

PROGRESSIVE DATA ANALYTICS IN HEALTH INFORMATICS USING AMAZON ELASTIC MAPREDUCE (EMR)

J.S. Shyam Mohan and P. Shanmugapriya

Department of Computer Science and Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University, India
E-mail: jsshayammohan@kanchiuniv.ac.in

Department of Information Technology, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University, India
E-mail: priya_prakasam@yahoo.co.in

Abstract

Identifying, diagnosing and treatment of cancer involves a thorough investigation that involves data collection called big data from multi and different sources that are helpful for making effective and quick decision making. Similarly data analytics is used to find remedial actions for newly arriving diseases spread across multiple warehouses. Analytics can be performed on collected or available data from various data clusters that contains pieces of data. We provide an effective framework that provides a way for effective decision making using Amazon EMR. Through various experiments done on different biological datasets, we reveal the advantages of the proposed model and present numerical results. These results indicate that the proposed framework can efficiently perform analytics over any biological datasets and obtain results in optimal time thereby maintaining the quality of the result.

Keywords:

Big Data, Data Analytics, MapReduce, Amazon EMR, Predictive Analysis

1. INTRODUCTION

Data collected from various sources like satellites, sensors, mobile phones, etc. is known as big data as most of the data is unstructured. Big data analytics aims at discovering hidden patterns for quick and effective decision making. Majority of the analytics are done by batch processing systems built on top of the Hadoop [1].

Health Informatics field is drastically gaining importance due to its importance and capability of handling big data [2]. Data mining along with big data analytics are helpful for diagnosing and treatment of various diseases. Healthcare Informatics deals with a variety of researches like Bioinformatics, clinical informatics and Trans Bioinformatics called as TBI, etc. Example, Zika virus caused by mosquito can be effectively handled by using these techniques that rely on advanced and sophisticated tools for processing.

1.1 RESEARCH IN HEALTH INFORMATICS

Research in health informatics is given below:

1. Bioinformatics - At molecular level
2. Neuro-informatics - At tissue level
3. Clinical Informatics - At patient level
4. Public Health Informatics - Utilizing population data

The main scope of health informatics deals with integrating data from various sources like laboratory level and clinical data and thereby striving to find the remedial measures for the

diseases [3]. This entire data access is possible by using data mining at all levels and hence making decision making easier.

Advancements in information technology have lead to increase in voluminous data that are being captured from small range sensors to astronomical data and also in health informatics [4]. Many microarray data repositories have been extensively used for gene prediction and mutations [5]. On the other side several sophisticated cameras generating a huge amount of data used in security surveillance. Square Kilometre Array Telescope generating several petabytes of astronomical data every year poses a great challenge [6, 7].

Computing systems for big data generally categorized into Batch processing and in memory stream processing. In batch processing large volumes of data are analysed and the in the later case data is analysed in real time or for a short period of time [8].

2. CHALLENGES AND PROBLEM STATEMENT

Medical data can be accessed by many sources that provide information relating to any gene viz., mRNA or protein sequence [9]. Most of the medical data is maintained and updated by individual organizations and are kept in house. DNA sequences, protein sequences are updated frequently for knowing the effects and causes of the above sequences. For example, cancer data is decentralized and is exponentially growing. For testing or for diagnosis of any medical data is a difficult task that involves periodic updation done at regular intervals.

Advancements in healthcare systems have brought radical changes in the era of big data. Performing analytics in healthcare consisting of huge voluminous and high dimensional biological data will provide predictive platforms for making effective decision making. Researchers have been constantly working on data analysis techniques for big data in medicine and healthcare systems using many advanced tools and techniques.

Typically, the researchers should have a clear idea or purpose of performing analytics for biological data [10]. Some the fundamental questions include:

1. At what optimal time can we get results?
2. What are the sources for data collection?
3. Where can I get biological data?

Some of the applications of Big Data are given below [11, 12]:

1. In Business- For customer personalization, reviews etc.
2. In Technology- optimal time from hours to minutes.

3. In Health –Finding DNA monitoring and improving the health aspects.
4. For Smart cities: Focuses on management of natural resources.

We resort ourselves on health informatics as mentioned in application three as our major aspect of discussion throughout the paper.

