Formation of Clusters with K – Means Algorithm Using Tumor Classification

[1]BhagyashreeRajgire,[2]Prerana P. Khobragade, [3]Bharti K.Haygune, [4]Gauri A. Dhoke

[1] Assistant professor, CSE Dept, SSCET, Bhadrawati. *Email: bhagya.rajgire29@gmail.com*

[2] CSE Dept, SSCET, Bhadrawati. *Email: prerana_khobragade@rediffmail.com*

[3]CSE Dept, SSCET, Bhadrawati. *Email: haygune34@gmail.com*

[4] CSE Dept, SSCET, Bhadrawati. *Email: gauridhoke27@gmail.com*

Abstract: Clustering is important data mining techniqueto extract useful information from various high dimensional datasets. A wide range of clustering algorithms is available in literature and still an open area for researcher. K-means algorithm is one of the basic and most simple partitioning clustering technique is given byMacQueen in 1967 and clustering algorithm aims todivide the dataset into disjoint clusters. After that many variations of k-means algorithm are given by different authors. Here in this paper we make analysis of k-mean based algorithms namely global k-means, efficient k-means.

*****

I. INTRODUCTION:

Handling datasets using computational methods has become an integral part of business, science and engineering practice. But the wide spread of computerization, on the other hand, increases the amount of data stored in terms of number of dimension, number of instances and data types that becomes problematic when one has to deal with a dataset with huge dimensions and or/huge instances. In the early 1990's the establishment of the internet made large quantities of data to be stored electronically, which was a great innovation for information technology.However, the question is what to do with all this data. Data mining is the process of discovering useful information (i.e. patterns) underlying the data. Powerful techniques are needed to extract patterns from large data because traditional statistical tools are not efficient enough anymore. Clustering is an important data mining technique that puts together similar objects into a collection in which the objects exhibit certain degree of similarities. Clustering also separates dissimilar objects into different groups. Clustering describes the underlying structure of the data by its unsupervised learning ability. Due to its supervised learning ability, it is able to discover hidden patterns of datasets. This is made clustering an important research topic of diverse fields such as pattern recognition, bioinformatics and data mining. It has been applied in many fields of study, from ancient Greek astronomy to present-day insurance industry and medical. The term classification and clustering is confusing, but they have the difference that in classification objects allocates in predefined classes while in the clustering classes is created. In database management, clustering is a process where, physically stored information is similar to logical information. To make efficient search and rescue in database, several disk admittance to be reduced. Objects having similar properties are grouped in the same class of objects. In the field of data mining, image processing, pattern recognition, machine learning, vector quantization, etc. Data clustering is used often, whose goal is to partition data into similar groups.

Clustering: Clustering is an important chore in data analysis and data mining applications. Data divides into similar object groups based on their features by clustering process. Each data group with similar objects are clusters. It means clusters are the ordered set of data which have the familiar characteristics. Clustering is a process of unsupervised learning. Highly superior clusters have high intra-class similarity and low inter-class similarity.

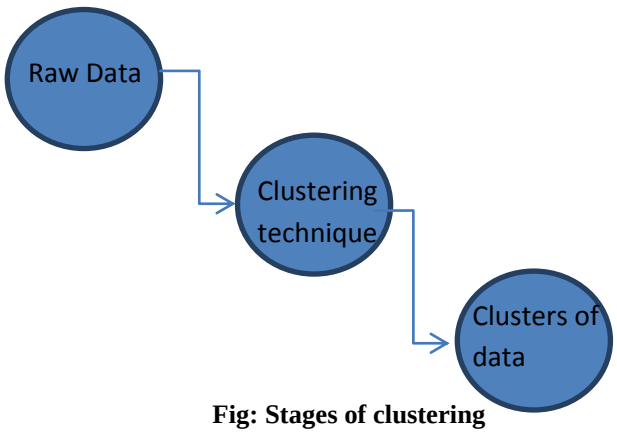

Fig: Stages of clustering

Clustering algorithms have many categories like hierarchicalbased algorithms, partition-based algorithms, density-based algorithms and grid based algorithms.

