THE INFLUENCE OF CLUSTERING TECHNIQUES IN THE DIAGNOSIS OF GENETIC DISORDERS

B. Lakshmipathi[1] and G. Kousalya[2]
[1]Department of Computer Science Engineering, University College of Engineering Kanchipuram, Kanchipuram, India
[2]Department of Computer Science Engineering, Coimbatore Institute of Technology, Coimbatore, India
E-Mail: lkpathi2008.ucek@gmail.com

ABSTRACT

Clustering is a process of putting similar data into groups. Clustering has considered the most important unsupervised learning technique so, as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data. This paper reviews the six types of clustering techniques namely k-Means, Hierarchical, DBSCAN, OPTICS, STING. All these algorithms have compared according to the factors: size of dataset, the number of clusters, types of dataset and the type of software used. Some conclusions that have extracted belong to the performance, quality, and accuracy of the clustering algorithms.

Keywords: data clustering, DNA, medical data mining.

1. INTRODUCTION

Clustering is a division of data into groups of similar objects. Each group, called clusters, consists of objects that are similar amongst them and dissimilar compared to objects of other groups. Representing data from fewer clustering necessarily loses certain fine details, but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Cluster analysis is the organization of a collection of patterns into clustering based on similarity (Simpson *et al*., 2002). Patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. It is important to understand the difference between clustering and discriminated analysis. In supervised classification, we had a collection of labeled patterns. Typically, the given labeled patterns used to learn the descriptions of classes, which in turn had used to label a new pattern. In the case of clustering, the problem is to group a given collection of unlabelled patterns into meaningful clustering. In a sense, labels are associated with clusters also, but these category labels are data driven, that is they are obtaining (Kamath *et al*., 2011) solely from the data.

Cluster analysis had be used as a standalone data-mining tool to gain insight into the data distribution, or as a preprocessing step for other data mining algorithms operating on the detected clusters (Liu *et al*., 2006). Many clustering algorithms had developed and categorized from several aspects such as partitioning methods, hierarchical methods, density-based methods, and grid-based methods. Further data set can be numeric or categorical. Inherent geometric properties of numeric data had exploited to naturally, define a distance function between data points. Categorical data have derived from either quantitative or qualitative data, where observations directly observed from the counts.

2. MATERIALS AND METHODS

In this section, we describe various clustering techniques towards diagnosis of genetic disorders.

2.1. k-Means

It is a partition method, technique that has finds mutual exclusive clusters of spherical shape. It generates a specific number of disjoint, flat (non-hierarchical) clusters. Statically method had used to cluster to assign rank values to the cluster categorical data (Liu *et al*., 2007). Here categorical data have been converting into numeric by assigning rank value.

K-Means algorithm organizes objects into k – partitions where each partition represents a cluster. We start out with the initial set of means and classify cases based on their distances to their centers. Next, we compute the cluster means again, using the cases that are assign to the clusters; then, we reclassify all cases based on the new set of means. We keep repeating this step until cluster means do not change between successive steps. Finally, we calculate the means of cluster once again and assign the cases to their permanent clusters.

k-MEANS algorithm properties

- There are always K clusters.There is always at least one item in each cluster.
- The clusters are non-hierarchical and they do not overlap.
- Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'centre' of clusters.

k-MEANS algorithm process

Given an initial set of k means $m_1^{(1)},\ldots,m_k^{(1)}$ (see below), the algorithm proceeds by alternating between two steps.

Assignment step

Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean.[8] (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means).

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \le \|x_p - m_j^{(t)}\|^2 \; \forall j, 1 \le j \le k\},$$

where each x_p is assigned to exactly one $S^{(t)}$, even if it could be assigned to two or more of them.

Update step

Calculate the new means to be the centroids of the observations in the new clusters.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

Since the arithmetic mean is a least-squares estimator, this also minimizes the within-cluster sum of squares (WCSS) objective.