Effective data analytics can be done only by knowledge discovery in databases (KDD). KDD is based on the following steps: data gathering, selection, preprocessing, transformation, data mining, evaluation and interpretation [13].

Effective analytics can be done in optimal time by cluster analysis because Biological datasets are huge in size and high-dimensional. Clustering provides insights into the huge biological datasets by automatically organizing distinct patterns. In genomics, cluster analysis has been used widely for finding gene expression profiles. Recent advancements in biology and high-throughput technology has brought a wider potential for cluster analysis in biological research. Example, identification of cell phenotypes based on quantitative image metrics [14].

3. BACKGROUND AND LITERATURE SURVEY

According to Human Genome Project estimates, the human genome DNA contains around 3.2 billion base of pairs distributed among twenty-three chromosomes translated about a gigabyte of information [15]. By adding gene data, X-ray and NMR spectroscopy data, the volume increases dramatically in gigabytes or petabytes.

Many works are done in clustering huge datasets in literature. Clustering based on swarm intelligence simulates the process of changes during population for biological data. Ex: ACO based algorithm [16]. DBCLASD algorithm works on the principle of data generated from similar distribution belongs to the same cluster if there are several distributions for the original data [17]. There are genetic algorithms that work over a variety of genes.

Some of the data collected from various repository are shown in Table.2 which is shown in page 3 and also can be found in references [18].

4. ABOUT AMAZON ELASTIC MAPREDUCE (EMR)

Amazon EMR works on Hadoop that is used for data processing and analysis. It provides a good way to perform in house clustering. The speciality of using MapReduce is that it provides a framework that allows users to write and run hundreds of programs that run simultaneously in parallel capable of handling huge datasets. Hadoop clusters are processed by using Amazon Elastic Compute Cloud (EC2). Some of the applications where Amazon EMR are log analysis, web indexing, machine learning, bioinformatics, etc. Due to this interesting aspect, we took Amazon EMR for healthcare systems [19].

Advantages of Amazon EMR:

- It is easy to use.

- Less cost.
- It is easy and simple to use.
- It is secure and flexible.

4.1 PROGRESSIVE ANALYTICS FOR BIOLOGICAL DATASETS

Progressive analytics for different and high dimensional biological datasets is proposed that performs analytics and used for effective and quick decision making. We take Progeny Clustering technique for clustering the high dimensional biological datasets. This make easy for EMR for further processing. This technique considers even the partial or missing data and aims to find missing patterns and thereby providing accuracy in results. By performing successive progressive analytics on different samples, they get incrementally processed automatically providing a significant performance benefit. No specific schedulers are needed for during processing [20]. Introducing traditional algorithm techniques into an existing relational engine is easy because majority of the data appears in text. Implementing the same on unstructured data is a challenging task. The table below shows the input data (taken in numeric) with assumed progressive intervals.

Table.1. Input data with progressive intervals

Interval	User	Ad
$(0, \infty]$	user 0	a0
$[1, \infty]$	user 1	a1
$[2, \infty)$	user 2	a2

4.2 DEPLOYING AMAZON EMR OVER CLOUD WITH HADOOP MAPREDUCE DISTRIBUTION

As mentioned before, MapReduce Distribution in Hadoop makes Hadoop in AWS Cloud. MapReduce is used in real time dataset handling and can be used across various health care organizations. This is useful in cross platform deployment and is proved as the best platform for Big Data. Many companies have already started using MapReduce for their organizations.

4.2.1 Running Parallel Hadoop Jobs in Amazon EMR Cluster Using Progeny Clustering Algorithm:

For high dimensional biological datasets, progeny clustering algorithm categorizes datasets into clusters based on the density, similarity, and other parameters. Inputs are passed over a multi-stage framework. Each job consists of several types of data viz., a partitioning key (or mapper), and a progressive reducer. Progressive analytics are done by taking insights for processing of cluster data. The Hadoop jobs can be run in parallel in progeny clusters using Amazon EMR. Cluster utilization can be increased using EMR. A scheduler takes care of Amazon EMR clusters and monitors Hadoop activities and assigns them to specific queues. New data arriving during real time processing can be specified to core Amazon EMR nodes and are automatically assigned to clusters.

