques such as DBSCAN, OPTICS and DENCLUE.

2.3.1 DBSCAN

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a clustering technique for density connected points DBSCAN starts with an arbitrary point *p* and retrieves all points density-reachable from *p* wrt. *Eps* and *MinPts*. If *p* is a core point, this procedure yields a cluster wrt. *Eps* and *MinPts* . If *p* is a border point, no points are density-reachable from *p* and DBSCAN visits the next point of the database. In density-based clustering each object of a cluster in the neighborhood of a given radius should have at least a minimum number *MinPts* of objects, i.e. the cardinality of the neighborhood has to exceed a given threshold. With so many advantages DBSCAN still suffers from a few drawbacks. Namely they are sensitivity to input parameters (the clustering result very much depends on the 'epsilon' parameter of the algorithm) and in some cases algorithm is not able to correctly identify clusters that are close to each other.

This algorithm requires minimal knowledge of domain to determine the Input parameters, because appropriate values are often not known in advance when dealing with large databases. They are able to find the clusters of arbitrary shapes.

2.3.2 OPTICS

OPTICS (Ordering Points to Identify the Clustering Structure) is the clustering technique in which the augmented order of the datasets for cluster analysis. Optics built dataset-using density based clustering structure.

The advantage of using optics is it in not sensitive to parameters input values through the user it automatically generates the number of clusters [12].

2.3.3 DENCLUE

DENCLUE (Clustering Based on Density Distribution Function) is the clustering technique in which the clustering method is dependent on density distribution function. The clustering technique is basically based on influence function (data point impact on its neighborhood), the overall density of data space can be calculated as the sum of influence functions applied to data points) and cluster can be calculated using density attractors (local maxima of the overall density function).

2.4 Grid Based Clustering

Among the existing clustering algorithms, grid-based algorithms generally have a fast processing time, which first employ a uniform grid to collect the regional statistic data and then, perform the clustering on the grid, instead of the database directly. The performance of grid-based approach normally depends on the size of the grid which is usually much less than the database. However, for highly irregular data distributions, using a single uniform grid may not be sufficient to obtain a required clustering quality or fulfill the time requirement. There are different types of grid based clustering technique such as Sting, Wave Cluster and Clique [7].

3. EXPERIMENTS

They are several series of experiments performed in this section to determine relevant pattern detection for medical diagnosis.

3.1 Clinical Database I

The dataset consists of number of cancer patients those who registered themselves to the www-dep.iarc.fr/globocan/database.htm. The dataset consists of basic attributes such as sex, age, marital status, height and weight. The data of age group was taken from (15 - 65+) years in this group major cancers were examined. A total of male and female cases were examined for the various cancers. The data were collected and substantial distribution was found for Incidence and Mortality by Sex and Cancer site. Perhaps analysis suggests that they were more male cases those who were suffering from cancer as per opposite sex. The patients from other sex were too small to be considered. The database analysis was done using TANAGRA tool kit. Figure 2 represents the statistical diagram for representation between number of male and female cases for cancer. The TANAGRA is software that has several data mining software for data analysis, statistical tools in data base [10]. The

data consists of discrete data sets with following attribute value types of cancer, male cases, female cases, cases of death pertaining to specific cancer. They were around 21 cancers that have been used as the part of analysis. The TANAGRA tool doesn't take the discrete value so it has to be transformed into continuous attribute value. The data was subdivided into X, Y values and the result was formed using K-means and HAC clustering algorithm. In TANAGRA, the low level clusters are formed using K-MEANS and SOM then HAC clustering builds the Dendogram using the low level clusters. Figure 3 and 4 specifies the number of cluster for male and female suffering from different cancers using TANAGRA.