- Calculate the distance from the data point to each cluster.
- If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.
- Repeat the above step until a complete pass through all the data point results in no data point moving from one cluster to another. At this point, the clusters are stable and the clustering process ends.
- The choice of initial partition can greatly affect the final clusters that result (Beck *et al*., 2008), in terms of inter-cluster and intra cluster distances and cohesion.

2.2. Hierarchical

A hierarchical method creates a hierarchical decomposition of the given set of data objects. Here the tree of clusters called as dendrograms has built (Ye *et al*., 2009). Every cluster node contains child clusters, sibling cluster partition the points covered by their common parent. In hierarchical clustering, we assign each item to a cluster such that if we have N items then we have N clusters. Find closest pair of clusters and merge them into a single cluster. Compute distance between new cluster and each of old clusters. We have to repeat these steps until all items have clustered into K no. of clusters (Ehrich *et al*. 2008). It is of two types:

Agglomerative

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc., It starts by letting each object forms, its own cluster and iteratively merges cluster into larger and larger clusters, until all the objects are in a single cluster or certain termination condition is satisfied. The single cluster becomes the hierarchy's root (Jason *et al*., 2005). For the merging step, it finds the two clusters that are closest to each other, and combine the two to form one cluster (Kaufman *et al*., 2005).

Let X = {x1, x2, x3, ..., xn} be the set of data points.

a) Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.
b) Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.
c) Increment the sequence number: m = m +1.Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)].
d) Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: d[(k), (r,s)] = min (d[(k),(r)], d[(k),(s)]).
e) If all the data points are in one cluster then stop, else repeat from step 2).

Divisive

A top-down clustering method and is less commonly used. It works in a similar way to agglomerative clustering, but in the opposite direction. This method starts with a single cluster containing all objects, and then successive splits resulting clusters until only clusters of individual objects remain (Joshua-Tor *et al*., 1992).

2.3. DBSCAN

Density Based Spatial Clustering of Application with Noise (DBSCAN), it grows clusters according to the density of neighborhood objects (Secker *et al*., 2010). It is based on the concept of "density reach ability" and "density connect ability", both of which depends upon input parameter- size of epsilon neighborhood e and minimum terms of local distribution of nearest neighbors. Here e parameter controls the size of neighborhood and size of clusters. It starts with an arbitrary starting point that has not been visited (Ladd-Acosta *et al*., 2007). The point's e-neighborhood has retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point has labeled as noise. The number of point's parameter impacts detection of outliers.

2.4. OPTICS

OPTICS (Ordering Points to Identify Clustering Structure) is a density based method that generates an augmented ordering of the data's clustering structure. It had a generalization of DBSCAN (Meissner *et al*., 2008) to multiple ranges, effectively replacing the e parameter with a maximum search radius that mostly affects performance. Essentially becomes the minimum cluster size to find and then mints. It is an algorithm for finding density-based clusters in spatial data, which addresses one of DBSCANS major weakness (Suzuki *et al*., 2008) i.e. of detecting meaningful clusters in data of varying density.

It outputs cluster ordering which is a linear list of all objects under analysis and represents the density-based clustering structure of the data. Here parameter epsilon is not necessary and set to maximum value. OPTICS abstracts from DBSCAN by removing this each point has assigned as core distance, which describes the distance to its Min Pts point. Both the core-distance and the reach ability-distance are undefined if no sufficiently dense cluster with respect to epsilon parameter is available (Meissner *et al*., 2008).

2.5. STING

STING (Statistical Information Grid) is a grid-based multi resolution-clustering technique in which the embedded spatial area of input object have divided into rectangular cells. Statistical information regarding the attributes in each grid cell, such as the mean, maximum, and minimum values are stored as statistical parameters in these rectangular cells. The quality of STING clustering depends on the granularity of the lowest level of grid structure as it uses a multi resolution approach to cluster analysis. Moreover, STING does not consider the spatial relationship between the children and their neighboring cells for construction of a parent cell (Schut *et al*., 2012). As a result, the shapes of the resulting clusters are aesthetic, that is, all the cluster boundaries are either horizontal or vertical, and no diagonal boundary has detected. It approaches to cluster result of DBSCAN if the granularity approaches 0. Using count and cell size information, dense clusters have identified approximately using STING (Mecca *et al*., 2007).

