IJCAT International Journal of Computing and Technology, Volume 1, Issue 4, May 2014
ISSN : 2348 - 6090
www.IJCAT.org

Human Genome Data Clustering Using K-Means Algorithm

[1] Amrita A. Kulkarni, [2] Prof. Deepak Kapgate

[1] Department of C.S.E., GHRAET, Nagpur University, Nagpur, Maharashtra, India

[2] Department of C.S.E., GHRAET, Nagpur University, Nagpur, Maharashtra, India

Abstract- In medical science field K-means algorithm can be applied to form clusters and data can be arranged in specific format. This paper explores the human genome data and indulgence of this data. Data from human genome project site is collected and model text is used. Data pre-processing steps in data mining are applied and inconsistencies in obtained data are removed. Using cluster analysis technique further this data is then grouped together using modified K-means algorithm. The formed clustered are arranged in group and availability of data is made easy in human genome information. Furthermore, outlier detection is possible and genetic disorders can be identified.

Keywords **- Human Genome, Clustering, K-means Algorithm, Pre-processing.**

1. Introduction

A large amount of computerized medical data is currently available in various forms such as text, images, numbers, video, audio reports. This data are used along with various analysis techniques to generate results that can be used by the health care professionals in efficient decision making.

The human genome is the complete set of genetic information for humans. This information is encoded as DNA sequences within the 23 chromosome pairs in cell nuclei and in a small DNA molecule found within individual mitochondria. Human genomes include both protein-coding DNA genes and non-coding DNA. . Coding DNA is defined as sequences that can be record into mRNA and translated into proteins during the human life cycle; these sequences occupy only a small fraction of the genome i.e. less than 2%. Non-coding DNA is made up of all of those sequences that are not used to encode proteins which are nearly 98%.

The data is acquired from human genome project site and steps in knowledge discovery process are applied. The concept of KDD states that knowledge is extracted from raw data given as input. However, the raw data is first transformed into a form that is appropriate for the Data Mining process using the KDD pre-processing steps. The data mining technique clustering is then applied on the preprocessed data to generate knowledge.

This paper discusses previous research on clustering in medical data (section 2). Later it describes data processing steps as collection, cleaning, clustering (section 3). Basic K-means algorithm is explained in section 4 and result of this paper is conferred at end.

2. Related Work

Kazuki Ichikawa et. Al. [1] analyze A simple but powerful heuristic method for accelerating k-means clustering of large-scale data in life science. Michael b. Eisen et. Al. [12] presents Cluster analysis and display of genome-wide expression patterns.

Rakesh Agrawal et. Al.[4] presents the ways of using sequences of clustering algorithms to mine temporal data. K.Y.Yeung et. Al. [6] analyses the performances of K-Means clustering algorithm with other clustering algorithms on the gene expression data.

Cios KJ et. Al. [7] discusses medical data mining with respect to the Heterogeneity of the data, ethical, legal and social issues, privacy and security issues related to handling medical Data.
Alp Aslandogan Y. And Gauri A. Mahajani (2004) discusses how various combinations of Data Mining Classification algorithms are used on medical data for efficient classification of the data.

Patricia Cerrito et. Al. [5] discusses on how the electronic medical records are created from Manual records and how data and text mining are used to improve the quality and reduce costs.

Anil K. Jain [13]presents a detailed study on well-known clustering methods, discuss the major challenges and key issues in designing clustering algorithms, and point out some of the emerging and useful research directions, including semi-supervised clustering, ensemble clustering and simultaneous feature selection during data Clustering, and large scale data clustering.

3. Data Pre-processing

A small portion from human genome database is collected and this sample is then used as an example in this paper. Using data pre-processing techniques noise, errors, inconsistencies in gained data can be identified and removed. The pre-processing includes Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Generation of Patterns and Knowledge Interpretation.

3.1 Data Cleaning

The process of detecting and correcting or removing inaccurate records from a record set, table, or database is Data Cleaning. The missing values in the human genome data cannot be replaced by any other value and hence those missing entries are replaced by the text "NA" referring "Not Applicable".

3.2 Data Selection

After cleaning step data is selected to form clusters in which attributes which contains only numbers are selected and "KOBIC", "SNV", "Variant_seq", "Reference_seq" and "Genotype" are removed.

3.3 Clustering

The process of grouping a set of objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression.

4. K-means Algorithm

The K-Means Clustering is a Centroid based clustering model in which the database is partitioned into K clusters in which each record belongs to the cluster with the nearest mean value. The algorithm starts with given initial set of mean values and allocates each object to a cluster with nearest mean value. The mean values for each cluster are calculated then using the elements in each cluster.