5. OPTIMIZING TIME FOR CLUSTERS

As mentioned before that there are no specific schedulers that specifically take care of jobs. Setting start and limit for the cluster data is done automatically based on the number of tasks assigned to Hadoop. We assume the progressive interval starts from 0. Optimal Stopping Theory (OST) is used to stop the cluster model [21][22] and the optimal or best time results are obtained based on sequential random variables defined as $T \in 0, 1, \dots, \infty$.

The collected data is considered as random independent variables and can be found in [23] which is considered as a finite or an infinite horizon. For finite horizon, the scheduler responds for a specific time interval and for an infinite horizon, the scheduler receives only a part of data and takes final decision in optimal time.

Table.2. Data Collected from various repository

S. No.	Repository	Sequence or category of Data	Data (Approx.)	Data Growth (Approx.)
1	Prism (Progressive sampling Model)	Progressive sampling data by tuples at explicit progress intervals	Progressive analytics on big data in the Cloud	Rely on pipelining techniques
2	Now	MapReduce computation	Binary Large Object Data(Blob)	Batch Processing
3	GeneBank (As on December 2014)	Nucleic acid sequences	178 million	Doubling in size for every 15 months
4	SWISS-PROT database	Protein sequences	18 million	Doubling in size for every 15 months
5	InSiteOne	Data archiving, storage, and disaster-recovery solutions to the health-care industry in US	4 billion medical images and 60 million clinical studies from 800 clinical sites	Increasing at an approximate rate of about 12% per year
6	ESG (Enterprise Storage Group)	Forecasting Medical image data	Grows at a rate of 35 percent per year	2.6 million terabytes (2014)

6. MULTI STAGE PROCESSING

6.1 METHOD

Let M be a finite dataset for M features for any N independent observations. K-Means clustering method partitions the data into

K clusters. Progeny clustering assumes heterogeneity of the datasets to reduce the computation costs. This entire setup itself is considered as a single dataset and is passed as input over the Amazon EMR for further processing. The process of Hadoop job assignment and scheduling of events is taken care by scheduler automatically. Further details about Progeny clustering can be found in [14].

Table.3. Parameters for K-Means clustering Method

Parameter values	Parameter
Lower limit	2
Upper Limit	10
Sampling size	20
iterations	100
Reference datasets	10

The configuration of scheduler is done by the following commands:

```
HADOOP_CONF_DIR/capacity-scheduler.xml
```

```
HADOOP_YARN_HOME/bin/yarn rmadmin -refreshQueues
```

6.1.1 Launching Instances in Amazon EMR:

Datasets can be launched on the top of Amazon EMR cluster with MapReduce version 4.0.2 from AWS Management Console. It supports many editions of the MapR viz., Community Edition (M3), Enterprise Database Edition (M7), etc. The algorithm below shows launching an instance in Amazon cluster [24].

6.1.2 Algorithm: Launching instances in EMR [25]:

Input: Mappers - $X = \{a, b, c, \dots\}$, $Y = \{1, 2, 3, \dots\}$

Output: new instances for collected datasets.

```
begin
id = collected datasets;
type = Cluster;
hadoopVersion = 0.20;
keypair = key value;
For each masterInstanceType = k1.xlarge do
If(coreInstanceType == k1.small )then continue
coreInstanceCount = 30;
if(taskInstanceType == k1.small)then continue
taskInstanceCount = 30
do bootstrapAction set from
D://elasticmapreduce/bootstrap-actions/configure-
hadoop,arg1,arg2,arg3", to
D3://elasticmapreduce/bootstrap-actions/configure-
hadoop/configure-other-stuff,arg1,arg2";
end
```

7. EVALUATION

Multi stage jobs are set up using Hadoop. Each job consists of input files, a partitioning key (or mapper). Progressive

analytics is done such that each job consists of a special reducer which uses progeny clustering algorithm to process cluster data. Amazon EMR can be deployed over a cluster of machines. Due to this interesting feature, we have considered Amazon EMR for performing progressive analytics for biological datasets.

