Categorizing high dimensional unlabelled genomic data

A dissertation submitted for the Degree of Master of Science in Computer Science

R.D.T.W.Ranasinghe
University of Colombo School of Computing
2019

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: R.D.T.W. Ranasinghe

Registration Number: 2016mcs088

Index Number: 16440882

____

Signature: Date:

This is to certify that this thesis is based on the work of

~~Mr.~~/Ms. R.D.T.W. Ranasinghe

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr. Ruvan Weerasinghe

____

Signature: Date:

Acknowledgement

Every work accomplished is a pleasure – a sense of satisfaction. However, a number of people always motivate, criticize and appreciate a work with their objectives, ideas and opinions. Hence, I use this opportunity to thank all, who have directly or indirectly helped me to conduct this study and prepare this report.

First, I would like to express my profound sense of gratitude to Dr. Ruvan Weerasinghe (Senior Lecturer, University of Colombo School of Computing) for assisting me in initiating this research study. He being my supervisor, despite of his busy schedule and workload, was able to find some time for me and helped me by giving necessary advices and guidance. I choose this moment to acknowledge his contribution gratefully.

I also feel very grateful to Miss Rupika Wijesinghe (Senior Lecturer, University of Colombo School of Computing) who is the co-supervisor of my research, for the guidance and constant supervision as well as for providing necessary information regarding the project & also for giving support in completing this research study.

Last but not least, I am thankful to all the lecturers of University of Colombo School of Computing for educating us throughout these years. I owe quite a lot to my family who provided me the chance to fulfill my career objectives and for the support throughout my studies. I would like to dedicate this study to them as an indication of their significance in this study as well as in my life.

Abstract

Since genomic data exploration became an important area with the completion of the Human Genome project, the tools and techniques that were used in genomic context were improved. These tools and techniques for data generation has increased the volume of data available to researchers and it is being increasing rapidly. However the high dimensional nature of these data make it difficult to analyze the presented data and make valuable conclusions or predictions.

These data are presented in different types of formats with several parameters in different data sources. Thousands of DNA combinations have been identified as indicators of susceptibility to specific diseases. Categorizing these data using there similarities which can be a hidden feature, will lead to reveal some important factors of these data collections.

Clustering is one of the major method is been used for data analyzing. In this study I present a novel approach to cluster the high dimensional genomic data in order to make important and valuable predictions on available data by taking into account the annotated information about genes on prostate cancers from online databases such as cBio portal.

These data has different characteristics as numerical, categorical, sparse and dense. Hence different normalizing methods and different clustering approaches. These different approaches were carried out having a base of three main clustering algorithms which are K-means, Hierarchical clustering and DBSCAN clustering. These clustering algorithms were used in different procedures using several dimensional reduction methods, different data normalizing methods. Each approach were evaluated using different measurements in order to find the better approach for genomic data clustering when the data are high dimensional.

Silhouette score and Davies–Bouldin index were used as the messurements of evaluation of each cluster in each approach. Selected novel hybrid approach of clustering genomic data gives the best scores for these meassurements confirming the validity of the novel approach in clustering high dimensional genomic data.

Table of contents

Chapter 1 Introduction 2
1.1 Overview 2
1.2 Background 2
1.3 Problem Overview 4
1.4 Problem Statement 6
1.5 Significance of the Study 7
1.6 Research Goals 8
1.7 Objectives 8
1.8 Research Scope 9
1.8.1 Only consider genomic data on Prostate cancer 9
1.8.2 Use CBIO portal as the data source 9
1.8.3 Consider only three selected Features on genomic data 10
Chapter 2 Literature Review 11
2.1 Overview 11
2.2 Related work 11
Chapter 3 Research Design & Methodology 15
3.1 Overview 15
3.2 Analyze nature of data and resources 15
3.2.1. Biological background 16
3.2.2. Data representation in selected data source 18
Chapter 4 Data Preparation 25
4.1 Overview 25
4.2 Data Collection and preparation 25
4.2.1 Data preparation 25
Chapter 5 Design & Implementation 29
5.1 Overview 29
5.2 Analyze the data set 29
5.3 Application of selected clustering methods 32
5.3.1. K- Means Clustering method 34
5.3.2. Hierarchical Clustering method 52
5.3.2. DBSCAN (Density-based spatial clustering of applications with noise) Clustering method .. 54

Chapter 6 Evaluation 65
6.1 Overview 65
6.2 External Evaluation 65
6.2 Internal Evaluation 70
Chapter 8 Results & Analysis 74
8.1 Overview 74
8.2 Results and analysis on each clustering approaches 74
Chapter 9 Conclusion 80
9.1 Overview 80
9.2 Summary 80
9.3 Limitations........ 80
9.4 Suggestions for Future Work 81
List of References: 83
Appendix A: Clusters when apply K-means on CNA data set 86
Appendix B: Clusters when apply K-means on CNA data set with TNSE dimensional reduction........ 88
Appendix C: Clusters when apply K-means on CNA data set with PCA dimensional reduction 90
Appendix D: Silhouette Score for each point 92
Appendix E: Clusters when apply K-means on CNA data set with ICA dimensional reduction........ 95
.... 95
Appendix F: Clusters performed by approach three........ 97
Appendix G: Clusters performed by approach four 99
Appendix H: Clusters performed by approach five........ 101
Appendix I: Clusters performed by approach six 103
Appendix J: Clusters performed by approach seven........ 106
Appendix K: Clusters performed by approach eight 109
Appendix L: Clusters performed by approach nine........ 112
Appendix M: Clusters performed by approach eleven 114

List of Figures

Figure 1: Increment of the total sequence in bp 3
Figure 2: Increment of data sources 5
Figure 3: Population scale cluster of 5 groups from chr 22 (between actual and predicted) 14
Figure 4: Block diagram of research design phase..... 15
Figure 5: Insertion 17
Figure 6: Deletion..... 17
Figure 7: Duplication 18
Figure 8: Meta file – Cancer study 20
Figure 9: Meta file – Cancer Type 20
Figure 10: Meta file – Clinical Data 20
Figure 11: Data file – Copy number alteration 21
Figure 12: Data file – Mutations 22
Figure 13: Data file – Expressions 23
Figure 14: Representation of data set combination 26
Figure 15: Before application of label encoder..... 27
Figure 16: Application of label encoder 27
Figure 17: Application of dummy variables 28
Figure 18: Cancer Type Classification 29
Figure 19: Sample Classification 30
Figure 20: Mutation count 30
Figure 21: Mutation count vs CNA..... 31
Figure 22: Tumor Disease anatomic site..... 32
Figure 23: Mutation types classification 34
Figure 24: Steps of K-means clustering..... 35
Figure 25: Selecting optimal number of clusters 37
Figure 26: Selecting optimal number of clusters after PCA 38
Figure 27: Approach one..... 38
Figure 28: Clusters using K-means without dimensional reduction 39
Figure 29: Approach two..... 40
Figure 30: Clusters after K- means with t-SNE dimensional reduction..... 42
Figure 31: Clusters after K- means with PCA dimensional reduction 45
Figure 32: Clusters after K- means with ICA dimensional reduction 46
Figure 33: Clusters applying one hot encoding on CNA data and then K-means clustering 49
Figure 34: Approach Three 50
Figure 35: Clusters after approach three..... 50
Figure 36: Approach four 51
Figure 37: Clusters after approach four..... 51
Figure 38: Steps of Hierarchical clustering algorithm..... 52
Figure 39: Dendogram using hierarchical clustering algorithm..... 53
Figure 40: Approach five 55
Figure 41: Clusters after approach five..... 57

Figure 42: Approach six 57
Figure 43: Clusters after approach six 58
Figure 44: Approach seven 58
Figure 45: Clusters after approach seven 59
Figure 46: Approach eight 59
Figure 47: Clusters after approach eight 60
Figure 48: Approach nine 60
Figure 49: Clusters after approach nine 61
Figure 50: Approach Ten 62
Figure 51: Clusters after approach ten 62
Figure 52: Approach eleven 63
Figure 53: Clusters after approach eleven 64
Figure 54: Elbow curve for the test data set 65
Figure 55: Dendogram for the test data set 66

List of Tables

Table 1: Different clusters form different methods.......... 33
Table 2: Data points which have negative values for silhouette score 40
Table 3: Data points which have negative values for silhouette score 44
Table 4: Data points which have negative values for silhouette score 46
Table 5: Data points which have negative values for silhouette score 48
Table 6: Selecting outliers 54
Table 7: Comparison of evaluation measurements for test data set 70
Table 8: Comparison of evaluation measurements 72
Table 9: Comparison of evaluation measurements 75
Table 10: Comparison of evaluation measurements 75
Table 11: Comparison of evaluation measurements 76
Table 12: Comparison of evaluation measurements 76
Table 13: Comparison of evaluation measurements 77
Table 14: Comparison of evaluation measurements 78
Table 15: Comparison of evaluation measurements 78

List of Graphs

Graph 1: Silhouette score for approach one 41
Graph 2: Silhouette score for approach two 43
Graph 3: Silhouette score for approach three 47

Abbreviations

DNA	Deoxyribonucleic acid
SNP	Single Nucleotide Polymorphism
DBSCAN	Density-based spatial clustering of applications with noise
CNA	Copy Number Alteration

Chapter 1 Introduction

1.1 Overview

Biology has become an important and a popular area among the sciences and fields in the twentieth and twenty first century as it helps to reveal some of the significant findings on human species and other species. It is more useful and predominant when knowledge of biology and knowledge of information science can be applied together. Bioinformatics can be identified as a hybrid field that brings these areas together. The contribution of bioinformatics advances made it possible to map the entire human genome and genomes of many organisms over a decade ago.

Bioinformatics is important to genetic research which involves the study of human DNA to find out what genes and environmental factors contribute to diseases, because genetic data has a context [21]. The large scale and very complex data generated in genetic research has to be analyzed to identify diseases and cures for diseases. Bioinformatics make it possible for researches to study these data and assist in researching.

Although bioinformatics plays a major role in genetic research, still there are gaps and inconsistencies those act as barriers to conduct genetic researches effectively. Hence, I will be addressing some of these issues in this research and analyze whether an information technology based solution will solve these issues.

1.2 Background

DNA sequencing and genome sequencing are considered as important area in genomic researches and has also become a useful subject to many other fields. Sequencing in genomic research refers to the process of determining the order of nucleotides or four bases – adenine, guanine, cytosine and thymine - of individual genes, large genetic regions, full chromosomes or entire genomes [14]. The genome sequence will represent a valuable shortcut, helping scientists find genes much more easily and quickly [15]. These sequences contains some clues on behaviors, characteristics and the abnormalities or diseases of a particular person [15].

With the introduction of Sanger sequencing technology by F. Sanger DNA sequencing became an important technique after 1977 [19]. Sanger sequencing is the traditional method for DNA sequencing and it was the most widely used sequencing method for approximately 25 years [16]. Since then the sequencing technology kept evolving and growing.

The Human Genome project which was started in 1990 was declared complete in 2003 by marking an important milestone in the genetic research history. The findings of the Human Genome Project led researchers to better understand the blueprint for building a person. Since then the demand for cheaper and faster sequencing methods has increased greatly. As a result of this high demand, Next Generation Sequencing (NGS) or Second-generation sequencing methods were developed [21].

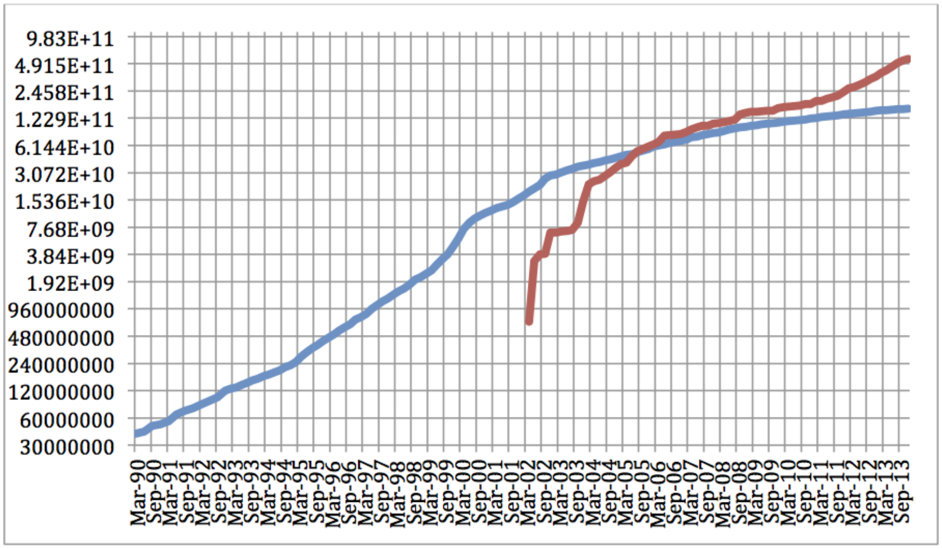


Figure 1: Increment of the total sequence in bp

The *y*-axis shows the total sequence in bp. (Blue = GenBank, red = whole genome shotgun [WGS] sequences.) Each line is double of the previous. The *x*-axis indicates time. Each line is 6 months after the previous. Source: http://www.ncbi.nlm.nih.gov/genbank/statistics.

Since genomic data exploration became an important area with the completion of the Human Genome project, the tools and techniques that were used in genomic context were improved. So the tools and techniques for data generation has increased the volume of data available to researchers and it is being increasing rapidly [21]. These data are being used in bioinformatics for collecting, storing and big data processing the genomes of organisms. Figure 1 shows how the number of sequences has increased over time.

From the above graph (figure 1), it is clear that the advent of new tools and technologies has significantly accelerated the pace of biological research & huge amounts of sequence data is becoming available for new researches.

These data are presented in different types of formats with several parameters in different data sources. Hence these are stored as high dimensional data from different biological and genetic studies. Thousands of DNA combinations have been identified as indicators of susceptibility to specific diseases. Some argue that you might go through life worrying needlessly about a disease that never appears. On the other hand, spotting those DNA variants and recognizing whether you are at risk can lead directly to early diagnosis and preventive strategies.

1.3 Problem Overview

The knowledge extracted from data is the key determinant of analyzing the functions of specific genes in genomic context. The outcome of these analysis will reveal some of the important aspects of human genome and its functions, such as human origins and disease risks as well as how they relate to environmental conditions, both past and present. Sometimes it is possible to predict the future as well when it comes to disease related genes such as cancer causing genes [13].

Genomic data explosion has been remaining as an important area in recent years with the advancement in several high-throughput biotechnologies such as RNA gene expression microarrays [16]. Researches which aim exploration of genomic data are mostly rely upon computational data and place the efforts to determine the entire DNA sequence of number of individuals in order to map and analyze individual genes [15],and particularly to discover how genes work in order to prevent or cure diseases.

However, with the rapid development of advanced technologies, the tools and techniques for data generation has increased the volume of data available to researchers, specifically in genomics [17]. These data are being used in bioinformatics for collecting, storing and big data processing the genomes of organisms to discover genome structures and other genomic parameters. These data sources can be categorized in to two categories as, Nucleotide sequence databases and Protein sequence databases [18]. Some of these data bases contain data from different studies separately and data on each study may present with several parameters [30]. Multiple measurements from multiple studies for each DNA sample can be extracted from these kind of data bases where high dimensional data set is required for analysis.

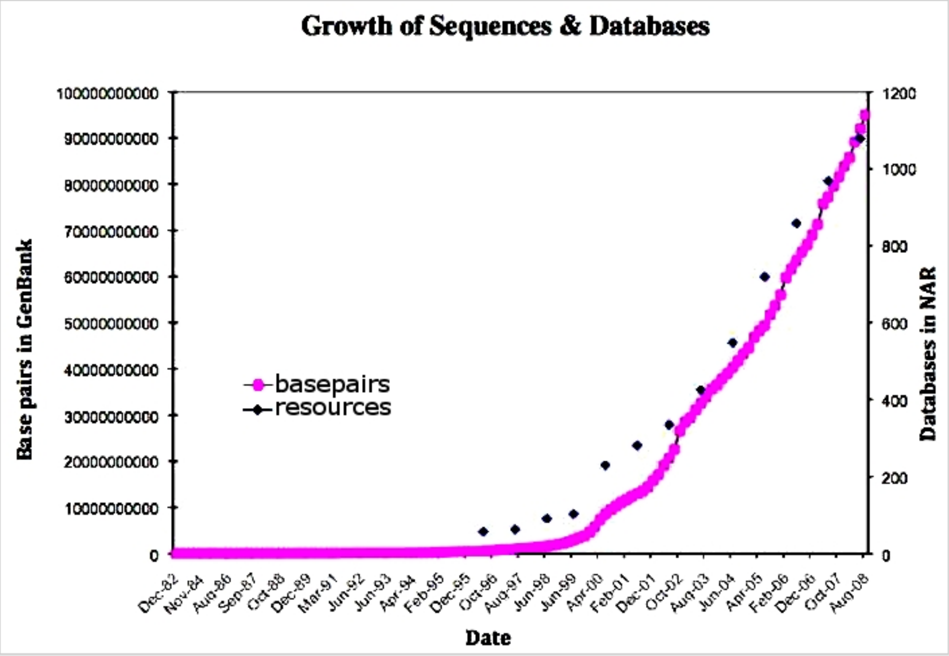


Figure 2: Increment of data sources

Above graph (Figure 2) shows how the data sources has increased in past few years and how the data that are available to the researches has been increased. The availability of large number of genomic data with multiple measurements makes 'meta' analysis possible, which can be used for study about the macro-level effects of DNA expression and mutations in the case of disease studies [32]. This in turn will potentially lead to the possibility of the early detection mortality rates and make predictions on expected life spans of terminally ill patients based on 'signatures' that could be derived from the DNA.

Make predictions on genomic data considering particular attribute or output has become a challenge as most of the genomic data is presented as unlabeled data in the data sources. To extract the information and predict on these information, it is important to cluster and label these data under different important aspect such as expected life spans. If the category of some gene data with known parameters can be decided using a particular method, it can be used to predict on those data in different aspect considering the data which already was in the same category [32].

Even though there are several researches which have one on genomic data in order to categorize them, most of those categorizing methods depend on the data set that they use. So there is a gap which needs to be filled, so that any kind of genomic data can be categorized and labeled irrespective of the characteristics of the data. In order to cluster and label these high dimensional genomic data it is require to explore a novel approach for analyzing and clustering, as a universal method which will be the main objective of this research [14].