3. EXPERIMENTAL SETUP

3.1. Genetic data set

An extensive web search has performed to find some of the data-clustering algorithm's implementations to test on. There are many protein data banks and many tools are available. After selections, we ended up with two of them. The dataset can be found at [27] [28] [29].

The Protein Data Bank (PDB) is a repository for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids [21]. The data typically obtained by X-ray crystallography or NMR spectroscopy and submitted by biologists and biochemists from around the world.

The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB. There are three different kind of protein database formats are available as below.

PDBe: PDBe is the European resource for the collection, organization and dissemination of data on biological macromolecular structures. In collaboration with the other worldwide Protein Data Bank (wwPDB) partners - the Research Collaboratory for Structural Bioinformatics (RCSB) and BioMagResBank (BMRB) in the USA and the Protein Data Bank of Japan (PDBj) - we work to collate, maintain and provide access to the global repository of macromolecular structure data [26].

PDBj: PDBj (Protein Data Bank Japan) maintains a centralized PDB archive of macromolecular structures and provides integrated tools, in collaboration with the RCSB, the BMRB in USA and the PDBe in EU. PDBj have supported by JST-NBDC and Institute for Protein Research IPR, Osaka University.

RCSB: The Research Collaboratory for Structural Bioinformatics (RCSB) is dedicated to improving our understanding of the function of biological systems through the study of the 3-D structure of biological macromolecules. RCSB members work cooperatively and equally through joint grants and subsequently provide free public resources and publications to assist others and further the fields of Bioinformatics and biology.

3.2. Bio software

UGENE: UGENE is free, open-source Bioinformatics software that helps biologists to analyze various biological data, such as sequences, annotations, multiple alignments, phylogenetic trees, NGS assemblies, and others. The data can be stored both locally (on a personal computer) and on a shared storage. UGENE integrates dozens of well-known biological tools and algorithms, as well as original tools in context of genomics, evolutionary biology, virology and other branches of life science. UGENE provides a graphical interface for the pre-built tools so biologists.

JMOL: Jmol is a free, open source molecule viewer for students, educators, and researchers in chemistry and biochemistry, which gives High-performance 3D rendering without hardware requirements and support more that 35 file formats of the biomedical data.

4. RESULTS AND DISCUSSIONS

The five clustering algorithms compared according to the following factors.

- Size of the dataset

- Number of clusters
- Dataset types
- Software used
- Accurancy of the detection

For each factor, four tests made one for each algorithm. For example, according to the size of data each of the five algorithms has executed twice, first by trying a huge dataset and then by trying a small data set. and totally 140 data sample were taken among that 40 are genetic disordered DNA.

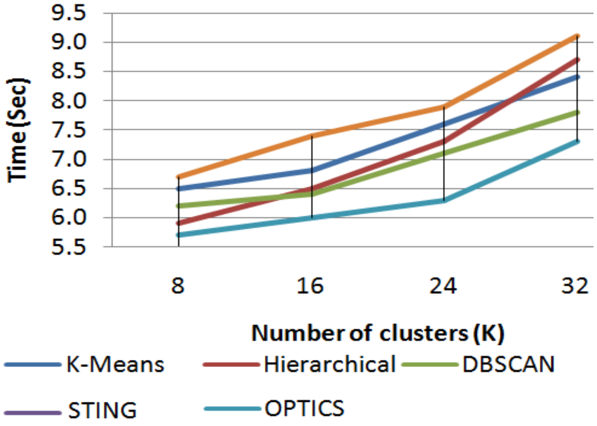

Figure-1. The relationship between number of clusters and the performance of the algorithms.

Table-1 explained how the four algorithms are compared. The total number of times the algorithms have executed is 32. For each 8-runs group, the results of the executions studied and compared and the conclusions written down. This step has repeated for all the factors.

Table-1. The factors consider for comparison.