8. EXPERIMENTAL SETUP

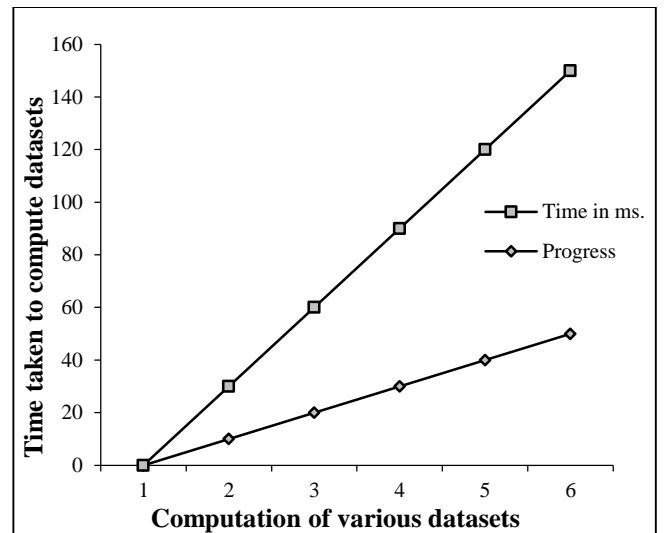
8.1 SYSTEM CONFIGURATION

EMR machines are configured using Virtual Private Cloud (VPC) by changing the host name settings. Instances communicate via EMR-managed security groups. Input and output data splits are stored as clusters. The cluster Id is known using VPC. The local system (Dell Laptop) is of the configuration, 16GB RAM, 2TB Hard Disk, and 2Gbps allocated I/O bandwidth. We took over 1000 instances for our tests.

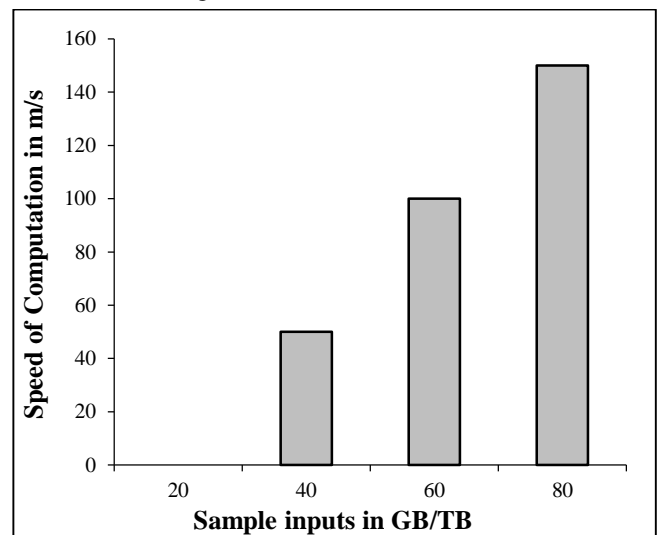
8.2 DATASETS

We used the datasets available in [26] for our evaluation. Under the brain classification, we choose MicroRNA datasets for performing analytics. Sample data sets for over 10000 patients for first iteration and then 20000 and so on based on the performance of the algorithm. Similarly we choose different data sources where the biological datasets are available and even cross mutation gene virus are considered that are collected from different sources which can be freely downloaded from UCI repository [27]. Intergenic factors are also considered for data analysis. Input splits are created by data shredding by partitioning with their corresponding Id [28]. Deterministic Stopping Model (DSM) is used for stopping the process that is considered for optimality. Missing data splits can be found using parametric, non- parametric approach or Weibull distribution approach. Evaluation can be performed on any type of datasets. Datasets can be downloaded for free from online. Users interested to perform evaluation for different datasets can apply the same methodology and test for the outcomes.

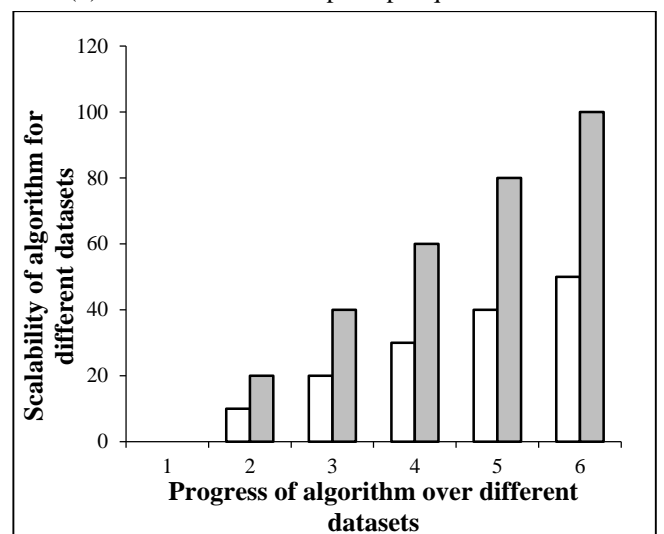
The Fig.1(a) shows computation for available biological datasets Amazon EMR. The Fig.1(b) shows the performance of sample input queries passed over Amazon EMR. The Fig.1(c) shows scalability of algorithm for increasing data sets. The Fig.1(d) shows throughput of the machines.