But it contains many challenges when analyzing a massive genomic data set. The high dimensional nature is one of the main attribute of biological data that has been identified as one of the challenges in many genomic studies [4] [12]. Normalizing the collected high dimensional data will be needed a novel approach in order to be successfully applied a naïve clustering method. Hence this research may lead to find out a better approach to normalize the high dimensional data as one of the research deliverables.

With the aim of addressing the aforementioned problems, this research will carry out to introduce a new customized clustering method using data from multiple studies, which will reveal macro-level effects of DNA expression and mutations, and can be used on any of the genomic data by performing thorough examinations of the existing methods and approaches and the results of them followed up with a critical evaluation and validation.

1.4 Problem Statement

In bioinformatics community, acquiring of genomic sequence data is usually followed by the computational analysis in order to draw scientific insights and thereby use them in several domains such as in development of personalized medicine, make predictions on some diseases, predict life span of patients. Unfortunately, even though there are very advanced sequencing technologies available today, most of the data that are found using these technologies are unlabeled. It will be very useful if those can be categorize using several parameters to give a valuable label to each data set such as life span of a cancer patient. Therefore, the following two main problems are addressed in this research study.

Problem 01

Absence of a normalizing method to normalize high dimensional genomic data by eliminating redundancy and noisy data without a significant loss of information [4] [12]

Problem 02

Absence of a categorizing method to categorize normalized high dimensional genomic data into clusters and label them considering characteristics of the data in each cluster [2].

1.5 Significance of the Study

As mentioned in the problem statement, since there are no appropriate method to categorize any type of genomic data irrespective of the characteristics of the data, for different purposes we may have to go through different methods. Since the available data is present as unlabeled data some of the valuable information may be missed if they are not analyzed. But it is difficult when analyzing these data with the characteristic of high dimensional and large number of data.

Hence to overcome this problem, by using the data from different, multiple studies it is aimed to carry out a meta-analysis and present an effective, comprehensive clustering method which will be able to apply on any type of genomic data. The clusters which will be performed by the novel method will reveal the information on the macro-level effects of DNA expression and mutations. In the study of diseases it will be really valuable to study macro level effects since it will helpful to maximize the possibility of the early detection mortality rates and expected life spans of terminally ill patients based on 'signatures' that could be derived.

This will lead the scientist to decide the method of cure like medicines and therapy methods. Further these extracted information are important, so that screening can be done early in the case of terminal diseases such as cancer, even differentiated treatment regimens possibly developed for those in the different categories and make prediction on valuable aspects like life span of the patient.

1.6 Research Goals

Research goal defines the main aim of a research study. The final intension of this research study is to achieve the following research goals.

- Identify different categories of genomic data which will lead to make predictions on valuable aspects (Ex: Life span of a cancer patient) by presenting a novel clustering method.
- Introduce a data normalizing method for high dimensional genomic data by eliminating redundancy and noisy data without a significant loss of information.

1.7 Objectives

The goals of a research are attained via research objectives. Research objectives support the achievement of research goals.

The main aim of this research study is to find a novel clustering method to cluster the high dimensional genomic data to label them under valuable aspects.

In this research study, the above defined research goals are attained via the following main research objectives.

1. Identify the online biomedical sources from which information on high dimensional genomic data can be acquired.
2. Identify the parameters of high dimensional genomic data that can be used to categorize the data set under important labels
3. Identify a most efficient method to normalize high dimensional genomic data to combine the data from several studies.
4. Design and implement a method to categorize and label the genomic data
5. Validate the new categorizing method using existing categorized data from several researches

1.8 Research Scope

1.8.1 Only consider genomic data on Prostate cancer

Since it will be difficult to study all type of genomic data, it is planned to narrow down the research scope to only one specific type of genomic data which will be cancer related data. According to the studies conducted, during the year of 2017, 1,688,780 new cancer cases and 600,920 cancer deaths are estimated to take place in the United States [10]. Also, it was remarked that the rate of cancer incidence is 20% greater in men than women, and the rate of cancer death is 40% higher in men [10].

Hence it is aimed to use only the data on prostate cancer as it is the most common cancer type among male population, even though there are lots of research on breast cancers, comparatively researches on prostate cancers are low [2].

1.8.2 Use CBIO portal as the data source

As mention in the problem definition, there are large amount of genomic data which is freely available with an internet access [6]. It is aimed to use CBIO portal as the data source for this research as it is specially stored cancer specific multidimensional genomic data with an open access. The CBIO Portal facilitate rapid, intuitive, and high-quality access to molecular profiles and clinical attributes from large-scale cancer genomics projects and empowers researchers to translate these rich data sets into biologic insights and clinical applications [9].

1.8.3 Consider only three selected Features on genomic data

Data analysis will be only used selected features in genomic data. They are,

- Expression level
- Copy number alteration/variation
- Mutation

Genomic data with these features will be extracted from Data_CNA, Data_clinical and Data_fusion files which can be downloaded from CBIO portal. One data file may contain about 25000 of data sets with above features which will create a high dimensional data set.

Chapter 2 Literature Review

2.1 Overview

As new sequencing technologies promise a new era in the use DNA sequence, a large number of genes related to human have been identified along with disease-causing mutations. Today, different computational methods are available for recording, capturing, analyzing and distribution of this information which continues to grow exponentially in size and complexity.

In this chapter, we are presenting latest technologies and methods available for gene clustering and categorizing and their benefits as well as drawbacks.

2.2 Related work

There has been a quite number of gene clustering methods proposed and applied in the literature. Some of the clustering methods are known as traditional clustering methods, such as Hierarchical clustering, K-means[13], K-medoid, self-organizing maps (SOM). Some of the clustering methods such as Model-based clustering and tight clustering are considered as the methods which allow a noise set of genes [7]. But these methods perform the cluster as a more false positive outcome [7].

K-means algorithm, which is known as a traditional clustering method is a vastly used algorithm among above methods, as there are considerably high number of researches have conducted on K-means algorithm. Most of them have been conducted on genomic data and medical data [21]. There are some researches which have been carried out in order to accelerate the performance of K-means algorithm on large scale data in life science by analyzing a simple heuristic method. [19]

When it comes to gene clustering, selecting the most suitable clustering method from many available methods and selecting the corresponding parameters is a challenge. So in literature there are studies which compare and demonstrate the effectiveness of the clustering methods and their feasibility to gene clustering. These studies have concluded that a method works well in some datasets may perform poorly in other datasets as there are different data structure and characteristics [7].

Some researches have compared a selected clustering method with other clustering methods in order to identify the benefits and the drawbacks of that method, as well as to identify the most suitable data set and its characteristics to apply the selected clustering method. K-means algorithm has been analyzed in this manner with some selected clustering methods in the literature [18]. Discussing how various combinations of data mining classification algorithms are used on medical data for efficient classification of the data is another research which carried out under comparing the existing clustering methods [19]. There are some of the researches which have been conducted to present the major challenges and key issues in designing clustering methods, hence point out some of the emerging and useful research directions, considering semi – supervised clustering, ensemble clustering and simultaneous feature selection during data clustering and large scale data clustering [20].

There are number of studies which have been done on genomic data which can be identified in a vast range of gene types. Some have used the genomic data from human genome while others have been conducted on microorganisms [14]. Supervised clustering methods and unsupervised clustering methods such as hierarchical clustering method, K- Means, SOM have been investigated to model the relationships between gene expression data and gene functions automatically in microorganism [14].

Studying the behavior of cancer causing genes is common in study of genomic data. There are some of the studies which have used images of the scanned slides of breast cancer tumors. Log (base 2) ratios were used to flag the aberrant spots and slide regions [1] of those images. Then the hierarchical agglomerative clustering using the statistical package BRB-ARRAYTOOLS software was applied to these normalized log ratios [1]. Also both compact linkage and average linkage and both Euclidean and one minus Pearson correlation distance metrics were used for the analysis [1]. Natural subclasses of breast tumors were classified using unsupervised, hierarchical clustering approach considering ER status of the tumor as the end result of this research.

Cancer classification is another area where analyzing genomic data will be interested. Cancer classification will be important when identifying new cancer classes or when assigning tumors to known classes [2]. Classify the cancers based on gene expression monitoring for leukemia cancer and predict cancer classes independent from previous biological knowledge is one of the research comes in the literature. Before apply the classification method in order to find the correlation between the expressions patterns of the genes in their data set, a method called “neighborhood analysis” were developed [2]. Then the self-organizing map (SOM) technique was applied on the data set to classify the tumor classes.

Using deep learning is another trend in classification genomic data in the literature of classification human genomic data. This comes under the unsupervised classification methods. There are quite few methods have been proposed to detect cancer using gene expression data. This method has also mainly applied on gene expression data aiming the cancer detection and cancer type analysis. The main advantage of this method over other cancer detection approaches is the possibility of applying data from various types of cancer to automatically form features which help to enhance the detection and diagnosis of a specific one [12].

In most of the methods which used deep learning method, the focus was on how to learn features and reduce the dimensionality of the gene expression data. The majority of these methods use manually designed feature selectors to reduce the dimensionality of gene expression and select informative sets of genes. The potential problems with these feature selection methods are scalability and generality of features. But there are researches which were aimed to provide the potential to overcome problems of traditional approaches with feature dimensionality as well as very limited size data set.

Doing predictions is one of the main target of classification genomic data. These predictions can be on type of cancer, type of medicine, group of geographic population. A research has been conducted to successfully predict geographic population groups and is consistent against all the genotypic data set consisting of all the chromosomes (Figure 3). It shows that the inferred featured from the genotypic data with higher clustering and classification accuracy [21].

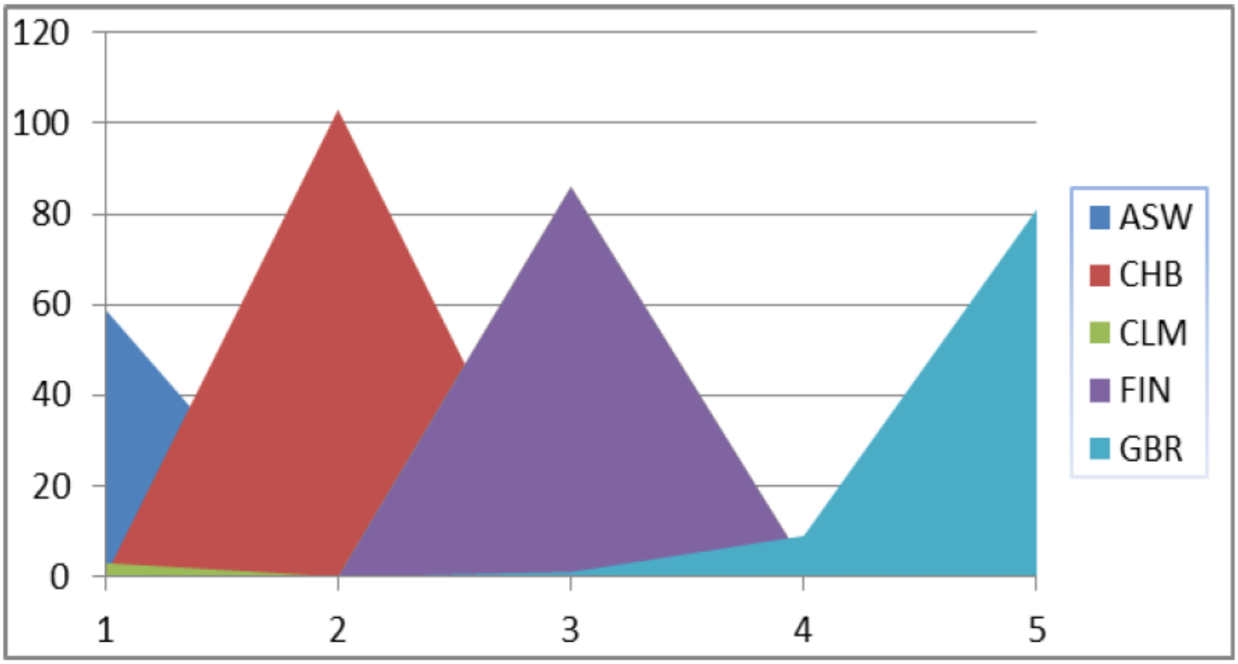


Figure 3: Population scale cluster of 5 groups from chr 22 (between actual and predicted)

Integrating the data from different types of cancers to automatically form features which help to enhance the detection and diagnosis of a specific cancer type is one of the researches which has addressed the nature of high dimensional data in the area of gene expression data [12]. Principal component analysis (PDA) has been used to reduce the dimensionality of the feature space. The approach used in this research consists of two parts. Feature learning phase and classifier learning phase. For the second part of the feature learning phase they have used an auto encoder neural network which is an unsupervised feature learning method [12].

Most of the researches in cancer classification, gene expression is one of the key features which is used for the classification under clinically relevant subgroups. Refining the results of these studies is another trend in the research field. Hierarchical clustering based on patterns of expression has been used on breast tumor genomic data in such researches and have concluded the idea that many of these breast tumor subtypes represent biologically distinct disease entities [8].

Chapter 3 Research Design & Methodology

3.1 Overview

The research design phase plays a major role in a research. It defines the structure that is followed in a research and thereby giving direction and systemizing the research. In this chapter, the design phases of the research methodology that we are adopting for our research is discussed. We have divided the research design and methodology into four main phases as shown in Figure 4. The first phase will be discussed in detail in this chapter itself whereas the other three phases will be discussed in the consecutive chapters in detail.

3.2 Analyze nature of data and resources

The focus of this research in biological scenario is to categorize genomic data on cancers in a convenient manner and thereby to predict on some of the important area which will be effect on patient's status such as life span. Therefore, it is crucial to understand the biological terms, data and resources that are needed to conduct the research. This phase focuses on analyzing and studying the nature of data being used and the sources that we will be using in the study.

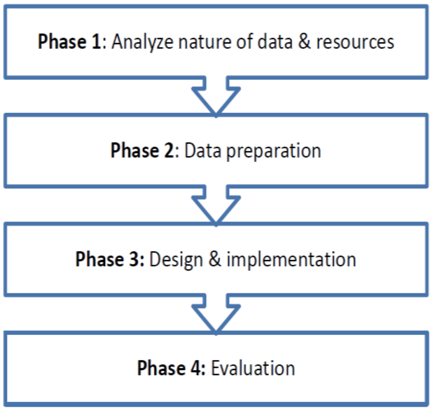

Phase 1: Analyze nature of data & resources
Phase 2: Data preparation
Phase 3: Design & implementation
Phase 4: Evaluation

Figure 4: Block diagram of research design phase

3.2.1. Biological background

What is DNA

Our bodies have around 210 different types of cells. Each cell does a different job to help our body to function. There are blood cells, bone cells, and cells that make our muscles. Cells get their instructions on what to do from DNA. DNA acts sort of like a computer program. The cell is the computer or the hardware and the DNA is the program or code.

What is gene?

The DNA Code

The DNA code is held by the different letters of the nucleotides. As the cell "reads" the instructions on the DNA the different letters represent instructions. Every three letters makes up a word called a codon.

ATC TGA GGA AAT GAC CAG

Genes

Within each string of DNA are sets of instructions called genes. A gene tells a cell how to make a specific protein. Proteins are used by the cell to perform certain functions, to grow, and to survive.

CNA (Copy number alteration)

Copy number alteration is one of the main feature in DNA that is going to be used in this research. This copy number alteration can be happen in three ways. They are,

Insertion – In insertion some different nucleotide will be added to the DNA sequence and the DNA sequence will be altered because of this newly added nucleotide (Figure 5).

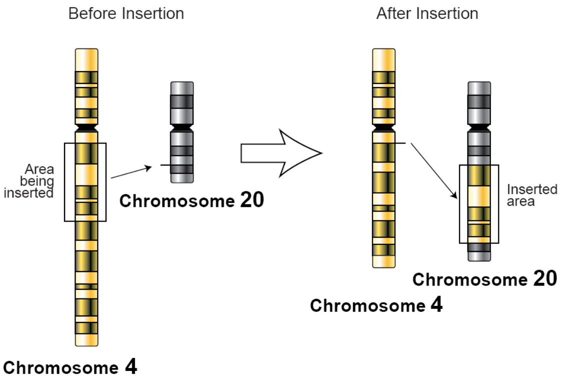


Figure 5: Insertion

Deletion – In deletion one or more nucleotide will be deleted from the DNA sequence so that the DNA sequence will get altered (Figure 6).

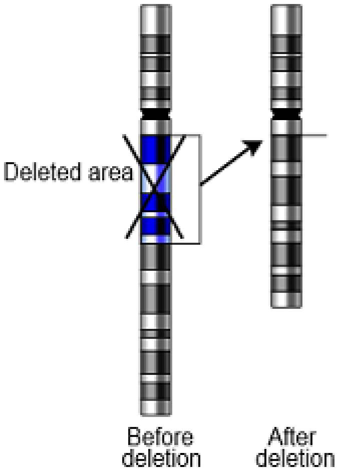


Figure 6: Deletion

Duplication – In duplication particular DNA part will get duplicate so that the DNA sequence will be altered (Figure 7).

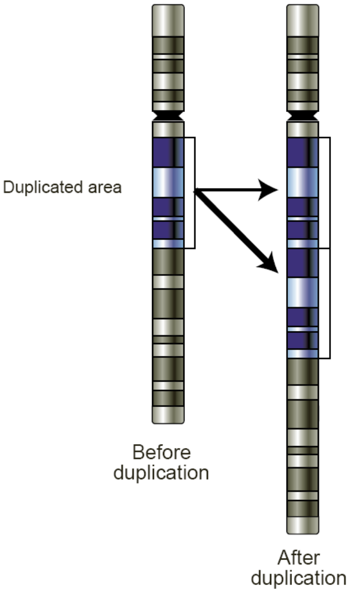


Figure 7: Duplication

Expression Levels

The process by which the heritable information in a gene, the sequence of DNA base pairs, is made into a functional gene product, such as protein or RNA. Interpret this using z-score

Mutations

A mutation is a mistake or a change in a living thing's DNA. DNA, or deoxyribonucleic acid, is a chain of chemical units found in each cell of a living thing. The chemical units are arranged in a particular sequence, or order. This sequence forms a kind of code, called a genetic code, which tells cells what to do. If the chain gets out of order, breaks, or changes in some other way, a mutation

3.2.2. Data representation in selected data source

In this research we are using cBioPortal as the main data resource which is an open-access, open-source resource. It allows the users to access multidimensional cancer genomic data sets as it stores the data on one type of cancer from several studies. It allows the users rapid, intuitive, and high-quality access to molecular profiles and clinical attributes from large-scale cancer genomics projects so that the researches can reach the relevant data easily. So that researches are encouraged to analyze these data and apply the investigated, observed outcomes into biologic insights and clinical applications.