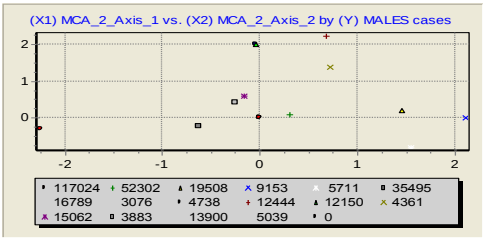


Figure. 3. Male cases of Cancer

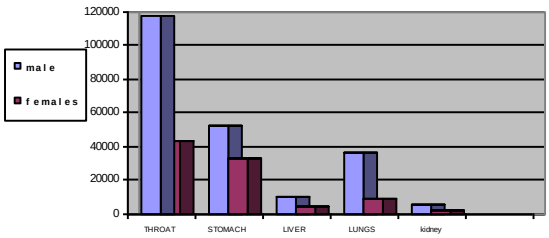


Figure 2 cases of male and female

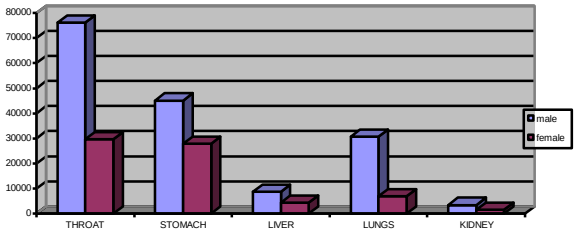
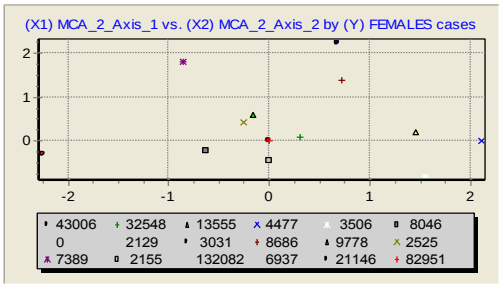


Fig 5 cases of death male and females



. **Figure. 4. Female cases of Cancer**

Note: that the other sample was collected for patients those who couldn't survive for the disease. The analysis suggests the male ratio was large in percentage as compared to the female. Perhaps by analyzing the data we can develop certain measures for the better procurement of disease.

Figure 6 and 7 specifies the number of death in males and female cases of death due to cancer using TANAGRA. The clusters have the number of higher cases of male as per female m

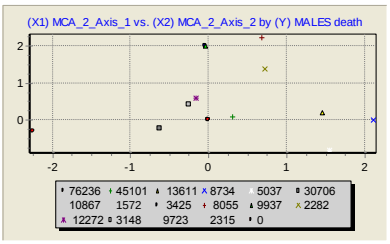


Figure 6. Male cases death

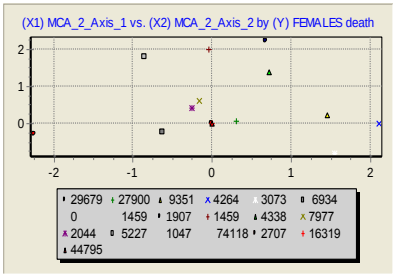


Figure. 7. Female cases death

3.2 Clinical Database II

In this study, the data was taken from SEER datasets which has record of cancer patients from the year 1975 – 2001 from Ref. [11]. The data was again classified into two groups that are spatial and non spatial dataset. Spatial dataset consists of location collected include remotely sensed images, geographical information with spatial attributes such as *location*, digital sky survey data, mobile phone usage data, and medical data. The five major cancer areas such as lung, kidney, throat, stomach and liver were experimented. After this data mining algorithms were applied on the data sets such as K-means, SOM and Hierarchical clustering technique

The K-means method is an efficient technique for clustering large data sets and is used to determine the size of each cluster. After this the HAC (hierarchical agglomerative clustering), is used on our datasets in which we have used tree based partition method in which the results has shown a tree structure and the gap between the nodes has been highlighted in the Table 3. The HAC has proved to have for better results than other clustering methods. The principal component analysis technique has been used to visualize the data. The X, Y coordinates identify the point location of objects. The coordinates were used and the clusters were determined by appropriate attribute value. The mean and standard deviation of each cluster was determined.

They were interesting facts that suggests that number of cancer cases that occur during the time interval.

Table 1 represents the size of each cluster determined by K-means clustering technique for dataset 2 .In Table 2 shows the number trials generated for the cluster determination.

Table 1. Result of K-means

Clusters	3	
Cluster	**Description**	**Size**
cluster n°1	c_kmeans_1	8
cluster n°2	c_kmeans_2	17
cluster n°3	c_kmeans_3	2

Table 2 Number of trials

Number of trials	5
Trial	**Ratio explained**
1	0.457312
2	0.480544
3	0.456888
4	0.520696
5	0.205954

Table 3 presents HAC (hierarchical agglomerative clustering) in which the cluster were determined with appropriate size. Table 4 represents the best cluster selection in which the gap is defined as the space in between the clusters.