Parameters	Consideration
The size of the dataset	Huge & small data set
Number of clusters	Large and Small number of clusters
Type of dataset	Ideal and random dates
Type of software	UGENE & JMOL

According to the number of clustering, K except for hierarchical clustering, all clustering algorithm compared here require setting k in advance. Here, the performance of different algorithms for different k's has compared in order to test the performances that have related in order to test the situation and to make the comparison easier, k has chosen equal to 8, 16, 24 and 32.

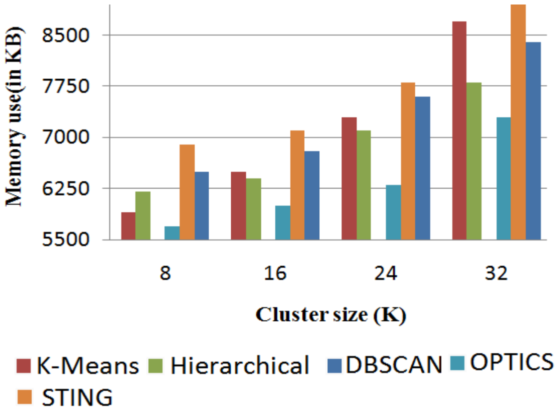

Figure-2. Memory usage of the algorithms.

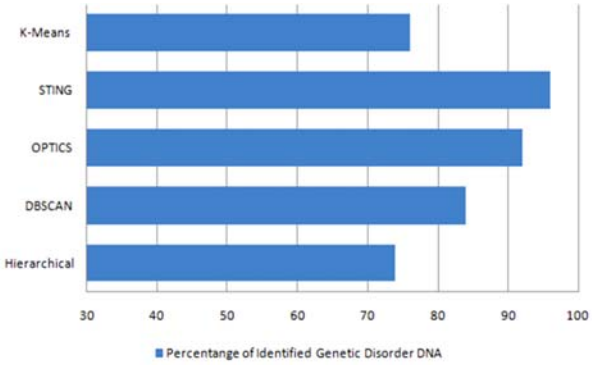

Figure-3. Percentage of identified genetic disorder DNA.

However, running the clustering algorithms using any one of the software gives the same results even when changing any of the other three factors (Dataset size, clustering number and dataset type). This, we believe, is because most software uses the same procedures and ideas in any algorithm implemented by them.

5. CONCLUSIONS

After analyzing the results of testing the clustering algorithms and running them under different factors and situation, the following conclusion has obtained:

- As a number of clusters, k becomes greater the performance of algorithms becomes lower
- The performance of k-means and optics is better than hierarchical clustering, dbscan and sting algorithms.
- Optics shows more accuracy in classifying most the objects into their suitable clustering than other algorithms.
- As a value of k becomes greater, the accuracy of the optics clustering becomes better until it reaches the accuracy.

- All the algorithms have some ambiguity in some (noisy) data when clustered.
- As a general conclusion, optics and k-means algorithms are giving better results compared to others, when using random datasets and the vice versa

6. FUTURE WORK

As a future work, comparison between these five algorithms (or may other algorithms) can be attempted according to different factors other than those considered in this paper. One important factor is accuracy and building of decision tree. This experiment, we performed for the genetic disorders of course we can test using other DNA samples also these may affect the performance of the algorithm and the quality of the result.

REFERENCES

P. E. Paramo and *et al*. 2012. Large-scale population structure of human commensal escherichia coli isolates. Applied and Environmental Microbiology. 70: 5698-5700.

D. M. Gordon and A. Cowling. 2013. The distribution and genetic structure of Escherichia coli in Australian vertebrates: host and geographic effects. Microbiology. 149: 35-75.

D. M. Gordon. 2010. Strain typing and the ecological structure of escherichia coli. Journal of AOAC International. 93: 974-984.

J. M. Simpson, J. W. S. Domingo and D. J. Reasoner. 2002. Microbial source tracking: State of the science. Environmental Science and Technology. 36: 5279-5288.

K. Y. Kamath and J. Caverlee. 2011. Transient crowd discovery on the real-time social web. In Proceedings of the fourth ACM international conference on Web search and data mining, ser. WSDM '12. New York, NY, USA: ACM. pp. 585-594.