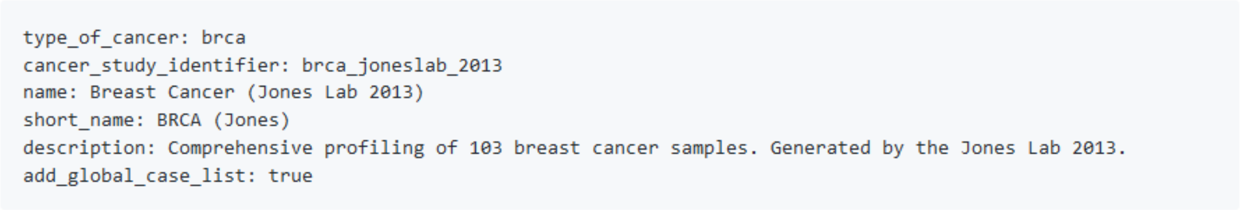
This data source is currently using hg19/GRCh37 as the version of human reference genome. The data sets are categorized under each cancer types such as Adenoid Cystic Carcinoma, Bladder Cancer, Breast cancer, Prostrate cancer etc…so that the researches can easily access the required cancer type. All these data are presented using different types of file formats for different types of data.

Analyze file types and the meaning of presented data

One of the main objectives of the research is to identify a data normalizing method for the data set which is going to be use for the research. The data set is also from biological back ground and it contains thousands of data. Hence it is important to have good understanding on the presented data and how they have presented in the data source.

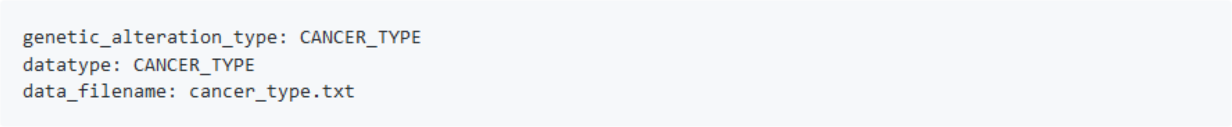
cBioportal contains data sets from different studies for different cancer types. One type of cancer contains data from different studies and one study contains different types of data, such as DNA data, RNA data, Sequence data and clinical data. The arrangement is to download one set of data of one study under preferred cancer type. These downloaded data is stored in different file formats.

These files are mainly in two formats one is meta files and other one is data files. There are three main types of meta files, cancer study, cancer type and clinical data. Cancer study data file contains meta data about cancer study such as type of cancer, cancer study identifier, name, description etc… (Figure 8). Cancer type data file contains some basic information on cancer type (Figure 9). In clinical data file is used to capture both clinical attributes and the mapping between patient and sample ids (Figure 10). The software supports multiple samples per patient.



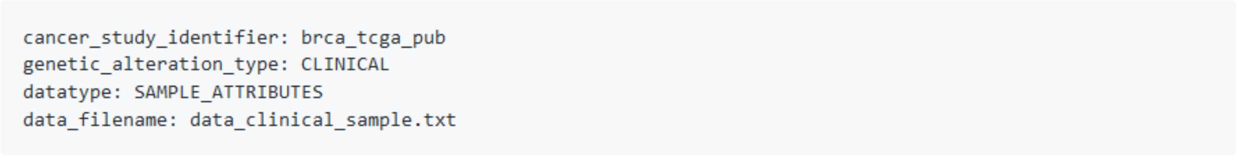
```
type_of_cancer: brca
cancer_study_identifier: brca_joneslab_2013
name: Breast Cancer (Jones Lab 2013)
short_name: BRCA (Jones)
description: Comprehensive profiling of 103 breast cancer samples. Generated by the Jones Lab 2013.
add_global_case_list: true
```

Figure 8: Meta file – Cancer study



```
genetic_alteration_type: CANCER_TYPE
datatype: CANCER_TYPE
data_filename: cancer_type.txt
```

Figure 9: Meta file – Cancer Type



```
cancer_study_identifier: brca_tcga_pub
genetic_alteration_type: CLINICAL
datatype: SAMPLE_ATTRIBUTES
data_filename: data_clinical_sample.txt
```

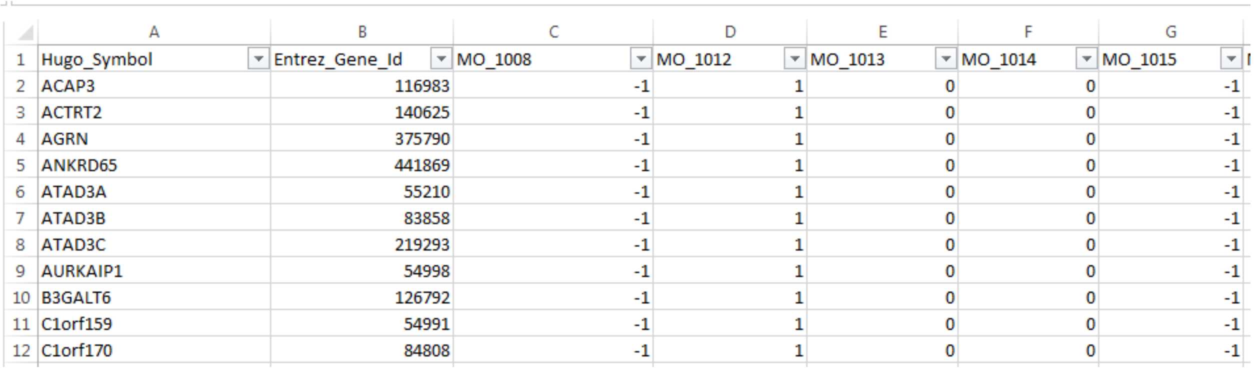
Figure 10: Meta file – Clinical Data

As shown in above examples, these files only contain some of the basic information on the study and the clinical data. But data in data files are more interested for researches. There are three main areas which have being used for several analysis and consider as valuable data sets which can be used for different research works.

Copy number alteration, Mutations and Expression are the three main features that are focused on the data files. These data files may contains about 40000 rows and 300 columns in one file.

Copy number alteration is one of the predominant feature researches use for their experiments. These data is presented in one of the data files in cBioportal as data_CNA file. CNA – Copy number alterations and copy number variations can be considered as the same meaning but the context that they are being using is different. Copy number alterations/aberrations (CNAs) are changes in copy number that have arisen in somatic tissue like in a tumor, copy number variations (CNVs) originated from changes in copy number in germline cells (and are thus in all cells of the organism). In the data_CNA file, they have used a standard convention to store these CNAs corresponding to each cancer gene. In the file the rows represent the cancer genes and the columns represent the positions of the particular gene. Each cell has a value which represent the type of the CAN in that position (Figure 11) [9]. For this representation they have used -2,-1, 0, 1, 2 these indicate the copy-number level per gene as below,

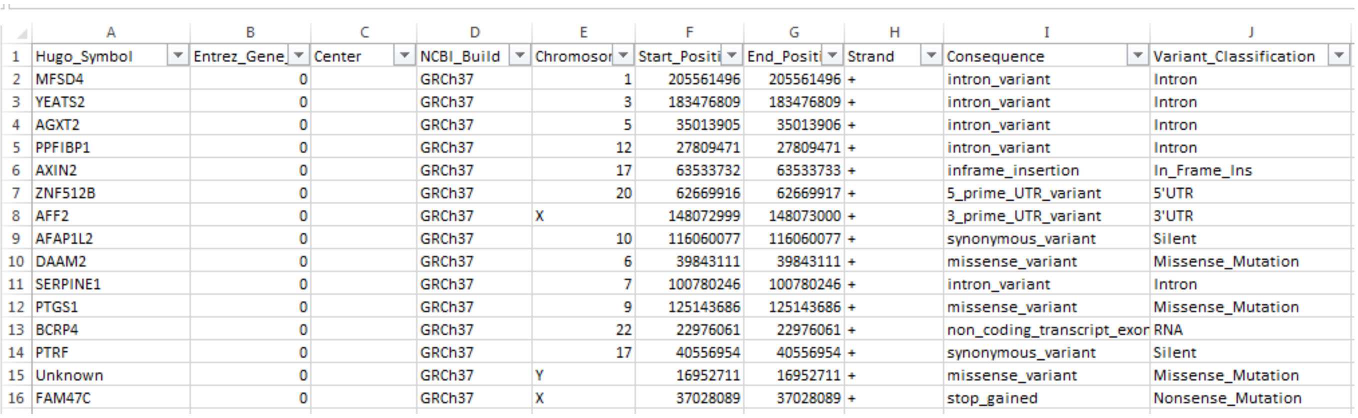
- -2 or Deep Deletion indicates a deep loss, possibly a homozygous deletion
- -1 or Shallow Deletion indicates a shallow loss, possible a heterozygous deletion
- 0 is diploid
- 1 or Gain indicates a low-level gain (a few additional copies copies, often broad)
- 2 or Amplification indicate a high-level amplification (more copies, often focal)



	A	B	C	D	E	F	G
1	Hugo_Symbol	Entrez_Gene_Id	MO_1008	MO_1012	MO_1013	MO_1014	MO_1015
2	ACAP3	116983	-1	1	0	0	-1
3	ACTRT2	140625	-1	1	0	0	-1
4	AGRN	375790	-1	1	0	0	-1
5	ANKRD65	441869	-1	1	0	0	-1
6	ATAD3A	55210	-1	1	0	0	-1
7	ATAD3B	83858	-1	1	0	0	-1
8	ATAD3C	219293	-1	1	0	0	-1
9	AURKAIP1	54998	-1	1	0	0	-1
10	B3GALT6	126792	-1	1	0	0	-1
11	C1orf159	54991	-1	1	0	0	-1
12	C1orf170	84808	-1	1	0	0	-1

Figure 11: Data file – Copy number alteration

Mutations are another predominant feature that is popular among researches. data_mutation is the file all the mutations are stored in a particular study in cBioportal (Figure 12). Unless the data_CNA file this data_mutation file contains different types of data such as consequence, variant type and variant classification. Compared to data_CNA file it is difficult to normalize the data mutation file as it contains categorical data and numerical data.



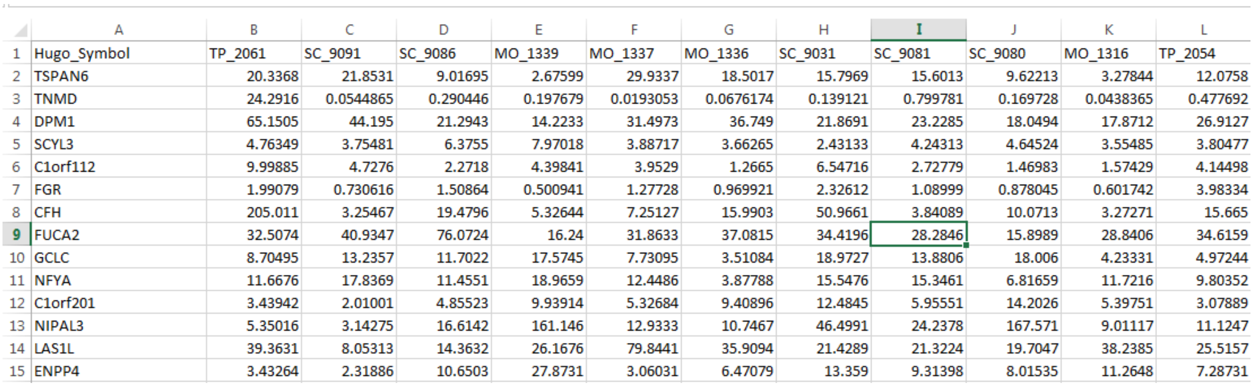
	A	B	C	D	E	F	G	H	I	J
1	Hugo_Symbol	Entrez_Gene_	Center	NCBI_Build	Chromosor	Start_Positi	End_Positi	Strand	Consequence	Variant_Classification
2	MFSD4	0		GRCh37	1	205561496	205561496	+	intron_variant	Intron
3	YEATS2	0		GRCh37	3	183476809	183476809	+	intron_variant	Intron
4	AGXT2	0		GRCh37	5	35013905	35013906	+	intron_variant	Intron
5	PPFIBP1	0		GRCh37	12	27809471	27809471	+	intron_variant	Intron
6	AXIN2	0		GRCh37	17	63533732	63533733	+	inframe_insertion	In_Frame_Ins
7	ZNF512B	0		GRCh37	20	62669916	62669917	+	5_prime_UTR_variant	5'UTR
8	AFF2	0		GRCh37	X	148072999	148073000	+	3_prime_UTR_variant	3'UTR
9	AFAP1L2	0		GRCh37	10	116060077	116060077	+	synonymous_variant	Silent
10	DAAM2	0		GRCh37	6	39843111	39843111	+	missense_variant	Missense_Mutation
11	SERPINE1	0		GRCh37	7	100780246	100780246	+	intron_variant	Intron
12	PTGS1	0		GRCh37	9	125143686	125143686	+	missense_variant	Missense_Mutation
13	BCRP4	0		GRCh37	22	22976061	22976061	+	non_coding_transcript_exor	RNA
14	PTRF	0		GRCh37	17	40556954	40556954	+	synonymous_variant	Silent
15	Unknown	0		GRCh37	Y	16952711	16952711	+	missense_variant	Missense_Mutation
16	FAM47C	0		GRCh37	X	37028089	37028089	+	stop_gained	Nonsense_Mutation

Figure 12: Data file – Mutations

In cBioportal data repository, the expression data will be stored in a text file which will be comes under data file category, named "data_RNA_Seq_expression_median". mRNA expression data (Figure 13) will be captured in this data file. Relative expression of an individual gene and tumor to the gene's expression distribution in a reference population are computed for mRNA and microRNA expression data. These values indicates the number of standard deviations away from the mean of expression in the reference population (Z-score). When determine whether a gene is up-regulated or down regulated compared to the normal sample this value is used. Positive values are considered as up-regulated and the negative values are considered as down – regulated. Usually up-regulated means more highly expressed compared to the reference whereas down-regulated means expressed lower compared to the reference.

In the data file of expression it contains gene name in the rows, the gene positions in the columns and the corresponding mRNA z-score in the each cell.

For all above mentioned data files, there are no any duplicate rows as the cBioPortal assumes that gene samples or the patients under the same ID are actually same. This feature is important for those whom interested on doing cross – cancer queries where each sample should only be counted once.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Hugo_Symbol	TP_2061	SC_9091	SC_9086	MO_1339	MO_1337	MO_1336	SC_9031	SC_9081	SC_9080	MO_1316	TP_2054
2	TSPAN6	20.3368	21.8531	9.01695	2.67599	29.9337	18.5017	15.7969	15.6013	9.62213	3.27844	12.0758
3	TNMD	24.2916	0.0544865	0.290446	0.197679	0.0193053	0.0676174	0.139121	0.799781	0.169728	0.0438365	0.477692
4	DPM1	65.1505	44.195	21.2943	14.2233	31.4973	36.749	21.8691	23.2285	18.0494	17.8712	26.9127
5	SCYL3	4.76349	3.75481	6.3755	7.97018	3.88717	3.66265	2.43133	4.24313	4.64524	3.55485	3.80477
6	C1orf112	9.99885	4.7276	2.2718	4.39841	3.9529	1.2665	6.54716	2.72779	1.46983	1.57429	4.14498
7	FGR	1.99079	0.730616	1.50864	0.500941	1.27728	0.969921	2.32612	1.08999	0.878045	0.601742	3.98334
8	CFH	205.011	3.25467	19.4796	5.32644	7.25127	15.9903	50.9661	3.84089	10.0713	3.27271	15.665
9	FUCA2	32.5074	40.9347	76.0724	16.24	31.8633	37.0815	34.4196	28.2846	15.8989	28.8406	34.6159
10	GCLC	8.70495	13.2357	11.7022	17.5745	7.73095	3.51084	18.9727	13.8806	18.006	4.23331	4.97244
11	NFYA	11.6676	17.8369	11.4551	18.9659	12.4486	3.87788	15.5476	15.3461	6.81659	11.7216	9.80352
12	C1orf201	3.43942	2.01001	4.85523	9.93914	5.32684	9.40896	12.4845	5.95551	14.2026	5.39751	3.07889
13	NIPAL3	5.35016	3.14275	16.6142	161.146	12.9333	10.7467	46.4991	24.2378	167.571	9.01117	11.1247
14	LAS1L	39.3631	8.05313	14.3632	26.1676	79.8441	35.9094	21.4289	21.3224	19.7047	38.2385	25.5157
15	ENPP4	3.43264	2.31886	10.6503	27.8731	3.06031	6.47079	13.359	9.31398	8.01535	11.2648	7.28731

Figure 13: Data file – Expressions

Identifying data types and categorizing

In order to categorize a data set according to their similarities and dissimilarities, it is important to have a clear view on the data set as well as the types of data since the process which is going to be used will depend on the structure of the data set and the types of the data. These collected data mainly can be categorized into two categories as numerical data and categorical data.

Numerical data

This has the meaning of measurement of something, such as height, weight, number of shares. Statisticians also call numerical data quantitative data. Numerical data can be further broken into two types: discrete and continuous.

Discrete data represent items that can be counted; they take on possible values that can be listed out. The list of possible values may be fixed (-2, -1, 0, 1, 2); or it may go from 0, 1, 2, on to infinity. In the data that is focused for the research, some of the data can be categorized under this discrete data which is numeric data. The first data set which comes under data files, CNA data is one example, CNA data set only contains the values of -2, -1, 0, 1, and 2.

Continues data is another data type which comes under numerical data, their possible values cannot be counted and can only be described using intervals on the real number line. In the data set mRNA Z- score which comes under expression data, can be identified as continues data but it has both negative and positive values.

Categorical data

Categorical data represent characteristics such as a person's gender, marital status, and hometown. Categorical data can take on numerical values (such as "1" indicating male and "2" indicating female), but those numbers don't have mathematical meaning.

Most of the values which is in mutation file are comes under this data type. Variant type (DEL, INS, SNP), Variant classification (RNA, Silent, Intron..), consequence are some of the data fields comes under categorical data.