Partition-based clustering:

It is centroid based clustering in which data pointsspilits into k partition and each partition represents a cluster. Different methods of partitioning clustering are k-means, bisecting k-means method, Medoids method, Portioning AroundMedoids (PAM), CLARA (Clustering large Applications) and the Probabilistic centroid.

K-means Algorithm:

The k-means algorithm (MacQueen, 1967) is one of a group of algorithms called partitioning methods. The k - means algorithm is very simple and can be easily implemented in solving many practical problems. The k-means algorithm is the best-known squared error-based clustering algorithm.

K-means clustering technique is a technique of clustering which iswidely used. This algorithm is the most popular clustering tool that is used in scientific and industrial applications. It is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean . The basics algorithm is very simple :

1. Select K points as initial centroids.

2. Repeat

3. Form K cluster by assigning each point to its closest centroid.

4. Recomputed the centroid of each cluster until centroid does not change.

Properties of k-means algorithm:

1. Large data set are efficiently processed.

2. It often terminates at a local optimum.

3. It works only on numeric values.

4. The shape of clusters are convex.

II. LITERATURE REVIEW:

K-mean is the most popular partitioning method of clustering. Mac Queen in 1967, firstly proposed this technique, though the idea goes back to Hugo Steinhaus in 1957. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation. Sometimes it is referred as Lloyd-Forgybecause In 1965, E.W.Forgy published essentially the same method .

K. A. Abdul Nazeer et al.discuss in this paper about the one major drawback of the k-means algorithm. K-means algorithm, for different sets of values of initial centroids, produces different clusters. Final cluster quality in algorithm depends on the selection of initial centroids. Two phases includes in original k means algorithm: first for determining initial centroids and second for assigning data points to the nearest clusters and then recalculating the clustering mean. An enhanced clustering method is discussed in this paper,in which both the phases of the original k-means algorithm are modified to improve the accuracy and efficiency

Tumor:A tumor, also known as a neoplasm, is an abnormal mass of tissue which may be solid or fluid-filled. Atumor does not mean cancer – tumors can be benign (not cancerous), pre-malignant 9pre-cancerous), or malignant (cancerous). There are many types of tumors and a variety of names for them – their names usually reflect their shape and the kind of tissue they appear in. Put simply, a tumor is a kind of lump or swelling, it does not necessarily pose a health threat. The terms tumors and cancer are sometimes used interchangeably which can be misleading. A tumor is not necessarilya cancer. The word tumor simply refers to a mass.

A tumor is a commonly used, but non-specific, term for a neoplasm. The word tumor simply refers to a mass. This is general term that can refer to benign or malignant growths.

Benign tumors are non-malignant / non-cancerous tumor. A benign tumor is usually localized, and does not spread to other parts of the body. Most benign tumors respond well to treatment. However, if left untreated, some benign tumors can grow large and lead to serious disease because of their size. Benign tumors can also mimic malignant tumors, and so for this reason are sometimes treated.

Malignant tumors are cancerous growths. They are often resistant to treatment , may spread to other parts of the body and they sometimes recur after they were removed.

Tumor classification:A classification is an organization of everything in a domain by hierarchical groups, according to features generalizable to the members of the groups. Four terms with distinctly different meanings have been used interchangeably with “Classification, “leading to considerable confusion among pathologists and cancer researchers. These terms are: identification, discrimination, taxonomy, and ontology. Identification (also known as diagnosing or naming) is the act of placing something into its correct slot within an existing classification. Discrimination is finding features that separate members of a group according to expected variations in group behaviour. Example of discrimination are “grading and staging.” Grading and Staging involve reporting additional features morphology (grading) or clinical behaviour (staging) that help predict a particular tumor’s clinical course or response to therapy. A taxonomy is a complete listing of aa the members of a classification. In the case of neoplasia, a taxonomy would be the complete listing of all the different name tumors. An Ontology is a rule-based grouping of some portion of a taxonomy. Ontology is support queries and logical inferences pertaining to the [ontologic] group members.