J. Liu, Q. Zhang, W. Wang, L. McMillan and J. Prins. 2006. Clustering pair-wise dissimilarity data into partially ordered sets. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD '06. New York, NY, USA: ACM. pp. 637-642.

J. Liu, Q. Zhang, W. Wang, L. Mcmillan and J. Prins. 2007. Poclustering: Lossless clustering of dissimilarity data. In SIAM International Conference on Data Mining.

B. Liu. 2006. Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data. Springer.

Beck S. and V.K. Rakyan. 2008. The methylome: approaches for global DNA methylation profiling. Trends Genet. 24:231-237.

Christian Rohde. 2009. New clustering module in BDPC bisulfite sequencing data presentation and compilation web application for DNA methylation analyses. Bio Techniques. 47(3): 781-783.

Ehrich, M., J. Turner, P. Gibbs, L. Lipton, M. Giovanneti, C. Cantor and D. van den Boom. 2008. Cytosine methylation profiling of cancer cell lines. Proc. Natl. Acad. Sci. USA. 105: 4844-4849.

Farthing, C.R., G. Ficz, R.K. Ng, C.F. Chan, S. Andrews, W. Dean, M. Hemberger and W. Reik. 2008. Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes. PLoS Genet. 4:e1000116.

Jason T.L. Wang, Mohammed J. Zaki, Hannu T.T. Toivonen and Dennis Shasha (Eds), 2005. Data Mining in Bioinformatics, ISBN 1852336714, Springer-Verlag London Limited.

Joshua-Tor L., Frolow F., Apella E., Hope H., Rabinovich D. and Sussman J. L. 1992. The three-dimensional structures of bulge-containing DNA fragment Journal of Molecular Biology. 225, 397.

Kaufman L. and P.J. Rousseeuw. 2005. Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley and Sons, Inc., Hoboken, NJ.

Ladd-Acosta C., J. Pevsner S. Sabunciyan, R.H. Yolken, M.J. Webster, T. Dinkins, P.A. Callinan J.B. Fan. 2007. DNA methylation signatures within the human brain. Am. J. Hum. Genet. 81: 1304-1315.

Meissner A., T.S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B.E. Bernstein. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 454: 766-770.

Suzuki M.M. and A. Bird. 2008. DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev, Genet. 9: 465-476.

G. Mecca et al. 2007. A new algorithm for clustering search results. Data and Knowledge Engineering. 504-522 (Pubitemid 46726850).

Ben-Dor, D. Lipson, A. Tsalenko , M. Reimers, L. Baumbusch, M. Barrett, J. Weinste in, A. Borresen-Dale and Z. Yakhini. 2007. Framework for Identi fying Common Aberrations in DNA Copy Number Data. RECOMB, LNBl4453. pp. 122-136.

Abdullah Alqallaf and Ahmed Tewfik. 2009. Maximum Likelihood Principle for DNA Copy Number Analysis. IEEE Int'l Conference on Acoustics, Speech, and Signal Processing, IEEE/ICASSP, Taipei, Taiwan.

X. Ji, J. Bailey, and G. Dong. 2007. Mining minimal distinguishing subsequence patterns with gap constraints. Knowl. Inf. Syst. 11(3): 259-286.

L. Ye and E. Keogh. 2009. Time series shapeletes: A newprimitive for data mining. In KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.

A. Secker, E. B. Clark. 2010. Hierarchical classification of GProtein-Coupled Receptors with data driven selection of attributes and classifiers. Journal of data mining, Inderscience.

M. H. Schut, J. Bullock, S. Patassini, E. Kim. 2012. Analysis of huntingtin protein fragments in post mortem human Huntington's disease brain tissue. Journal of Neurology.

R. D. Delima, A. C. G. Chua, J. E. E. Tirnitz Parker. 2012. Disruption of hemochromatosis protein and transferring receptor 2 causes iron induced liver injury in mice. Wiley Online Library.

www.rcsb.org.

www.ebi.ac.uk.

www.wwpdb.org.