Chapter 4 Data Preparation

4.1 Overview

The previous chapter provided an outline of the design and methodology of the research study. The purpose of this chapter is to discuss the data preparation phase of the research in detail. This chapter discusses how data was collected and how the data was prepared according to the data types.

4.2 Data Collection and preparation

Since this research is based on genomic data which needs high technology and expertise knowledge to be collected data collection is focused on collecting secondary data which is already in data resources and data repositories. cBioportal is the selected data source as the convenient data repository as it is focused on vast range of cancer genomic data and it gives the free access to interactive exploration of multidimensional cancer genomics data sets.

In the cBioportal the data set of different studies have listed under corresponding cancer type. One cancer type may contains data sets from about five, six studies. These data can be easily downloaded to the local environment. For this research three data sets from three studies under prostate cancer are downloaded. Each data set has meta data files as well as data files. The meta files only contains the meta data such as information on the study and the information on the cancer. The data files which are downloaded are interested on this research as it contains inside data on the cancer (Please refer 3.2.1.1. for more information on data). Data on copy number alteration (CNA), Mutations and Expression take in to consideration throughout the research.

4.2.1 Data preparation

Since one of the main objective of the research is normalizing the collected data in order to use the data set for new approach to a clustering method data preparation, preprocessing and data analysis are considered as important steps and may have several iteration according to the approach of the clustering method.

The three main files which are going to be used for the research should be combined together in order to normalize one data set. But these three data sets may contain different gene set from one another, also different sample sets from one another. To combine all the data in to one set, the common genes of all the three data set and the common sample sets in three data sets.

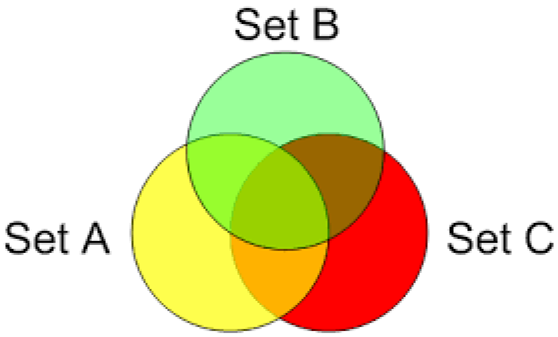


Figure 14: Representation of data set combination

Intersection of gene sets of all three data sets and intersection of sample sets of all three data sets as showed in figure 14 will be taken in to consideration of this research.

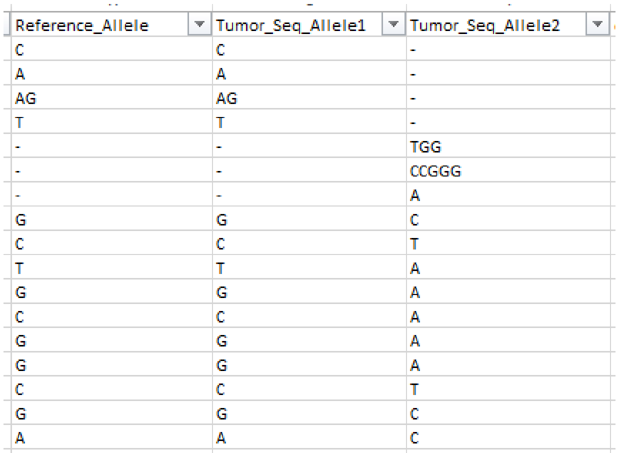
Both numerical and categorical data are included in the data set as explained in chapter 3. Converting categorical data in to numerical data is one of the main challenge in this data set as the data on mutation mostly contains categorical data and one feature contains considerable number of parameters. Some of the convenient methods which were in the literature were carried out in the process of converting the categorical data to numerical data.

As discussed above, to compare two entries of data and find the similarities and the dissimilarities in order to categorize the data categorical data need to be converted in to comparable format. Thus, we convert them into numerical variables. Below are the some of the methods which were used to convert a categorical (string) input to numerical nature.

All the categorical data comes under the mutation data set. Consequence, Variant classification, variant type, Tumor_Seq_Allele1, Tumor_Seq_Allele2 are some of the identified categorical data which will be interested in clustering process. Below are some of the used methods for converting categorical data in to numerical data.

Label Encoder

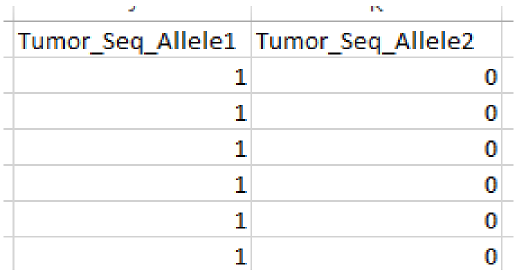
This method is used to transform non-numerical labels to numerical labels (or nominal categorical variables). This method was applied on Tumor_Seq_Allele1, Tumor_Seq_Allele2. In the mutation data set Tumor_Seq_Allele1 and Tumor_Seq_Allele2 are compared with the Reference_Allele and store whether the allele has copied correctly. (Figure 15)



Reference_Allele	Tumor_Seq_Allele1	Tumor_Seq_Allele2
C	C	-
A	A	-
AG	AG	-
T	T	-
-	-	TGG
-	-	CCGGG
-	-	A
G	G	C
C	C	T
T	T	A
G	G	A
C	C	A
G	G	A
G	G	A
C	C	T
G	G	C
A	A	C

Figure 15: Before application of label encoder

In the mutation data set reference allele was removed and Tumor_Seq_Allele1, Tumor_Seq_Allele2 allele only present whether it has copied correctly or not. (Figure 16) If the Tumor_Seq_Allele has not changed, it will be represented using '1' if the Tumor_Seq_Allele has not changed it will be represented by '0'.

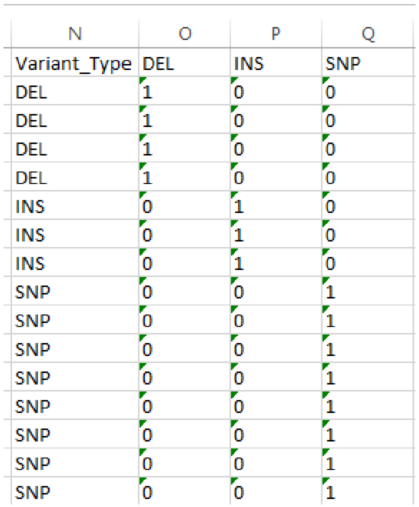


Tumor_Seq_Allele1	Tumor_Seq_Allele2
1	0
1	0
1	0
1	0
1	0
1	0

Figure 16: Application of label encoder

Dummy Coding

Another method used for converting categorical data in to numerical data is dummy coding. Variant type was converted to numerical variable using this dummy coding method. Variant type feature has three main values as SNP. DEL and INS. Three different dummy variables were created for to represent those values. Presence of a value is represent by 1 and absence is represented by 0. For every value present, one dummy variable will be created. Look at the representation below to convert a categorical variable using dummy variable. (Figure 17)



N	O	P	Q
Variant_Type	DEL	INS	SNP
DEL	1	0	0
DEL	1	0	0
DEL	1	0	0
DEL	1	0	0
INS	0	1	0
INS	0	1	0
INS	0	1	0
SNP	0	0	1
SNP	0	0	1
SNP	0	0	1
SNP	0	0	1
SNP	0	0	1
SNP	0	0	1
SNP	0	0	1
SNP	0	0	1

Figure 17: Application of dummy variables

There are some of the drawbacks with this normalized data set. Some of the features were eliminated as those features have considerable number of values hence it is difficult to normalize it using dummy variables. Replace those values with a number is also not worked as levels can't be defined for those values. Also there is a known challenge with nominal categorical variable it may decrease performance of a model. As the value '1' in CNA data and value '1' in dummy variables are represent two different meanings.

Hence for the first phase of the research, to apply different clustering methods and compare them in orders to get an approach to a new clustering method we decided to use only one set of data. After get to a conclusion on novel approach of the clustering method, normalized data set will be used and the validity of the data set is also can be measured by analyzing the result we will get after applying the clustering method.

Chapter 5 Design & Implementation

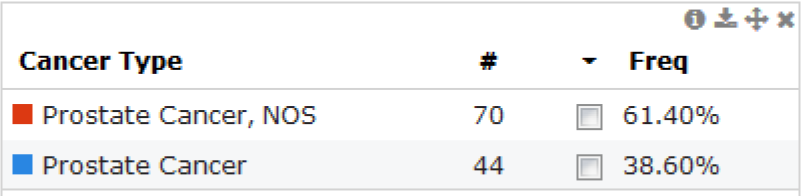
5.1 Overview

This chapter illustrates the process followed in order to construct a new approach for a clustering method by analyzing collected data and analyzing known clustering methods by applying them on the prepared data set. Initially, the focus is given to categorize data according to different measurements under medical aspect so that clear view on the data set and the high level idea on nature of the clusters can be explained. Then apply selected clustering methods for the data set and compare the outcomes was carried out.

5.2 Analyze the data set

Data set was analyzed according to the features before approach with the clustering methods as it gives an over view of the arrangement of the data set. The data set was analyzed and visualized in different aspects.

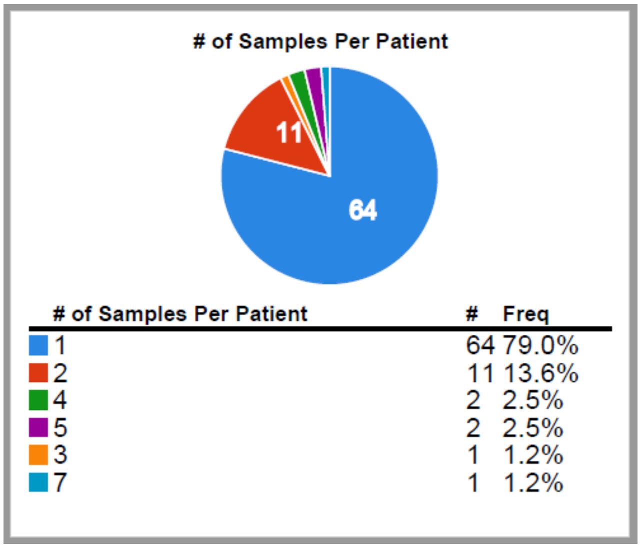
In one data set that is focused for the 1st phase of the clustering contains 114 samples of 81 patients. In the analysis we could identify that there are two categories according to the cancer type. They are Prostate cancer NOS and Prostate Cancer. The distribution among those two cancer types is as shown in figure18. According to this feature the data set will be clustered in to two.



Cancer Type	#	▾	Freq
🟥 Prostate Cancer, NOS	70	□	61.40%
🟦 Prostate Cancer	44	□	38.60%

Figure 18: Cancer Type Classification

The selected data set contains 114 samples from 81 patients. Several samples may contain from one patient. Figure 19 shows how the sample distribution has happened across patients.



# of Samples Per Patient	#	Freq
🟦 1	64	79.0%
🟥 2	11	13.6%
🟩 4	2	2.5%
🟪 5	2	2.5%
🟧 3	1	1.2%
🟦 7	1	1.2%

Figure 19: Sample Classification

All the genes in the data set contains some kind of mutation in it. These mutations play a major role in having the particular cancer or not. Analysis on this aspect is also will be more important when understanding the clusters that will be formed in the next phase of the research.

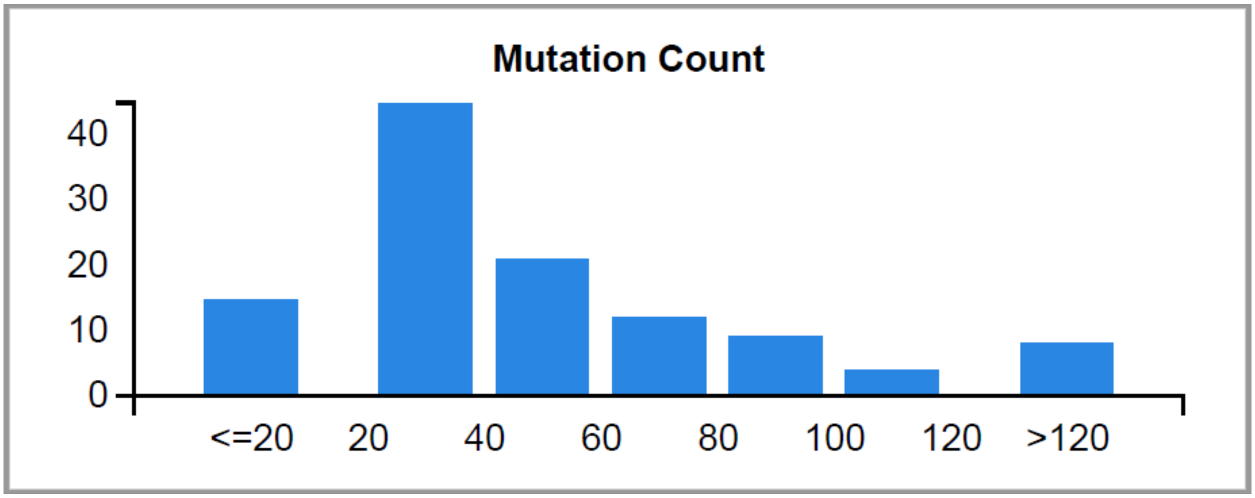


Figure 20: Mutation count

According to the mutation count of the each sample the distribution is as shown in the figure 20. Most of the samples (about 45 samples) contains about 20-40 mutations in one sample. The amount of mutation that the least number of samples contain is 100 – 120. Depending on the number of mutation we can define seven clusters as shown in the graph.

Copy number alteration and mutations are important aspects which are considered throughout this research. It was interested checking whether those two features have any connection in this sample set. Number of mutations in each sample and the fraction of copy number altered genome were considered as the variables and check the connection by plotting a graph between those two axes. Most of the samples (About 100) were led in an area which is parallel to the x- axis (Fraction of copy number altered genome). It implies most of the sample sets have number mutations in a particular range of values and there are small number of outliers as shown in figure 21.

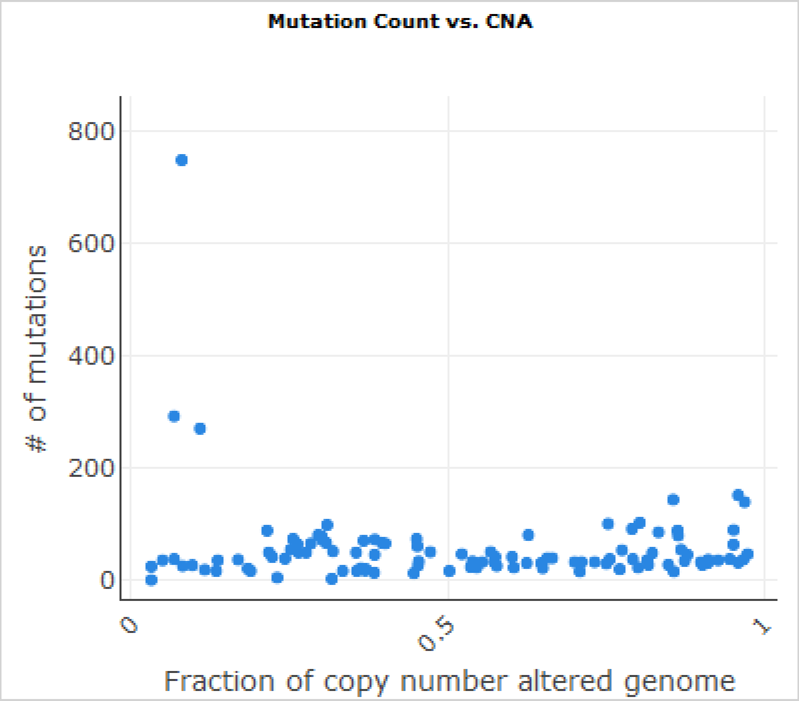
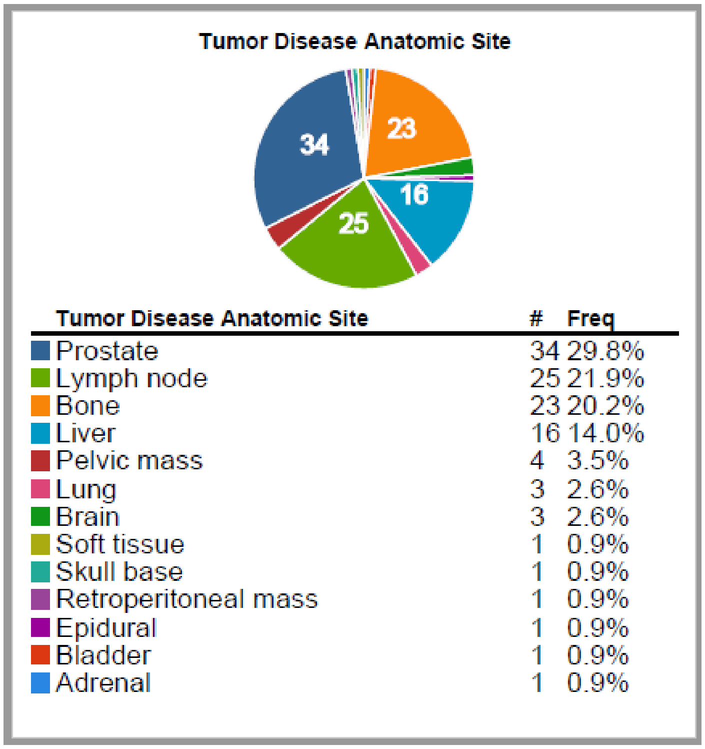


Figure 21: Mutation count vs CNA

Considering the area, the identified tumor is again the data set can be categorized. Some of the features of the genes differ according to these areas when clustering this data set considering all the similarities and the dissimilarities this fact may also may considered.



Tumor Disease Anatomic Site	#	Freq
🟦 Prostate	34	29.8%
🟩 Lymph node	25	21.9%
🟧 Bone	23	20.2%
🟦 Liver	16	14.0%
🟥 Pelvic mass	4	3.5%
🟥 Lung	3	2.6%
🟩 Brain	3	2.6%
🟨 Soft tissue	1	0.9%
🟩 Skull base	1	0.9%
🟪 Retroperitoneal mass	1	0.9%
🟪 Epidural	1	0.9%
🟥 Bladder	1	0.9%
🟦 Adrenal	1	0.9%

Figure 22: Tumor Disease anatomic site

Using all the three sets, copy number alteration, mutations and expression level analysis was carried out where we could find out the different levels of mutations and the samples which those mutations are with. According to the figure 22. In the graph the amplifications are the most prominent mutation type then the deletions. It is clear that there are different types of mutations and some of the sample sets contains amplifications in them and other samples don’t include them.