Table 3. HAC Cluster size

Clusters	3	
Cluster	**Description**	**Size**
cluster n°1	c_hac_1	16
cluster n°2	c_hac_2	2
cluster n°3	c_hac_3	9

Table 4 Cluster Selection

Clusters	BSS ratio	Gap
1	0	0
2	0.4365	3.205
3	0.5817	0.5671
4	0.6753	0.5323
5	0.7205	0.1321
6	0.7536	0.0937
7	0.7783	0.0247
8	0.8007	0.0348
9	0.82	0.028
10	0.8367	0.0176
11	0.8518	0.016
12	0.8655	0.0185
13	0.8775	0.0083
14	0.8887	0.0073
15	0.8993	0.0147
16	0.9085	0.011
17	0.9167	0.0065
18	0.9244	0.0114
19	0.931	0.0107
20	0.9366	0.005

Figure 8 represents the dendogram in which the dataset has been partitioned into three clusters with the K-means. The HAC clustering algorithm is applied on K-means to generate the dendogram. In a dendogram, the elements are grouped together in one cluster when they have the closest values of all elements available. In the diagram the cluster2 and cluster 3 are combined. The subdivisions of clusters are then analyzed.

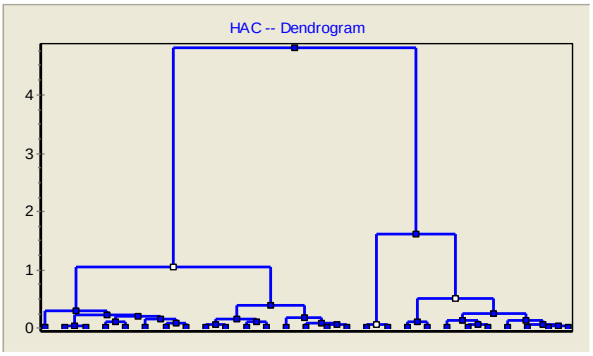


Figure. 8. Dendrogram

Table 5 characterizes the cluster according to the mean and standard deviation of each object and cluster were determined. The first comparison in between cluster 1 objects. The second comparison was between the objects of cluster 2 and cluster 1.

The third comparison was determined in between cluster 3 and cluster 1. The results show the mean and standard deviation of each cluster and also among the objects in each cluster. The cluster 1 has the lowest number of cancer cases the cluster 2 has largest number of cancer cases where as the cluster 3 has average number of cancer cases.

Table 5. Representation of Cluster Mean and Standard Deviation

Cluster_HAC_1=c_hac_1				Cluster_HAC_1=c_hac_2			
Examples		[59.3 %] 16		Examples		[7.4 %] 2	
	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
Year ofDiagnosis	0.5	82.50 (4.76)	80.59 (24.14)	moral	2.9	64.90 (0.99)	55.78 (4.52)
Att - Desc	-1.7	11.79 (4.34)	13.03 (4.47)	mliver	2.7	13.60 (0.00)	6.53 (3.72)
fstomach	-2.1	24.98 (5.31)	26.61 (4.84)	mstomach	2.7	27.60 (2.97)	19.57 (4.22)
mstomach	-3.5	17.15 (1.90)	19.57 (4.22)	fliver	2.6	21.05 (1.77)	13.03 (4.47)
flungs	-3.6	18.88 (1.08)	19.83 (1.64)	flungs	2.4	22.55 (0.78)	19.83 (1.64)
foral	-3.7	62.41 (4.19)	65.60 (5.32)	mlungs	2.3	16.25 (0.64)	14.56 (1.05)
mliver	-3.9	4.20 (1.92)	6.53 (3.72)	mkidney	2.1	70.00 (0.99)	62.13 (5.30)
mlungs	-3.9	13.90 (0.68)	14.56 (1.05)	fkidney	2.1	77.75 (2.05)	66.97 (7.28)
mkidney	-4	58.71 (3.83)	62.13 (5.30)	fstomach	1.7	32.20 (1.56)	26.61 (4.84)
moral	-4.1	52.80 (1.63)	55.78 (4.52)	foral	1.6	71.45 (0.92)	65.60 (5.32)
fkidney	-4.1	62.14 (5.17)	66.97 (7.28)	Year ofDiagnosis	-4.8	0.50 (0.71)	80.59 (24.14)

Figure 9 shows the clusters were formed using K-Means and HAC. The comparison was made in the attribute value and the scatter plot was formed. To find the most appropriate cluster we use HAC clustering technique.