Let D be a data set containing n elements or objects to be clustered and k the number of clusters to be formed where $k \leq n$. The clusters formed are represented as Gi where i = 1 to k. The mean values of each cluster Gi is represented as Mi. When the clustering process is started initial mean values m1, m2 …… mk are identified from the given set of elements or objects. In the next step, the elements or objects are selected one by one and the distance between the element and each of the cluster means mi is found. The mean squared error E for each element p from the various clusters Gi is calculated for finding the strength of the clustering technique as follows:

$$E = \sum_{i=1}^{k} \sum_{p \in Gi} | p - Mi |^2 \qquad (1)$$

The elements are added to the clusters for which the mean squared error is lesser. The mean values of the clusters are recalculated and the process is repeated until there is no change in the cluster means.

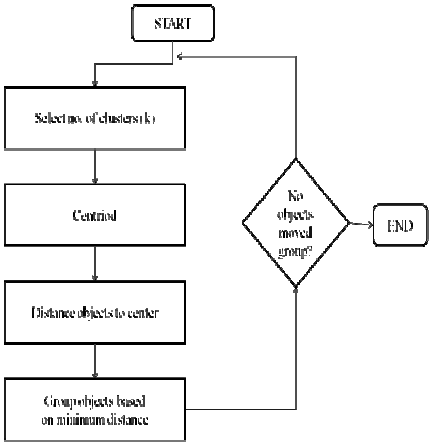

Fig. 1 Flowchart of k-means algorithm

File Edit Format View Help

chr	KOBIC	SNV	LENGTH	ID	Variant_seq	reference_seq	Variant_reads	Total_reads	Genotype
chr10	KOBIC	SNV	1230519	rs7081523	T	C	11	11	homozygous
chr10	KOBIC	SNV	1230596	rs6560719	T	A	20	20	homozygous
chr10	KOBIC	SNV	1230966	rs7099511	G	A	16	16	homozygous
chr10	KOBIC	SNV	1231340	rs7085017	T	G	13	14	homozygous
chr11	KOBIC	SNV	187337	rs3802984	A	G	19	20	homozygous
chr11	KOBIC	SNV	187557	rs7103852	G	A	23	23	homozygous
chr11	KOBIC	SNV	188986	rs4980320	C	T	16	16	homozygous
chr11	KOBIC	SNV	189673	rs4029235	C	T	13	13	homozygous
chr12	KOBIC	SNV	102988	rs4980977	A	G	18	19	homozygous
chr12	KOBIC	SNV	103513	rs4980978	G	A	19	19	homozygous
chr12	KOBIC	SNV	103523	rs4980979	C	T	18	18	homozygous
chr12	KOBIC	SNV	103677	rs11503141	A	G	19	22	homozygous

Fig. 2 A sample input file, k=3

The basic algorithmic steps are as follows:

1. Select k points as the initial centroids.
2. Repeat
3. Form k clusters by assigning all point to the closest centroid.
4. Re-compute the centroid of each cluster.
5. Until the centroid don"t change.

5. Result

Data format of sample human data

Table 1: Attributes of input file

Attribute name	Description
Chr	Chromosome number
KOBIC	Korean Bio Information Center
SNV	Single nucleotide variants
Length	Length of chromosome
ID	ID of chromosome
Variant_seq	All sequences found in individual at different locations
Reference_seq	The sequence from the reference sequence corresponding to the start and end coordinates of this feature
Variant_reads	No. of Variant_seq occurred
Total_reads	Variant longer than one nucleotide should be average read counts for each position over length of variant
Genotype	Genome type

File Edit Format View Help

chr	NAME	LENGTH ID	Variant_reads	Total_reads
chr11	chr11	187337.000000	19.000000	20.000000
chr11	chr11	187557.000000	23.000000	23.000000
chr11	chr11	188986.000000	16.000000	16.000000
chr11	chr11	189673.000000	13.000000	13.000000
chr12	chr12	102988.000000	18.000000	19.000000
chr12	chr12	103513.000000	19.000000	19.000000
chr12	chr12	103523.000000	18.000000	18.000000
chr12	chr12	103677.000000	19.000000	22.000000
chr10	chr10	1230519.000000	11.000000	11.000000
chr10	chr10	1230596.000000	20.000000	20.000000
chr10	chr10	1230966.000000	16.000000	16.000000
chr10	chr10	1231340.000000	13.000000	14.000000

Fig. 3 Output obtained on first run

File Edit Format View Help

Chr	NAME	LENGTH	Variant_reads	Total_reads
chr11	chr11	187337.000000	19.000000	20.000000
chr11	chr11	187557.000000	23.000000	23.000000
chr11	chr11	188986.000000	16.000000	16.000000
chr11	chr11	189673.000000	13.000000	13.000000
chr12	chr12	102988.000000	18.000000	19.000000
chr12	chr12	103513.000000	19.000000	19.000000
chr12	chr12	103523.000000	18.000000	18.000000
chr12	chr12	103677.000000	19.000000	22.000000
chr10	chr10	1230519.000000	11.000000	11.000000
chr10	chr10	1230596.000000	20.000000	20.000000
chr10	chr10	1230966.000000	16.000000	16.000000
chr10	chr10	1231340.000000	13.000000	14.000000