5.3 Application of selected clustering methods

If different clustering methods are applied to a same data set, different cluster sets will be created. There is no any rule or procedure to verify which cluster set is the most appropriate when considering the features of those data. For example rows below table (Table 1) present the clustering methods and column of the table shows the clusters which were created by the corresponding clustering methods. When comparing each set of clusters of each clustering methods some similarities can be found in some clusters. For example A, B clusters which were formed by P clustering method and the Q clustering method shows some similarity same as C cluster which was formed by Q, R, and S clustering methods shows similarity. If we can find those similar clusters which we identified exploring through the clustering methods using one clustering method or using some particular steps that would be really helpful in data analysis. Exploring such method to identify the most appropriate cluster set or to perform most suitable cluster set by introducing some sequential steps is the main target of this research.

	A	B	C	D
P	🟦	🟦		
Q	🟦	🟦	🟨	
R			🟨	🟩
S			🟨	🟩
T				🟩

Table 1: Different clusters form different methods

With the idea of the similarities of samples which explained in section 5.2 as the next step some known clustering methods which were selected from the literature were applied to the data set. The selected clustering methods are K-means clustering method, hierarchical clustering method and DBSCAN method.

As the first phase of this process K- means clustering method and hierarchical clustering method were applied to the data set and compared the results gained from both clustering methods using cluster method validations.

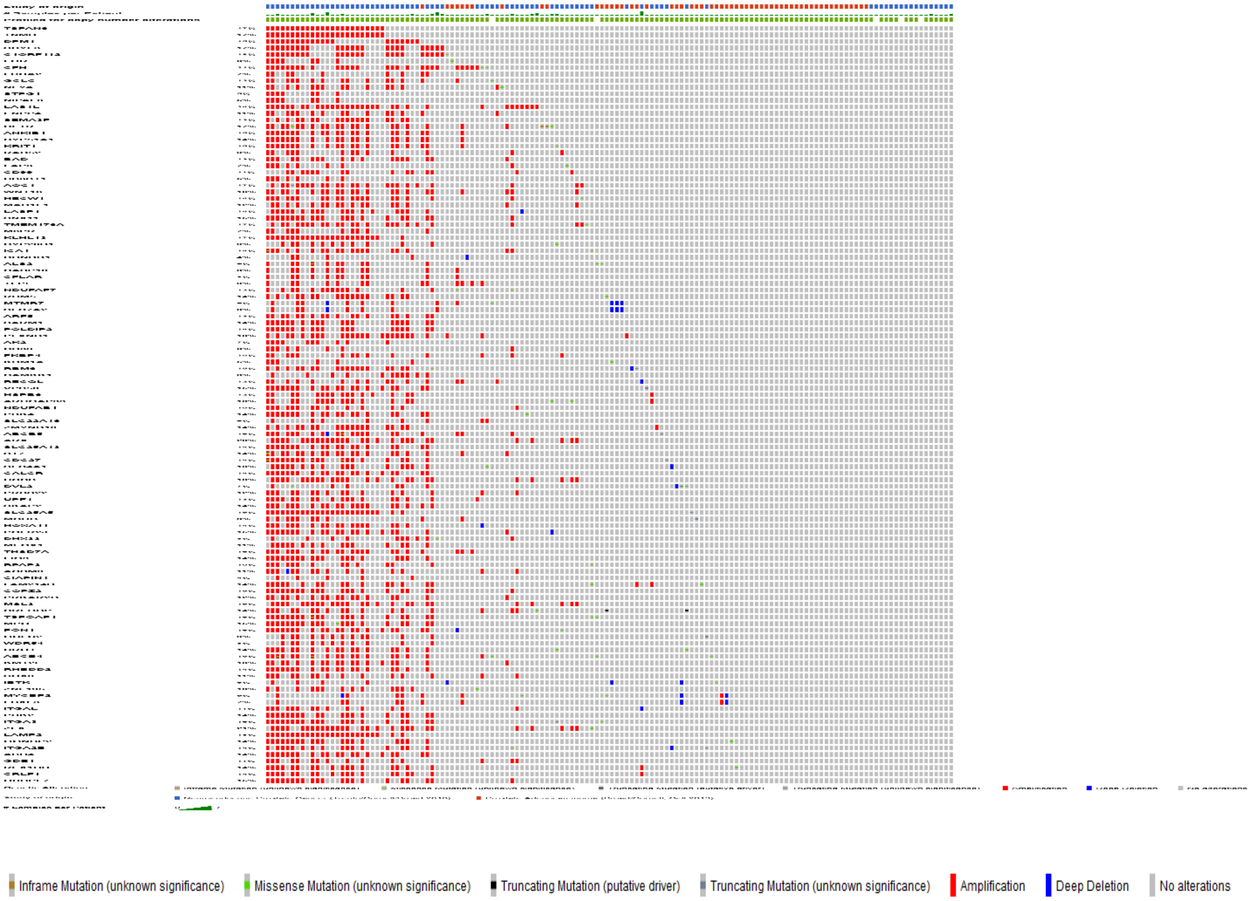


Figure 23: Mutation types classification

5.3.1. K- Means Clustering method

As per the literature on clustering algorithms, K-means clustering is one of the simplest and popular unsupervised machine learning algorithms [22]. These unsupervised clustering algorithms only use input data and form clusters without referring to known, or labelled, outcomes (Figure 24).

The algorithm works as follows:

1. First k points, called means, will be initialized randomly.
2. Each item will be categorized to its closest mean and the mean's coordinates will be, which are the averages of the items categorized in that mean so far.

3. Then the above two steps will be repeated for a given number of iterations and at the end, we have our clusters.

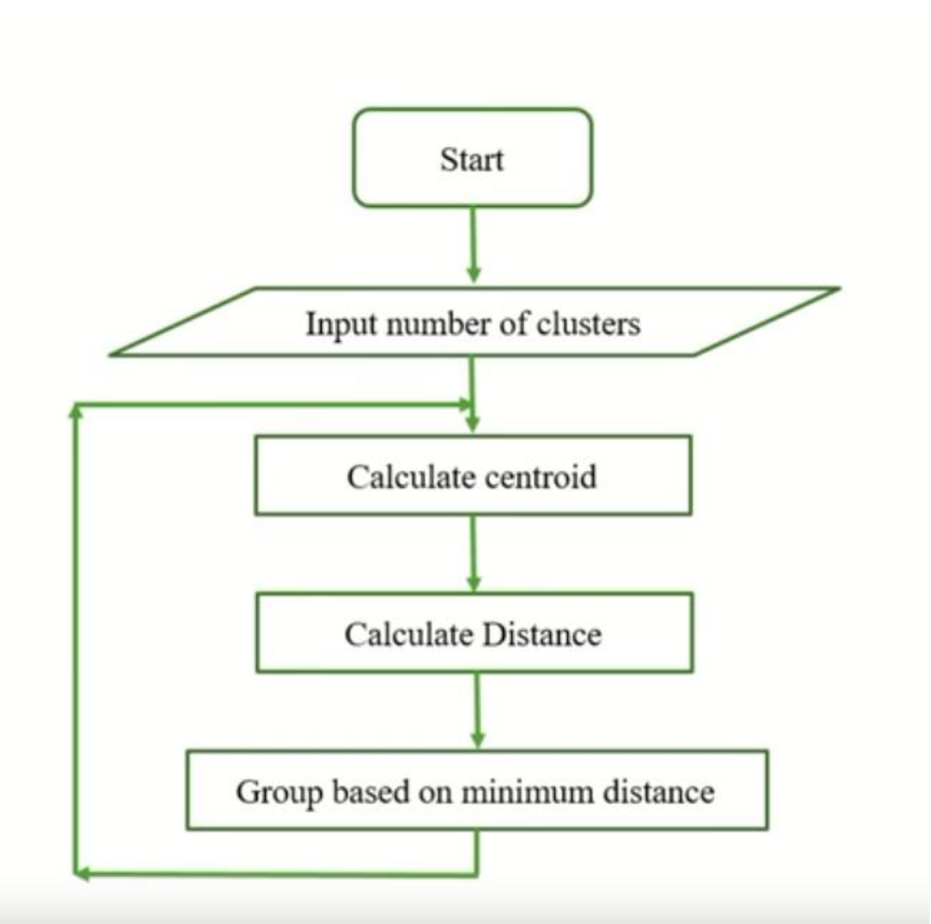


Figure 24: Steps of K-means clustering

Here target number of cluster will be defined as k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster. Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares [37].

In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

In partitioning clustering, where K means clustering include defining the number of cluster that we need our data set to be clustered should be given as an input to the algorithm. So the K- means clustering requires the user to specify the number of clusters k to be generated [37].

Even though the user can define any number as the number of clusters, there is an optimal number which we can define as the number of clusters [22]. The optimal number of clusters is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning.

One of the most popular method of finding optimal number of clusters is elbow method.

Elbow method

Elbow method is a method which looks at the percentage of variance explained as a function of the number of clusters [35]. This method exists upon the idea that one should choose a number of clusters so that adding another cluster doesn't give much better modelling of the data. In this method a plotted diagram will be created. Here the percentage of variance explained by the clusters will be plotted against the number of clusters [34]. The first clusters will add much information but at some point the marginal gain will drop dramatically and gives an angle in the graph.

```
1. Initialize $k=1$
2. Start
3. Increment the value of $k
4. Measure the cost of the optimal quality solution
5. If at some point the cost of the solution drops dramatically
6. That‟s the true $k$.
7. End
```

Based on pre-evaluated cluster number, the cluster nodes start the computations and divide themselves in the clusters according to the pre-evaluation. The cluster nodes divide themselves in the pre-evaluated number of clusters using Euclidean distance calculation. The cluster formation is performed using the K-Means algorithm.

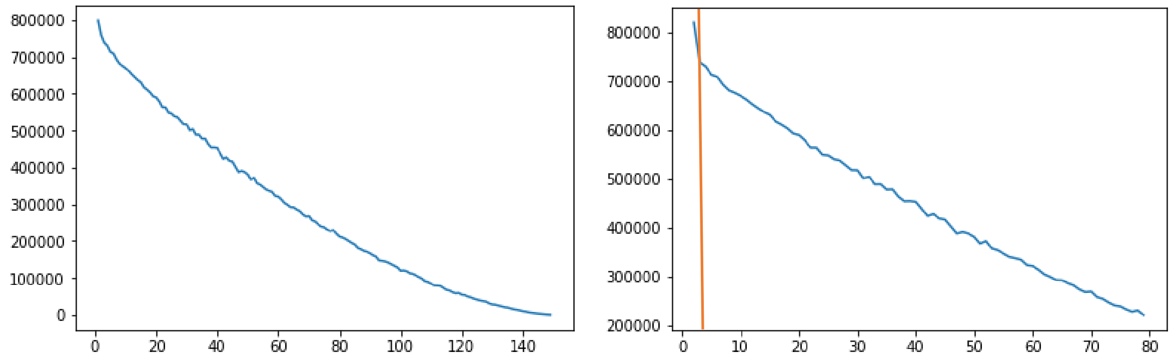


Figure 25: Selecting optimal number of clusters

In order to find the optimal number of clusters diagrams were plotted using the percentage of variance explained by the clusters against the number of clusters. The diagram was plotted several times graphs (Figure 25) reducing the number of cluster to ensure the optimal number of clusters by improving the elbow shape of the graph.

The graphs which were plotted without dimensional reduction show that the optimal number of clusters can be found in between 0 – 20. In order to have a clear number for this dimensional reduction process were carried out on top of the data set and then applied the elbow method [35].

Dimensional reduction will not be used when clustering the data as each value of each gene is important when deciding some kind of disease or deciding on cancers or any type of curing method.

Dimensional reduction using PCA (Principal Component Analysis)

PCA is often useful to measure data in terms of its principal components rather than on a normal x-y axis. They are the directions where there is the most variance, the directions where the data is most spread out.

As above mentioned even after the PCA applied on the data set, the diagram was plotted several graph (Figure 26) times to have a better elbow shape. The last plotted graph using 2-10 number of clusters shows a better elbow shape where the optimal number of clusters is shows as three.

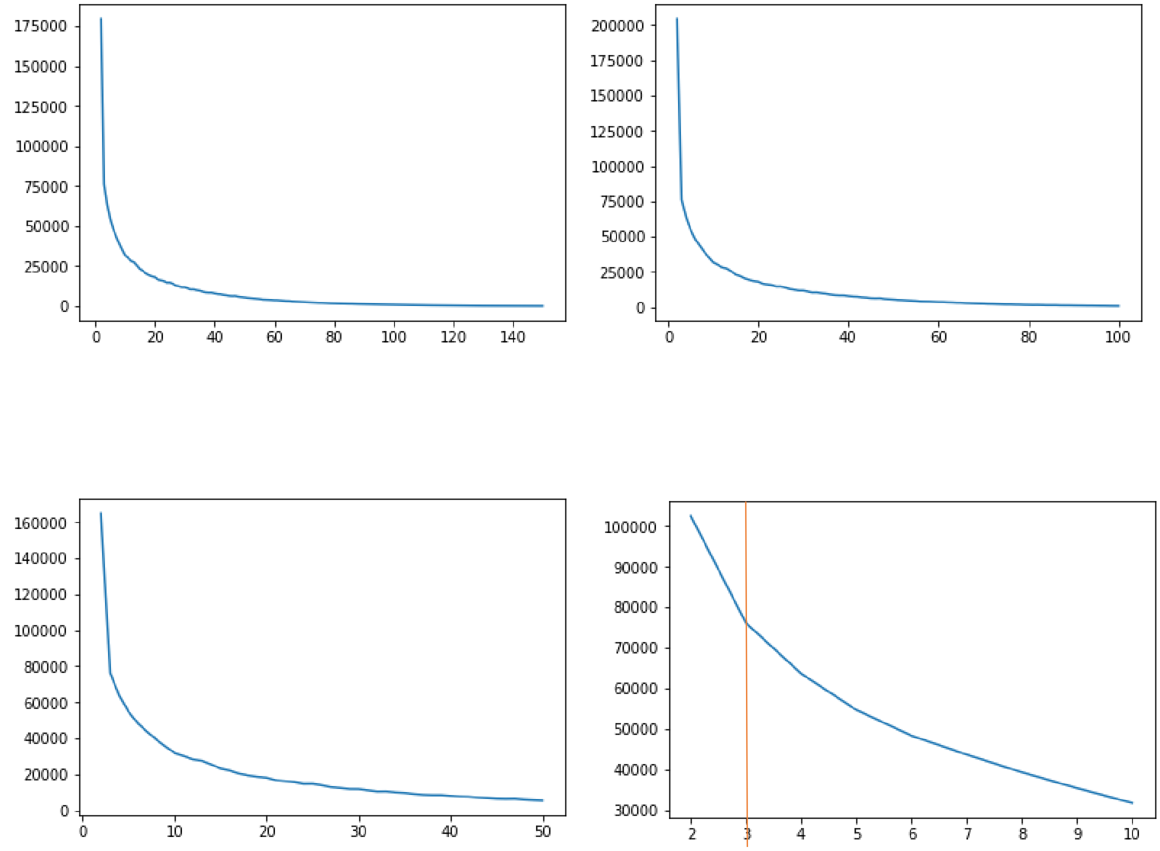


Figure 26: Selecting optimal number of clusters after PCA

Approach 1

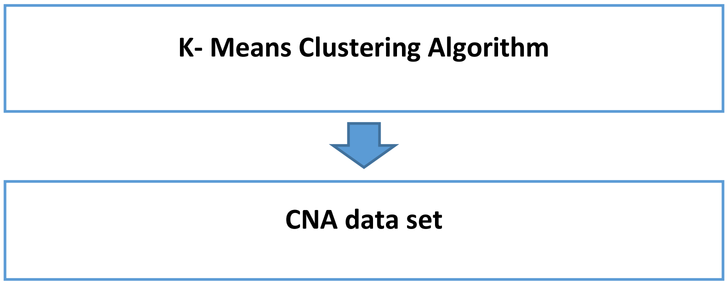


Figure 27: Approach one

Then the k- means clustering algorithm was applied to the data set giving the input of number of data as three. As the 1st step none of the dimension reduction methods were applied when applying the clustering method. After applying the K-means algorithm three clusters were performed using Euclidean distance measure as follows (Figure 27) [22]. Three clusters were performed (Figure 28) as 0, 1 and 2. Corresponding sample ID's are shown in below table (Refer Appendix A).

- Cluster 0 – 19 data points
- Cluster 1 – 57 data points
- Cluster 2 – 42 data points

After clustering the data using K means clustering algorithm, Silhouette score was calculated for each data point in order to find the how well the each data point matches with their clusters.

Silhouette score

Silhouette analysis can be used to study the separation distance between the resulting clusters. This measure has a range of 1- (-1). This measure gives the idea of how close each data point in one cluster to points in neighboring clusters [29].

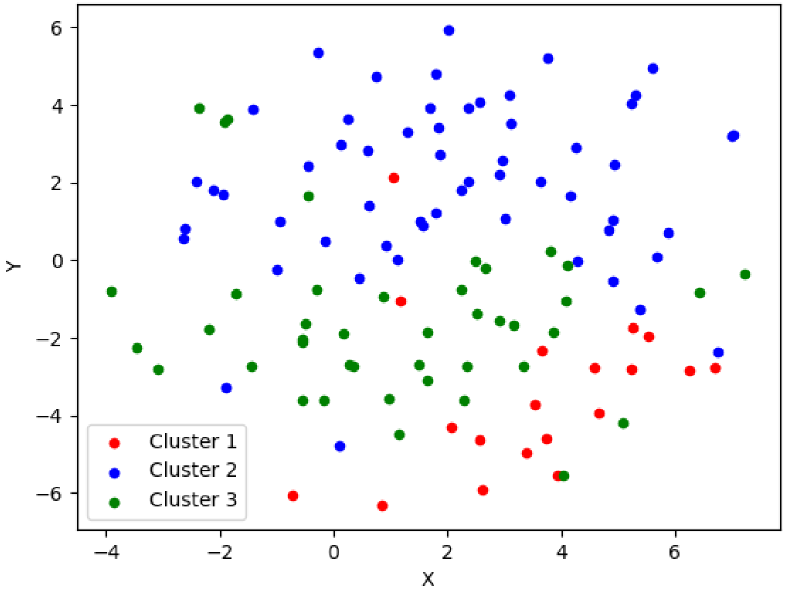


Figure 28: Clusters using K-means without dimensional reduction

Silhouette coefficients near +1 indicate that the sample is far away from the neighboring clusters [28]. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

As per the above diagrams (Graph 1) there are some data points which are positive values as well as negative values. These data points with the negative values might be in the wrong cluster. These data points are as follows (Table 2),

Sample	Score
'MO_1040'	-0.0204
'MO_1054'	-0.0061
'MO_1114'	-0.0148
'6115242'	-0.0072
'SC_9036'	-0.01
'SC_9055'	-0.0061
'SC_9073'	-0.0103
'SC_9094'	-0.0093

Table 2: Data points which have negative values for silhouette score

The average silhouette score for this set of cluster is calculated as 0.0258.