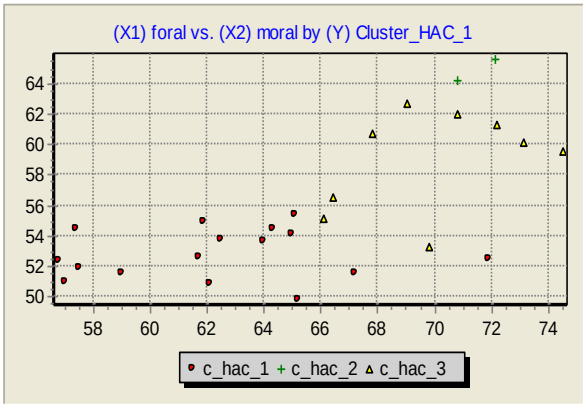


Figure. 9. Scatter plot of male and female Oral Cancer

The cluster compactness has been determined by standard deviation where the cluster becomes compact when standard deviation value decreases and if the value of standard deviation increases the cluster becomes dispersed.

4. CONCLUSIONS

This paper focuses on clustering algorithms such as HAC and K-Means in which, HAC is applied on K-means to determine the number of clusters. The quality of cluster is improved, if HAC is applied on K-means. The paper has referenced and discussed the issues on the specified algorithms for the data analysis. The analysis does not include missing records. The application can be used to demonstrate how data mining technique can be combined with medical data sets and can be effectively demonstrated in modifying the clinical research.

This study clearly shows that data mining techniques are promising for clinical datasets. Our future work will be related to missing values and applying various algorithms for the fast implementation of records. In addition, the research would be focusing on spatial data clustering to develop a new spatial data mining algorithm.

5. REFERENCES

[1] Rao, Y.N, Sudir Gupta and S.P. Agarwal 2003. National Cancer Control Programme:Current status and strategies, 50 years of cancer control in India,NCD Section, Director General of Health.

[2] Jain, A.K., Murty M.N., and Flynn P.J. (1999): *Data Clustering: A Review.*

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu.1996. A density-based algorithm for discovering clusters in large spatial databases. *KDD'96*.

[4] Ng R.T., and Han J. 1994. Efficient and Effective Clustering Methods for Spatial Data Mining, *Proc. 20th Int. Conf. on Very Large Data Bases*, Chile.

[5] W. Wang, J. Yang, and R. Muntz, STING: A Statistical Information grid approach to spatial data mining, *Proc. 23rd1nt. Conf. on Very Large Databases*, Morgan Kaufmann, pp. 186-195 (1997).

[6] T. Zhang, R. Ramakrishnan, and M. L1nvy, B1RCH: An Efficient Data C1ustering Method for Very Large Databases, Proc. ACM SIGMOD Int'L Conf. On Management of Data, ACM Press, pp. 103-114 (1996).

[7] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

[8] L. Kaufinan, and P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons1990.

[9] Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In CIKM, 2002.

[10] http://eric.univlyon2.fr/~ricco/tanagra/en/tanagra.html.

[11] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov)Public-Use Data (1973-2002), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2005.

[12] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander (1999). "OPTICS: Ordering Points to Identify the Clustering Structure". *ACM SIGMOD* international conference on Management of data.

[13] U.M. Fayyad and P. Smyth. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Menlo Park, CA, 1996.

[14] Kaur H, Wasan S K, Al-Hegami A S and Bhatnagar V, A Unified Approach for Discovery of Interesting Association Rules in Medical Databases, *Advances in Data Mining*, Lecture Notes in Artificial Intelligence, Vol. 4065, Springer-Verlag, Berlin, Heidelberg (2006).

[15] Kaur H and Wasan S K, An Integrated Approach in Medical Decision Making for Eliciting Knowledge, Web-based Applications in Health Care & Biomedicine, *Annals of Information Systems (AoIS), ed. A. Lazakidou*, Springer 2009.

[16] M. S. Chen, J. Han, and P. S. Yu. Data mining: an overview from database perspective. IEEE Trans. On Knowledge and Data Engineering, 5(1):866—883, Dec.1996