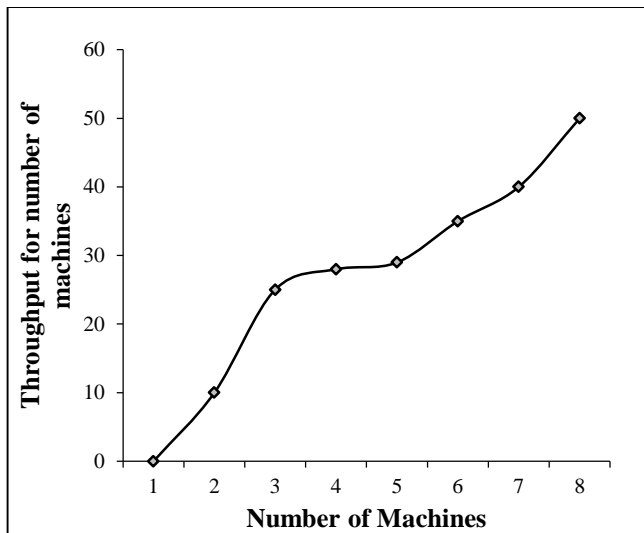
(a). Progressive Computation-Time taken to process available biological datasets in Amazon EMR



(b). Performance of sample input queries on EMR



(c). Algorithm scalability with increase in data size



(d). Throughput for number of machines

Fig.1. Analysis of various data sets on EMR

9. CONCLUSION

Progressive analytics on the available data is used to extract data for exploratory querying [29]. Due to lack of proper tools for progressive analytics, obtaining results is a tedious task. We took progeny algorithm for clustering the high dimensional biological datasets that when deployed over Amazon EMR [30] allows efficient and deterministic query processing for the passed datasets. For performing progressive sampling, we rely on Amazon EMR which provides a new framework for performing big data analytics over any huge and complex datasets where progress is the crucial part [31]. Combination of the all the above factors will lead to an effective and progressive big data analytics for any biological datasets in optimized time.