As the 2nd step same data set was clustered after dimension reduction. As the dimensional reduction techniques three popular techniques were selected. They are,

- t-SNE
- PCA
- ICA

Approach 2

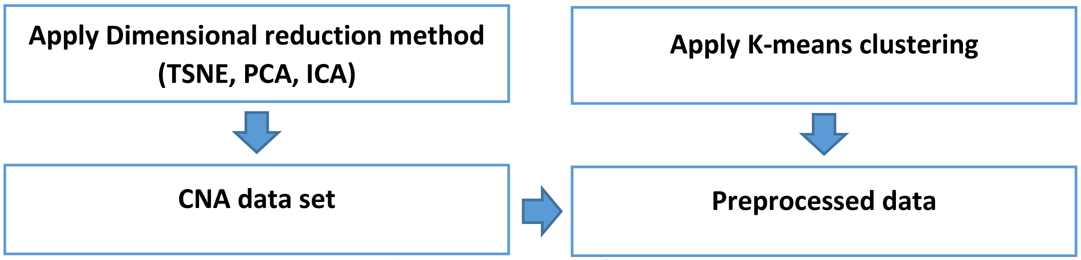
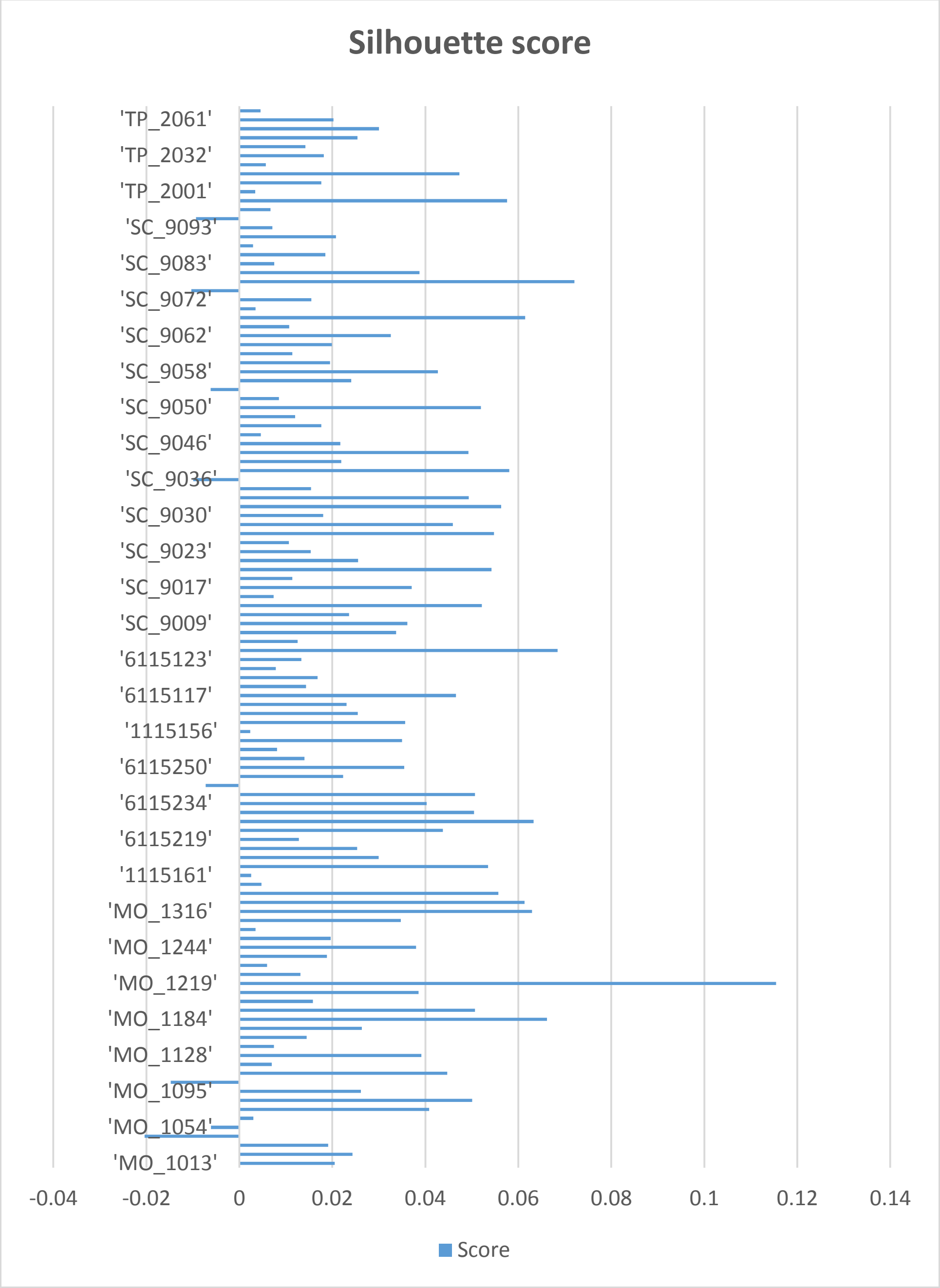


Figure 29: Approach two



Graph 1: Silhouette score for approach one

Most of the time when analyzing and visualizing the high dimensional data, dimensional reduction is carried out as a main step [26]. There are different types of methods which can be used for this dimensional reduction. t-SNE one of the most popular dimensionality reduction method which is proposed by Geoffry Hinton's group back in 2008.

t-SNE creates a low dimensional mapping using the local relationships between points. This is aiming to capture nonlinear structures not linear projections. t-SNE use Gaussian distribution to create a probability distribution which defines the relationships in dimensional space.

As the 2nd step t-SNE dimensional reduction technique were applied to the data set. Then again applied the k means clustering algorithm to analyze whether the clusters formed after t-SNE dimensional reduction is more accurate [27] (Figure 29).

Table in appendix B shows the clusters that were formed after t-SNE dimensional reduction. There are some data points which change the corresponding cluster which were performed in earlier step (Figure 30).

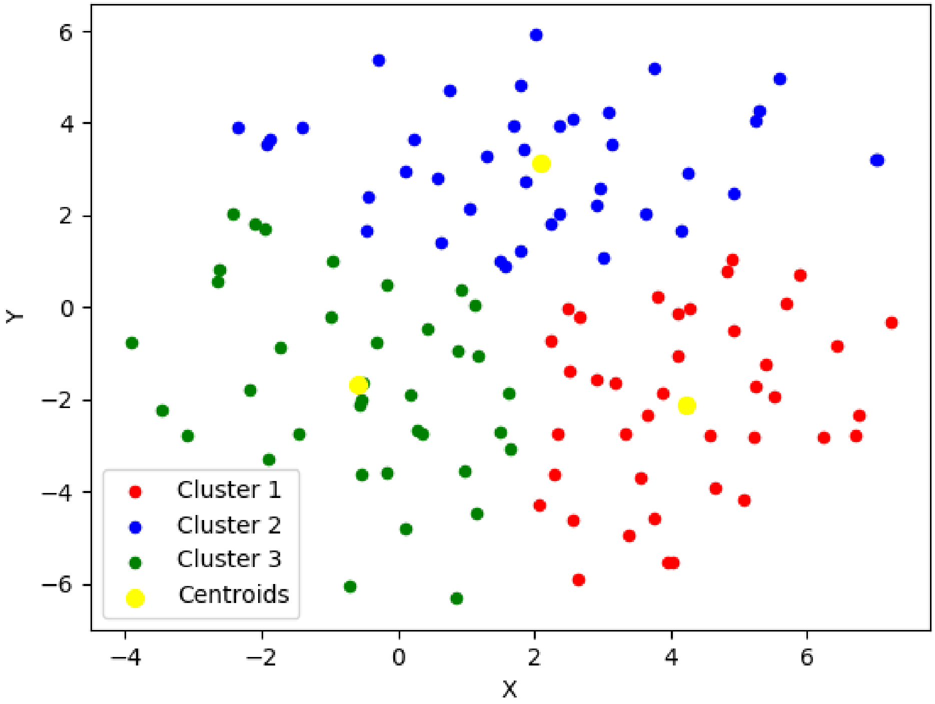
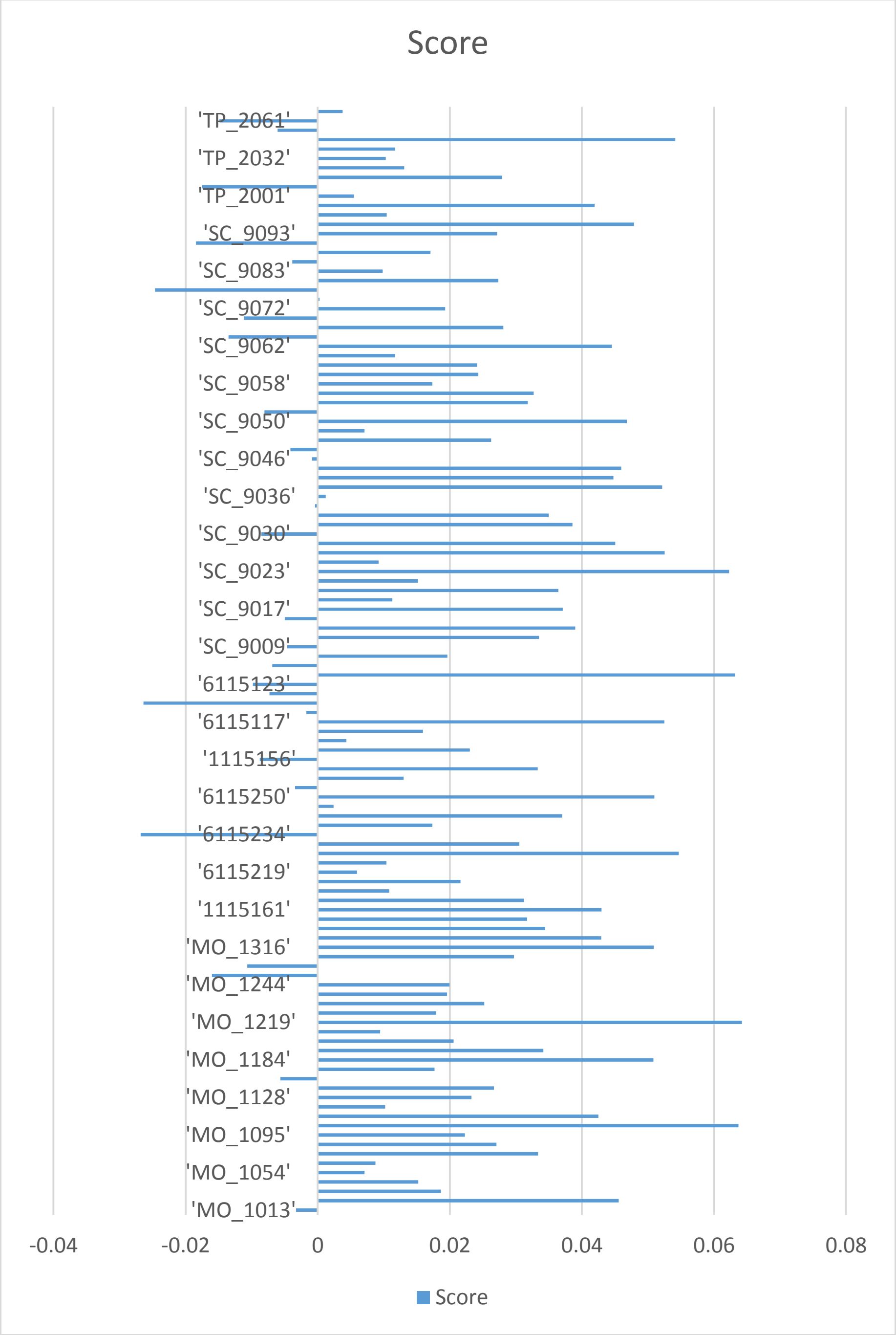


Figure 30: Clusters after K- means with t-SNE dimensional reduction

The average silhouette score and the separate silhouette score for each data point were calculated for these clusters in order to find out the correctness of the clusters. The average silhouette score for these clusters is 0.01922. Each silhouette score for each data point is as follows (Graph 2).



Graph 2: Silhouette score for approach two

Same as the 1st step negative values can be seen in this step also which means some of the data points might be in wrong cluster. Those data points are as follows (Table 3),

Data point	Score
'MO_1013'	-0.0033
'MO_1176'	-0.0057
'MO_1249'	-0.016
'MO_1262'	-0.0107
'6115234'	-0.0268
'6115251'	-0.0034
'1115156'	-0.0088
'6115118'	-0.0017
'6115121'	-0.0264
'6115122'	-0.0073
'6115123'	-0.0098
'SC_9007'	-0.0069
'SC_9009'	-0.0046
'SC_9016'	-0.005
'SC_9030'	-0.0085
'SC_9034'	-0.0004
'SC_9046'	-0.0009
'SC_9047'	-0.0041
'SC_9054'	-0.0081
'SC_9063'	-0.0135
'SC_9071'	-0.0112
'SC_9080'	-0.0246
'SC_9086'	-0.0039
'SC_9092'	-0.0184
'TP_2009'	-0.0175
'TP_2060'	-0.0061
'TP_2061'	-0.0149

Table 3: Data points which have negative values for silhouette score

Approach 3

When comparing with the data points which has negative values in the 1st step number of data point is higher than earlier. Which means the cluster set in step 1 is more accurate than these clusters.

As the 3rd step another dimensional reduction method were applied for the data set and then applied k means clustering. PCA dimensional reduction method were applied here.

Principal component analysis (PCA) is a method which is used to create set of linearly uncorrelated variables called principal components using orthogonal transformation (Figure 31).
Table in appendix C shows the clusters performed after PCA dimensional reduction.

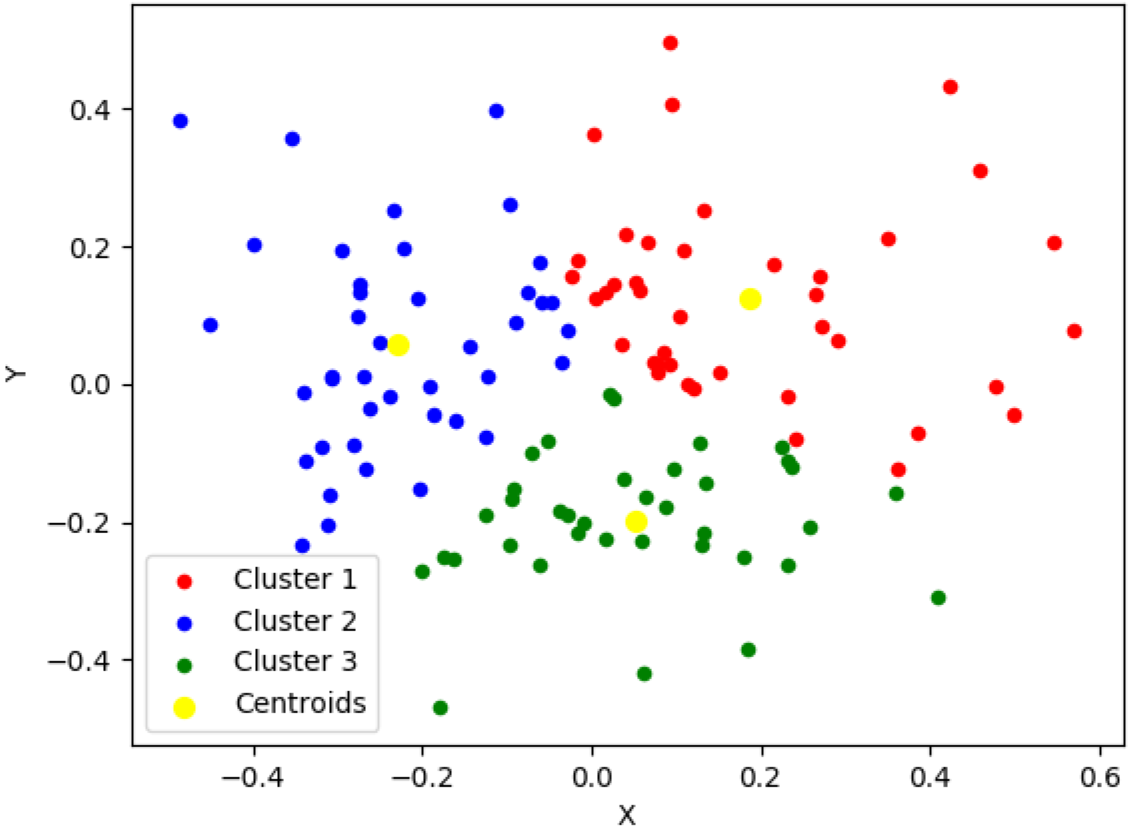


Figure 31: Clusters after K- means with PCA dimensional reduction

The average silhouette score for this cluster set is 0.02310. Corresponding silhouette score for each value is shown in appendix D.

Negative values as well as positive values can be also seen here. The data points which has negative values are as follows (Table 4),

Data point	Score
'MO_1014'	-4.39E-03
'1115154'	-1.43E-03
'1115157'	-4.69E-03

'6115118'	-5.84E-04
'SC_9034'	-6.54E-03
'SC_9092'	-4.02E-03
'SC_9097'	-5.93E-03

Table 4: Data points which have negative values for silhouette score

Approach 4

As the next step ICA dimensional reduction technique was used before applying the k means clustering method. ICA stands for Independent Components Analysis. In ICA it consider that each sample of data is a mixture of independent components and it aims to find these independent components (Figure 32).