Fig. 4 Output obtained on second run

File Edit Format View Help

chr	NAME	LENGTH ID	Variant_reads	Total_reads
chr10	chr10	1230966.000000	16.000000	16.000000
chr10	chr10	1231340.000000	13.000000	14.000000
chr12	chr12	102988.000000	18.000000	19.000000
chr12	chr12	103513.000000	19.000000	19.000000
chr12	chr12	103523.000000	18.000000	18.000000
chr12	chr12	103677.000000	19.000000	22.000000
chr11	chr11	187337.000000	19.000000	20.000000
chr11	chr11	187557.000000	23.000000	23.000000
chr11	chr11	188986.000000	16.000000	16.000000
chr11	chr11	189673.000000	13.000000	13.000000
chr10	chr10	1230519.000000	11.000000	11.000000
chr10	chr10	1230596.000000	20.000000	20.000000

Fig. 5 Output obtained on third run

Paragraph Insert

chr	GROUP
chr11	0
chr11	0
chr11	0
chr11	0
chr12	1
chr12	1
chr12	1
chr12	1
chr10	2
chr10	2
chr10	2
chr10	2

Fig. 6 Clusters formed at end

It can be observed from above screenshots that objects are assigned in different clusters each time and their mean is calculated using Euclidian distance. The k-means algorithm runs three times each time forming a new cluster. These objects are grouped together with respect to length of chromosome attribute.

6. Conclusion

Human genome data was analyzed in this paper. The format of the human genome data analysis result was described and few of the attributes were selected for processing, based on the knowledge. The KDD steps were explained and were applied on the Human genome Data to convert the raw data into a transformed data that was used for generating more knowledge from the system. Various clusters are formed based on the various numerical attributes of the human genome data. Observing the data if any attributes numerical value is absent an outlier detection technique applied. If missing value is error then value can be inserted however if value is not present then analysis regarding to missing value can be done and genetic disorders can be identified.

References

[1] Kazuki Ichikawa et. Al., "A simple but powerful heuristic method for accelerating k-means clustering of large-scale data in life science", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. no. - PP, Issue no -99, pp – 1- 12, 2014.

[2] D. Minnie et. Al., "Clustering the preprocessed bone marrow data using modified k-means algorithm", Indian Journal of Computer Science and Engineering, Vol. no.- 4, Issue no.- 2, pp- 196-203, 2013.

[3] Alp Aslan dogan et. Al. "Evidence Combination in Medical Data Mining", International Conference on Information Technology: Coding and Computing , Vol. no- 2, pp – 465-469 , 2004.

[4] Rakesh Agrawal, et. Al., "Database Mining: A Performance Perspective", IEEE Transactions on Knowledge and Data Engineering, Vol. no- 5, Issue no.- 6, pp – 914-925, 1993.

[5] Patricia Cerrito, et. Al., "Data and Text Mining the Electronic Medical Record to Improve Care and to Lower Costs", SUGI 31 Proceedings, paper- 077-31, 2006.

[6] K. Y. Yeung, et. Al., "Validating clustering for gene expression data", Bioinformatics Oxford Journal, Vol. no.- 17, Issue no.- 4, pp-309-318, 2001.

[7] Cios KJ, et. Al., "Uniqueness of Medical Data Mining", Artificial Intelligence in Medicine,pp-1-24, 2002.

[8] Berks, Georg, et. Al., "Fuzzy clustering-a versatile mean to explore medical database." Program on European Symposium on Intelligent Techniques, Aachen, Germany, 2000.

[9] J. Harrow, et. Al., "GENCODE: the reference human genome annotation for The ENCODE Project," Genome Research, vol. 22, no. 9, pp. 1760-74, 2012.

[10] M. H. Fulekar, Book on Bioinformatics: Applications in Life and Environmental Sciences: Springer, pp-1-11, 2009.

[11] F. De Smet, et. Al., "Adaptive quality-based clustering of gene expression profiles," Bioinformatics, vol. 18, no. 5, pp. 735-46, May, 2002.

[12] Michael B. Eisen, et. Al., "Cluster Analysis and Display of Genome-Wide Expression Patterns", Proc. Natl. Acad. Sci. USA Vol. 95, Pp. 14863–14868, 1998.

[13] Anil K Jain, "Data clustering: 50 years beyond K-means", Journal Pattern Recognition Letters, Volume no.- 31, Issue no.- 8,pp- 651-666, 2010.