REFERENCES

- [1] S. Singh, N. Singh, "Big Data Analytics", *Proceedings of International Conference on Communication, Information and Computing Technology*, pp. 1-4, 2012.
- [2] M. Herland, T.M. Khoshgoftaar and R. Wald, "A Review of Data Mining using Big Data in Health Informatics", *Journal of Big Data*, Vol. 1, No. 2, pp. 1-35, 2014.
- [3] J. Chen, F. Qian, W. Yan and B. Shen, "Translational Biomedical Informatics in the Cloud: Present and Future", *BioMed Research International*, pp. 1-8, 2013.
- [4] T. Barrett, et al., "Ncbi Geo: Archive for Functional Genomics Data Sets Update", *Nucleic Acids*, Vol. 41, pp. 991-995, 2012.
- [5] P.E. Dewdney, P.J. Hall, R.T. Schilizzi and T. Joseph L.W. Lazio, "The Square Kilometre Array", *Proceedings of IEEE*, Vol. 97, No. 8, pp. 1482-1496, 2009.
- [6] Jianqing Fan, Fang Han and Han Liu, "Challenges of Big Data Analysis", *National Science Review*, Vol. 1, No. 2, pp. 293-314, 2014.
- [7] Amardeep Kaur and Amitava Datta, "A Novel Algorithm for Fast and Scalable Subspace Clustering of High-Dimensional Data", *Journal of Big Data*, Vol. 2, No. 17, pp. 1-24, 2015.
- [8] H. Howie Huang and Hang Liu, "Big Data Machine Learning and Graph Analytics: Current State and Future Challenges", *Proceedings of IEEE International Conference on Big Data*, pp. 16-17, 2014.
- [9] Insights learned from the human DNA sequence, Available at: <https://www.genome.gov/10002171/insights-from-the-human-dna-sequence/>
- [10] Fran Berman, Geoffrey Fox, Tony Hey and Anne Trefethen, "The Data Deluge: An e-Science Perspective", *Grid Computing-Making the Global Infrastructure a Reality*, 2003.
- [11] Genbank Statistics, Available at: <http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html>.
- [12] Uniprotkb/Swiss-Prot Protein Knowledgebase Release 2011_04 Statistics, Available at: <http://expasy.org/sprot/relnotes/relstat.html>.
- [13] Alejandro Baldominos, Esperanza Albacete, Yago Saez and Pedro Isasi, "A scalable machine learning online service for big data real-time analysis", *Proceedings of IEEE Symposium on Computational Intelligence in Big Data*, pp. 1-8, 2014.
- [14] Uniprotkb/trembl Protein Knowledgebase Release 2011_04 Statistics, Available at: <http://www.ebi.ac.uk/uniprot/TrEMBLstats>.
- [15] Dilpreet Singh and Chandan K. Reddy, "A Survey on Platforms for Big Data Analytics", *Journal of Big Data*, Vol. 1, No. 8, pp. 1-20, 2014.
- [16] Charu C. Aggarwal, "Managing and Mining Sensor Data", Springer, 2013.
- [17] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos, "Big Data Analytics: A Survey", *Journal of Big Data*, Vol. 2, No. 1, pp. 1-21, 2015
- [18] Chenyue W. Hu, Steven M. Kornblau, John H. Slater and Amina A. Qutub, "Progeny Clustering: A Method to Identify Biological Phenotypes", *Scientific Reports*, Vol. 5, pp. 1-12, 2015.
- [19] Tamara Baluja, "Electronic Patient Records will Soon End Doctors Scrawl on Paper", Available at: <http://www.theglobeandmail.com/news/national/toronto/electronic-patient-records-will-soon-end-doctors-scrawl-on-paper/article1982647>.
- [20] Julia Handl and Bernd Meyer, "Ant-Based and Swarm-Based Clustering", *Swarm Intelligence*, Vol. 1, No. 2, pp. 95-113, 2007.
- [21] Surajit Chaudhuri, Gautam Das and Utkarsh Srivastava, "Effective use of Block-Level Sampling in Statistics Estimation", *Proceedings of the SIGMOD International Conference on Management of Data*, pp. 287-298, 2004.
- [22] "Nuclear Cardiology Markets", TriMark Publications, pp. 1-252, 2011.
- [23] <http://searchaws.techtarget.com/definition/Amazon-Elastic-MapReduce-Amazon-EMR>.
- [24] Badrish Chandramouli, Jonathan Goldstein and Abdul Quamar, "Scalable Progressive Analytics on Big Data in the Cloud", *Proceedings of the VLDB Endowment*, Vol. 6, No. 14, pp. 1726-1737, 2013.

- [25] Goran Peskir and Albert Shiryaev, “*Optimal Stopping and Free Boundary Problems (Lectures in Mathematics. ZTH Zürich)*”, Birkhäuser, 2006.
- [26] Kostas Kolomvatsos, Christos Anagnostopoulos and Stathes Hadjiefthymiades, “An Efficient Time Optimized Scheme for Progressive Analytics in Big Data”, *Big Data Research*, Vol. 2, No. 4, pp. 155-165, 2015.
- [27] O.Y. Al-Jarrah, P.D. Yoob, S. Muhaidat, G.K. Karagiannidis and K. Taha, “Efficient Machine Learning for Big Data: A Review”, *Artificial Intelligence*, Vol. 2, No. 3, pp. 87-93, 2015.
- [28] <http://searchaws.techtarget.com/definition/Amazon-Elastic-MapReduce-Azure-EMR>.
- [29] <https://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn-site/CapacityScheduler.html>.
- [30] <http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi>.
- [31] J.S. Shyam Mohan, P. Shanmugapriya and N. Kumaran, “Data Reduction Techniques for High Dimensional Biological Data”, *International Journal of Research in Engineering and Technology*, Vol. 5, No. 2, pp. 319-324, 2016.