Table in appendix E shows the clusters which were performed by k means clustering after applying ICA dimensional reduction technique for the data set.

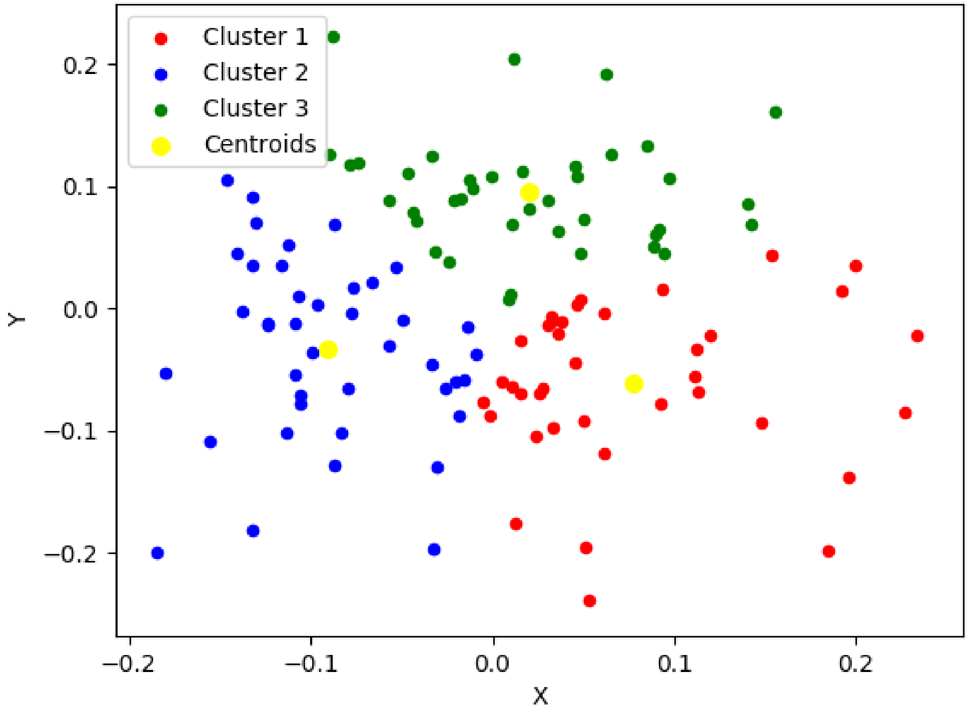
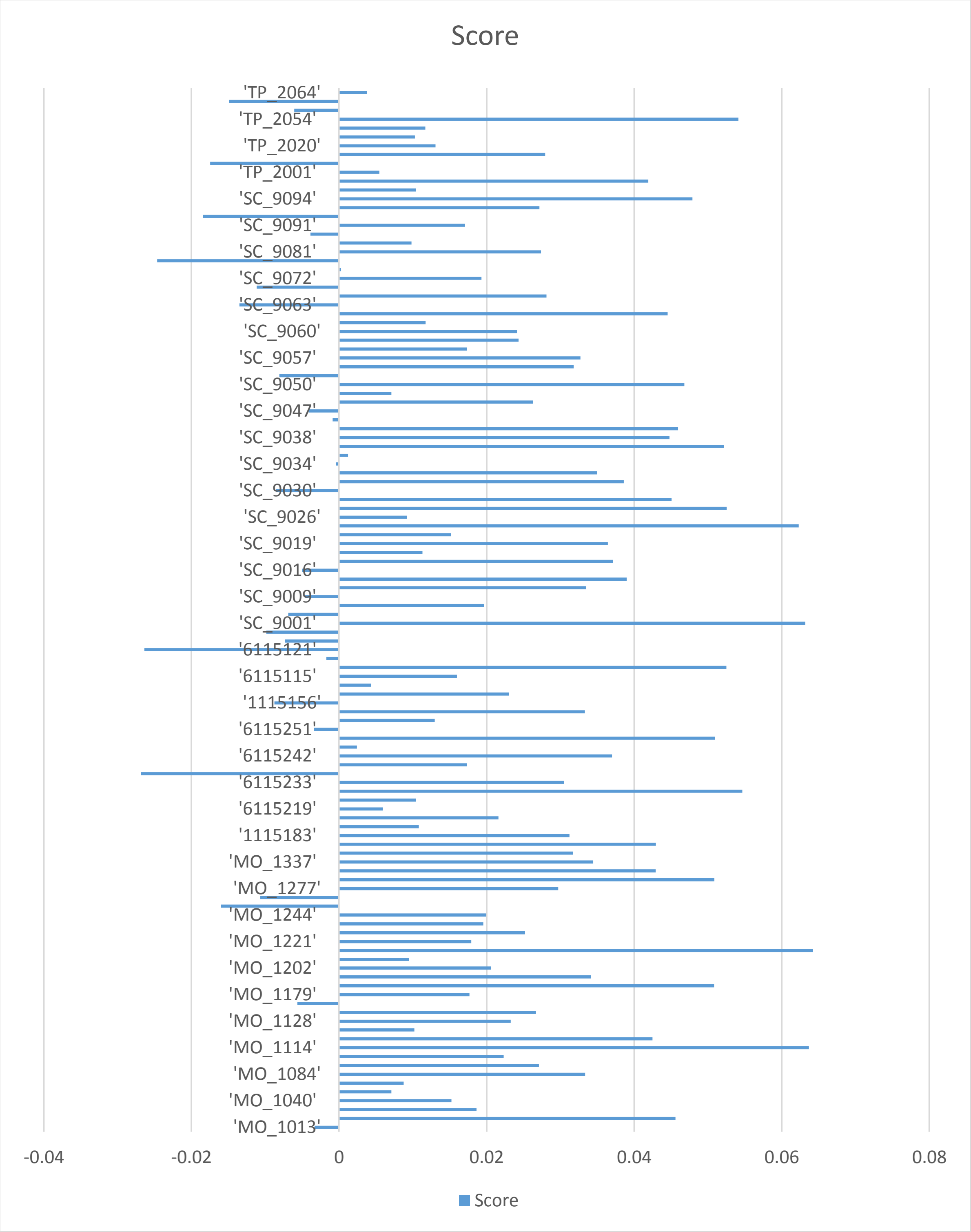


Figure 32: Clusters after K- means with ICA dimensional reduction

The average silhouette score for this cluster set is 0.023100687694212127. Corresponding silhouette score for each value is as follows (Graph 3),



Graph 3: Silhouette score for approach three

In this method also there are some data points which got negative values as the silhouette score.

Data point	Score
'MO_1013'	-0.0033
'MO_1176'	-0.0057
'MO_1249'	-0.016
'MO_1262'	-0.0107
'6115234'	-0.0268
'6115251'	-0.0034
'1115156'	-0.0088
'6115118'	-0.0017
'6115121'	-0.0264
'6115122'	-0.0073
'6115123'	-0.0098
'SC_9007'	-0.0069
'SC_9009'	-0.0046
'SC_9016'	-0.005
'SC_9030'	-0.0085
'SC_9034'	-0.0004
'SC_9046'	-0.0009
'SC_9047'	-0.0041
'SC_9054'	-0.0081
'SC_9063'	-0.0135
'SC_9071'	-0.0112
'SC_9080'	-0.0246
'SC_9086'	-0.0039
'SC_9092'	-0.0184
'TP_2009'	-0.0175
'TP_2060'	-0.0061
'TP_2061'	-0.0149

Table 5: Data points which have negative values for silhouette score

When considering all four steps which carried out k means clustering method, above (Table 5) are the data points with negative silhouette score in all the steps.

Approach 5

Apply one hot encoding on CNA data and then K-means clustering

CNA data set consists of categorical data but they have been represented as numerical data as explain in the ‘Data representation in selected data source’ section above. It only includes numerical numbers -2, -1, 0, 1, 2, representing five different categories. In above sections K-means clustering algorithm were carried out on top of these data without converting the categorical data into numerical data. As the next step these data will be converted to categorical data using one hot encoding method which, a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction (Figure 33).

With this data set three clusters were performed using K-means clustering with the average silhouette score of -0.0937. The distribution of the clusters are as below.

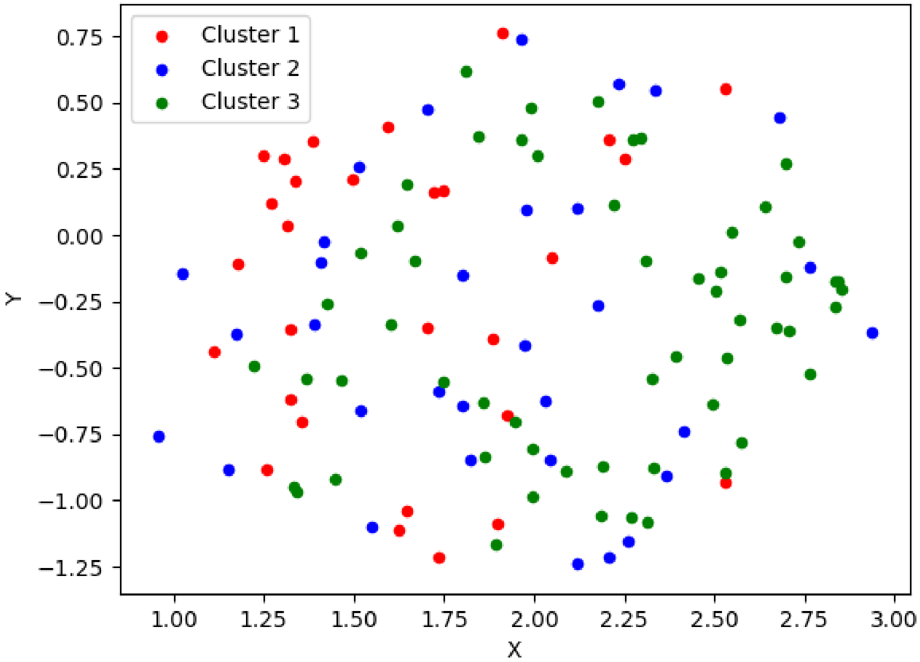


Figure 33: Clusters applying one hot encoding on CNA data and then K-means clustering

This distribution of cluster points doesn’t show good density of each cluster and the good separation of clusters. Hence dimensional reduction method were applied on top this data set in order to find whether any better cluster set can be performed.

Approach 6

Then dimensional reduction method was applied as applying one hot method will create a data set with large number of zeros. Hence dimensional reduction need to be carried out. As the dimensional reduction method PCA method was used. Then K-means clustering method was applied for this preprocessed data set (27). Clusters (Refer appendix F) were performed after this method with the average silhouette score of 0.3812.

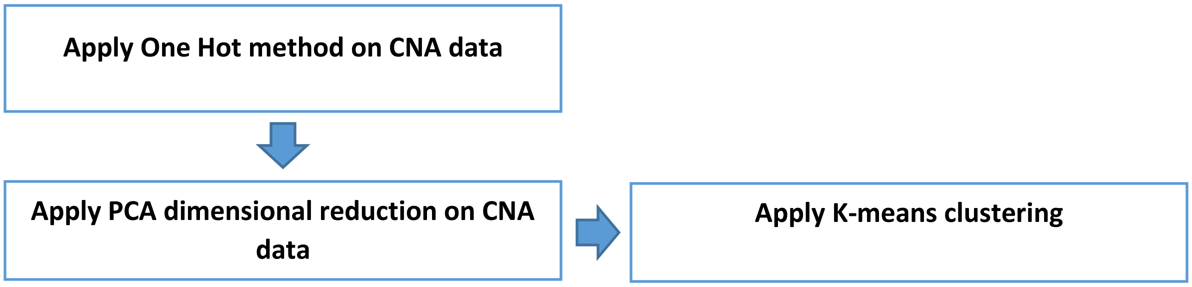

Apply One Hot method on CNA data
Apply PCA dimensional reduction on CNA data
Apply K-means clustering

Figure 34: Approach Three

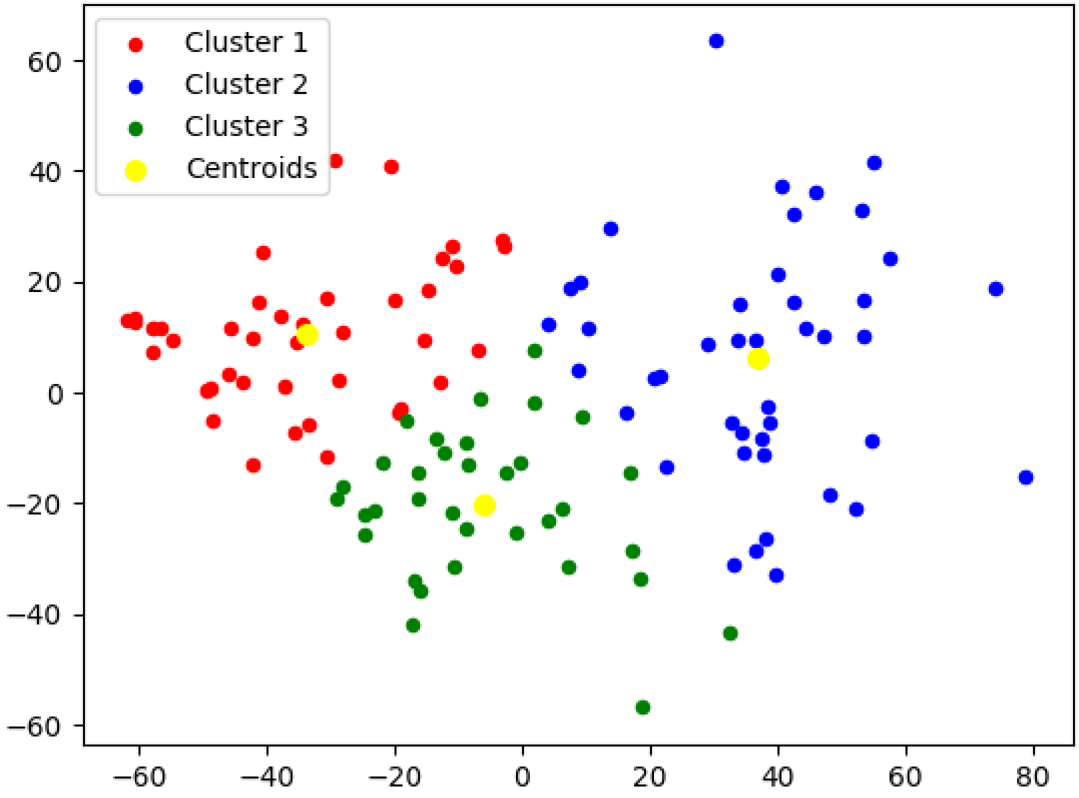


Figure 35: Clusters after approach three

As per the graph above (Figure 35) distribution of the clusters are better than the above phases which were carried out without converting the categorical data into numerical data.

Approach 7

Since above method gives better clusters, same method was applied on the corresponding expression data set for the same genes and the same sample set, which is a dense data set. PCA dimension reduction method was applied on this data set 1st then k-means clustering method was used. Created clusters (Refer appendix G) were compared with the above clusters in order to identify the similarities and the dissimilarities with those clusters (Figure 36).

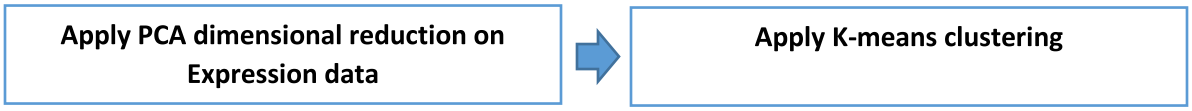


Figure 36: Approach four

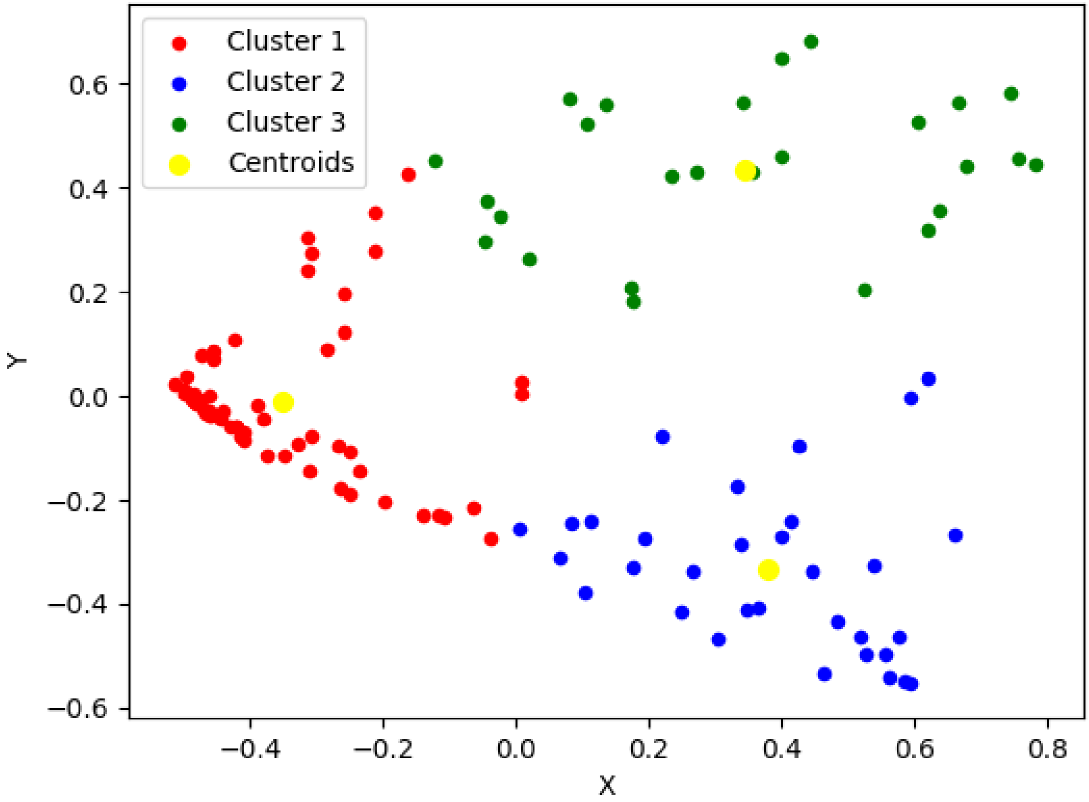


Figure 37: Clusters after approach four

For the dense data set above (Figure 37) three clusters were performed using k-means clustering method with the average silhouette score of 0.5847.