Traditionally, tumor have been classified by their morphologic appearances. Unfortunately, tumors with similar histologic features often follow different cleanical courses or respond differently to chemotherapy. Limitations in the clinical utility of morphology-based tumor classifications have prompted a search for a new tumor classification based on molecular analysis. Gene expression

array data an proteomic data from tumor samples will provide complex data i.e unobtainable from morphologic examination alone. The growing question facing cancer researchers is, " How can we successfully integrate the molecular, morphologic and clinical characteristics of human cancer to produce a helpful tumor classification?".

A tumor classification must include every type of tumor and must provide a unique place for each tumor within the classification. Groups within a classification inherit the properties of theirancestorscan impart properties to their descendants. A classification was prepared grouping tumors according to theirhistogenetic development. The classification is simple (reducing the complexity of information received from the molecular analysis of tumors), comprehensive (providing a place for every tumor of man), an consistent with recent attempts to characterize tumors by cytogenetic and molecular features.

III. EXISTING SYSTEM:

The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity.

Advantages:

- Simplicity
- Speed

Disadvantages:

Does not yield the same result with each run, since the resulting clusters depend on the initial random assignments Requirement for the concept of a mean to be definable which is not always the case.

IV. PROPOSED SYSTEM:

The aim of the proposed algorithm is to improve the computational efficiency of the K Means algorithm. The algorithm involves initial centroid selection, which is done randomly in existing algorithm. Hence we propose an algorithm which selects initial centroids based on the distances calculated from the origin. Moreover, the k-means algorithm is computationally very expensive also. The proposed algorithm is found to be more accurate and efficient compared to the original k-means algorithm. This proposed method finding the better initial centroids and provides an efficient way of assigning the data points to the suitable clusters and classify the tumors making doctors an easier task. The proposed algorithm produces the more accurate unique clustering results. The value of k, desired number of clusters is still required to be given as an input to the proposed algorithm

V. CONCLUSION:

In this review work most widely used k-means clustering techniques of data mining is analysed. This work shows that there are several methods to improve the clustering with different approaches. Various clustering techniques are reviewed which improve the existing algorithm with different perspective. Some limitations of existing algorithm will be eliminated in future work. This technique will be useful in extraction of useful information using cluster from huge Database.

REFERENCES:

[1] Alexander Genkin, David D. Lewis, David Madigan (2004). Large-scale bayesian logistic regression for text categorization, 2004.

[2] Alizadeh A., Eisen M.B, Davis R.E, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 403(6769):503–511,2000.

[3] Bentley J. L. and Friedman J.H., Fast algorithms for constructing minimal spanning trees in coordinate spaces. IEEE Transactions on Computers, C- 27(2): 97 – 105, February 1978.

[4] Cheeseman P., Stutz J.:Bayesian Classification(AutoClass): theory and Results. Advances in Knowledge Discovery and Datamining, 153-180,1996.

[5] Genkin A., Lewis D. Large-Scale Bayesian Logistic Regression for Text Categorization. http://www.stat.rutgers.edu/˜madigan/BBR/, 2004.

[6] Duda, P. E. Hart and D. G. Stork, Pattern Classification, Wiley, New York, 2001.

[7] Fort G., Lambert S., "Classification using partial least squares with penalized logistic regression", England: Bioinformatics-Oxford, 2005.

[8] Freund, Y., Mason, L.: The alternating decision tree learning algorithm. In: Proceeding of the Sixteenth International Conference on Machine Learning, Bled, Slovenia, 124-133, 1999.

[9] Golub T.R, Slonim D.K, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science. 286(5439): 531–37, 1999.

[10] Quinlan J.R., C4.5: Programs for Machine Learning. Morgan Kaufman, 1993

[11] Roozgard A, Cheng S, Liu H. Malignant Nodule Detection on Lung CT Scan Images with Kernel RX Algorithm. Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics; 2012. p. 499–502.

[12] Mann AK, Kaur N. Survey Paper on Clustering Techniques. IJSETR. 2013 Apr; 2(4):803–6.

[13] Saini A, Kumar V. Detection system for lung cancer based on neural network: X-Ray validation performance. International Journal of Enhanced Research in Management & Computer Applications. 2013; 2(9):40–7.