5.3.2. Hierarchical Clustering method

Hierarchical clustering is also a well-known clustering method in the literature. In this method a dendogram will be used to cluster the data points in to clusters [38]. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

In the very 1st step each data point will be consider as different clusters. Then it recursively execute the below steps (Figure 38).

(1) Identify the two clusters that are closest together, and

(2) Merge the two most similar clusters.

This continues until all the clusters are merged together. This is illustrated in the diagrams below.

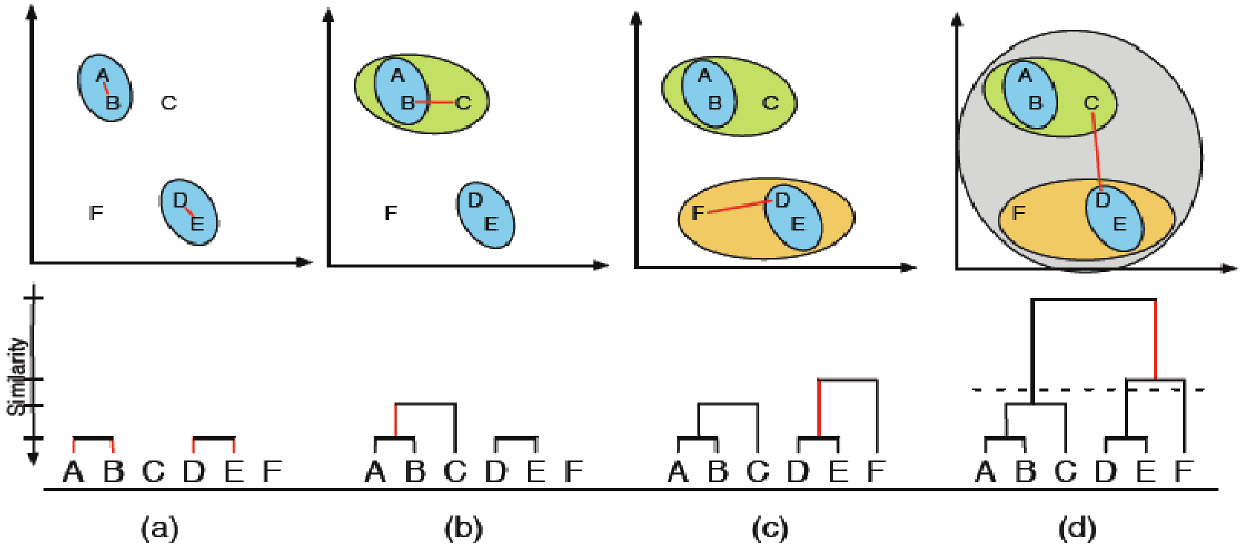


Figure 38: Steps of Hierarchical clustering algorithm

As the next step of the research the data set was clustered using this hierarchical clustering method. Eucledian distance, Manhatan distance and Minkowski distance measures were used when clustering the data set using hierarchical clustering method. Similar dendogram was populated by all three measures (Figure 39).

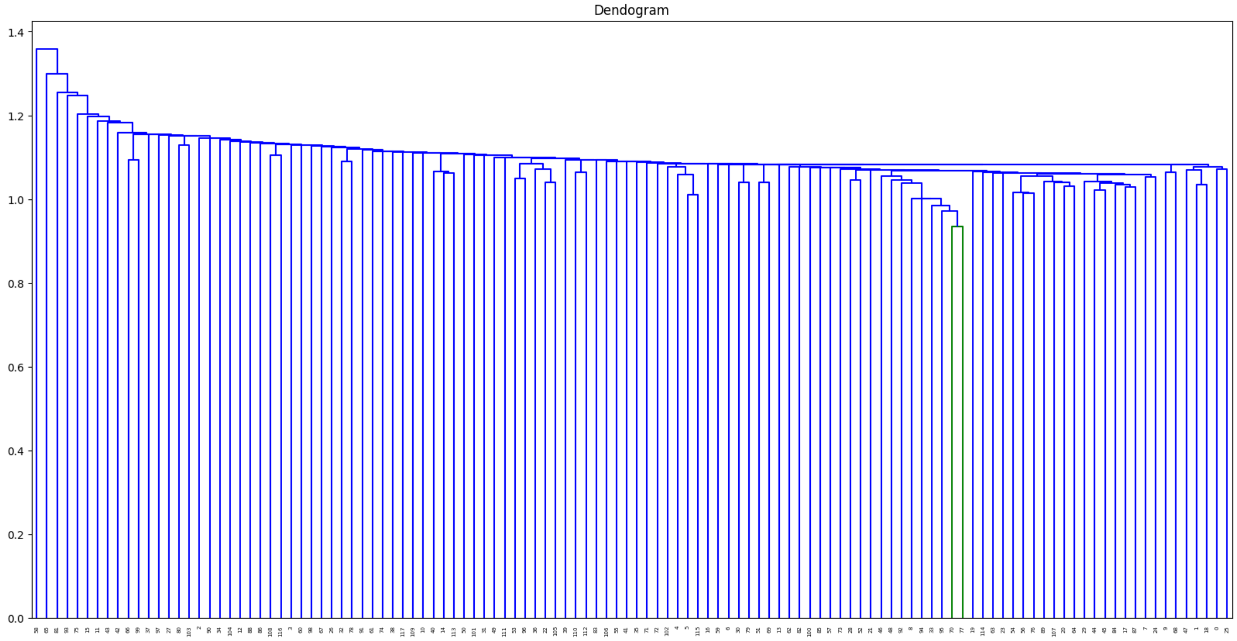


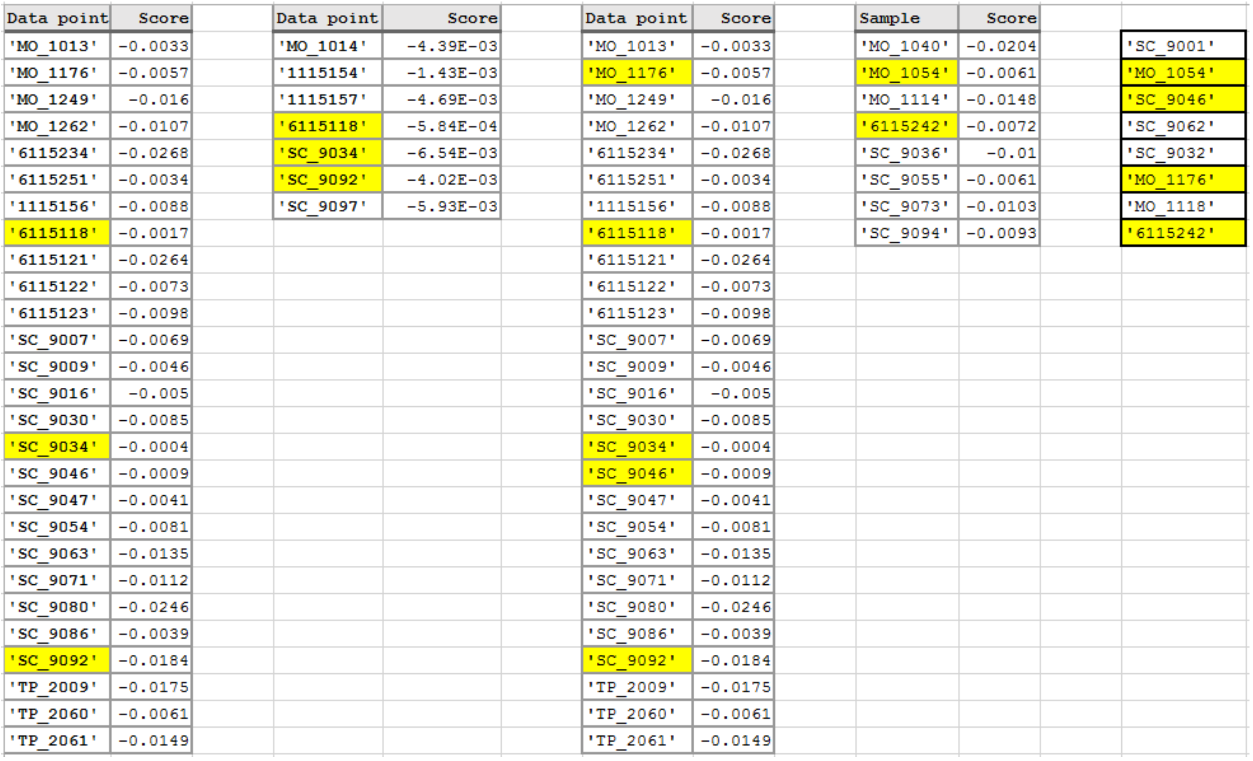
Figure 39: Dendogram using hierarchical clustering algorithm

Then the dendogram were cut where three clusters were generated. But here two clusters got one data point in each and other 116 data points were included in one cluster. Hence this dendogram was used to identify the outliers [35].

Result of this dendogram and results of all the clustering steps which used K- means clustering above were used to identify the outliers. Data points which has negative values as silhouette score in above steps and the data points which have connected to the dendogram very lastly were compared and the same data points which were in most of those categories were selected as the outliers.

Below highlighted cells (Table 6) shows the data points which were recognized as a outlier at least three times. Below are the data points which will be considered as outliers,

'6115118'
'SC_9034'
'SC_9092'
'MO_1054'
'SC_9046'
'MO_1176'
'6115242'



Data point	Score	Data point	Score	Data point	Score	Sample	Score	
'MO_1013'	-0.0033	'MO_1014'	-4.39E-03	'MO_1013'	-0.0033	'MO_1040'	-0.0204	'SC_9001'
'MO_1176'	-0.0057	'1115154'	-1.43E-03	'MO_1176'	-0.0057	'MO_1054'	-0.0061	'MO_1054'
'MO_1249'	-0.016	'1115157'	-4.69E-03	'MO_1249'	-0.016	'MO_1114'	-0.0148	'SC_9046'
'MO_1262'	-0.0107	'6115118'	-5.84E-04	'MO_1262'	-0.0107	'6115242'	-0.0072	'SC_9062'
'6115234'	-0.0268	'SC_9034'	-6.54E-03	'6115234'	-0.0268	'SC_9036'	-0.01	'SC_9032'
'6115251'	-0.0034	'SC_9092'	-4.02E-03	'6115251'	-0.0034	'SC_9055'	-0.0061	'MO_1176'
'1115156'	-0.0088	'SC_9097'	-5.93E-03	'1115156'	-0.0088	'SC_9073'	-0.0103	'MO_1118'
'6115118'	-0.0017			'6115118'	-0.0017	'SC_9094'	-0.0093	'6115242'
'6115121'	-0.0264			'6115121'	-0.0264			
'6115122'	-0.0073			'6115122'	-0.0073			
'6115123'	-0.0098			'6115123'	-0.0098			
'SC_9007'	-0.0069			'SC_9007'	-0.0069			
'SC_9009'	-0.0046			'SC_9009'	-0.0046			
'SC_9016'	-0.005			'SC_9016'	-0.005			
'SC_9030'	-0.0085			'SC_9030'	-0.0085			
'SC_9034'	-0.0004			'SC_9034'	-0.0004			
'SC_9046'	-0.0009			'SC_9046'	-0.0009			
'SC_9047'	-0.0041			'SC_9047'	-0.0041			
'SC_9054'	-0.0081			'SC_9054'	-0.0081			
'SC_9063'	-0.0135			'SC_9063'	-0.0135			
'SC_9071'	-0.0112			'SC_9071'	-0.0112			
'SC_9080'	-0.0246			'SC_9080'	-0.0246			
'SC_9086'	-0.0039			'SC_9086'	-0.0039			
'SC_9092'	-0.0184			'SC_9092'	-0.0184			
'TP_2009'	-0.0175			'TP_2009'	-0.0175			
'TP_2060'	-0.0061			'TP_2060'	-0.0061			
'TP_2061'	-0.0149			'TP_2061'	-0.0149			

Table 6: Selecting outliers

5.3.2. DBSCAN (Density-based spatial clustering of applications with noise) Clustering method

In density based clustering, areas which have higher density other than other areas are highlighted as the lusters. Other areas which have less number of data points which can be considered as sparse areas usually treated as the noise and the border points. DBSCAN is one of the most popular density based clustering methods.

One of the data set which was described in chapter 3 has dense data which is known as expression data. Since all the above clustering processes were applied on CNA data, in this step expression data will be used to apply the clustering algorithm. This data set includes expression data of selected genes for the selected samples. In the original data set there were about 40000 genes, when comparing to the corresponding CNA data set it only has about 25000 genes. Expression data set was prepared as the CNA and the expression data set have same gene set.

DBSCAN clustering method was used to cluster the expression data [31]. This algorithm was run on the expression data set changing the parameters which have defined in DBSCAN clustering method in order to find clusters with high similarity among the data points. Below are the parameters which were used for clustering the data set [33].

eps: the minimum distance between two points. It means that if the distance between two points is lower or equal to this value (eps), these points are considered neighbors.

minPoints: the minimum number of points to form a dense region. For example, if we set the minPoints parameter as 5, then we need at least 5 points to form a dense region.

Since the main objective of the research is explore a novel clustering method, might be a hybrid method connecting the outputs of each different clustering method, some of the outputs of K-means clustering method will be used as the input to the DBSCAN method [27]. Hence number of effective clusters which was found using elbow method will be used for the DBSCAN clustering method. As the first step DBSCAN clustering method was used to cluster the dense data set, expression data set for the same gene set and the same sample set as the CNA data set to make three clusters [31] (Figure 40).

Approach 8

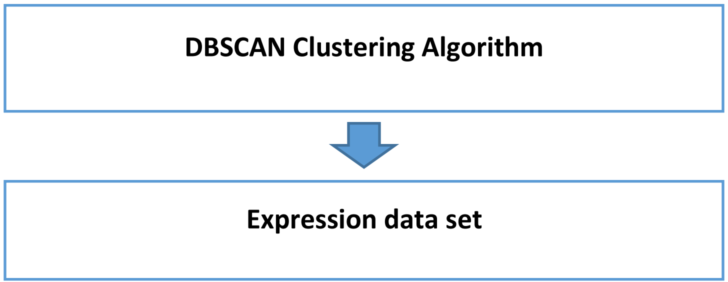


Figure 40: Approach eight

Since the number of clusters are defined by the parameters which have mentioned above, DBSCAN algorithm was tuned using different values for those parameters and found out the values which perform three clusters.

Below are the values which performed three clusters,

```
eps:  0.15151515151515152 min_samples:  2.0 Number of clusters:  3
eps:  0.15454545454545454 min_samples:  2.0 Number of clusters:  3
```

```
eps:  0.1575757575757576 min_samples:  2.0 Number of clusters:  3
eps:  0.1606060606060606 min_samples:  2.0 Number of clusters:  3
eps:  0.19090909090909092 min_samples:  3.0 Number of clusters:  3
eps:  0.19393939393939397 min_samples:  3.0 Number of clusters:  3
eps:  0.196969696969697 min_samples:  3.0 Number of clusters:  3
eps:  0.2 min_samples:  3.0 Number of clusters:  3
eps:  0.20303030303030306 min_samples:  3.0 Number of clusters:  3
eps:  0.20606060606060608 min_samples:  3.0 Number of clusters:  3
eps:  0.2090909090909091 min_samples:  3.0 Number of clusters:  3
eps:  0.21212121212121215 min_samples:  3.0 Number of clusters:  3
eps:  0.21515151515151518 min_samples:  4.0 Number of clusters:  3
eps:  0.2181818181818182 min_samples:  4.0 Number of clusters:  3
eps:  0.22121212121212125 min_samples:  4.0 Number of clusters:  3
eps:  0.22424242424242427 min_samples:  4.0 Number of clusters:  3
eps:  0.2272727272727273 min_samples:  4.0 Number of clusters:  3
eps:  0.23030303030303031 min_samples:  4.0 Number of clusters:  3
eps:  0.23333333333333336 min_samples:  4.0 Number of clusters:  3
eps:  0.2363636363636364 min_samples:  4.0 Number of clusters:  3
eps:  0.2393939393939394 min_samples:  4.0 Number of clusters:  3
eps:  0.24242424242424246 min_samples:  4.0 Number of clusters:  3
eps:  0.24545454545454548 min_samples:  4.0 Number of clusters:  3
eps:  0.2484848484848485 min_samples:  4.0 Number of clusters:  3
eps:  0.2515151515151516 min_samples:  4.0 Number of clusters:  3
eps:  0.2545454545454546 min_samples:  4.0 Number of clusters:  3
eps:  0.25757575757575757 min_samples:  3.0 Number of clusters:  3
eps:  0.2606060606060606 min_samples:  3.0 Number of clusters:  3
eps:  0.26363636363636367 min_samples:  3.0 Number of clusters:  3
eps:  0.2666666666666667 min_samples:  3.0 Number of clusters:  3
eps:  0.26969696969696977 min_samples:  3.0 Number of clusters:  3
eps:  0.27272727272727276 min_samples:  3.0 Number of clusters:  3
eps:  0.27575757575757576 min_samples:  3.0 Number of clusters:  3
eps:  0.2787878787878788 min_samples:  3.0 Number of clusters:  3
eps:  0.28181818181818186 min_samples:  3.0 Number of clusters:  3
eps:  0.2848484848484849 min_samples:  3.0 Number of clusters:  3
eps:  0.2878787878787879 min_samples:  3.0 Number of clusters:  3
eps:  0.29090909090909095 min_samples:  2.0 Number of clusters:  3
eps:  0.29090909090909095 min_samples:  4.0 Number of clusters:  3
eps:  0.29393939393939394 min_samples:  4.0 Number of clusters:  3
eps:  0.296969696969697 min_samples:  4.0 Number of clusters:  3
eps:  0.30000000000000004 min_samples:  4.0 Number of clusters:  3
eps:  0.3030303030303031 min_samples:  4.0 Number of clusters:  3
```

Each of these value pairs were used at a time and observed the distribution of the clusters and calculate the average silhouette score. Then select the parameters which gives the best silhouette score and the best distribution of the clusters.

eps=0.3030303030303031, min_samples=4.0 were selected as the most effective parameter values for this data set. Below are the clusters (Refer appendix H) (Figure 41) which were perform by DBSCAN clustering for the above parameter values with the average silhouette score of 0.0474806

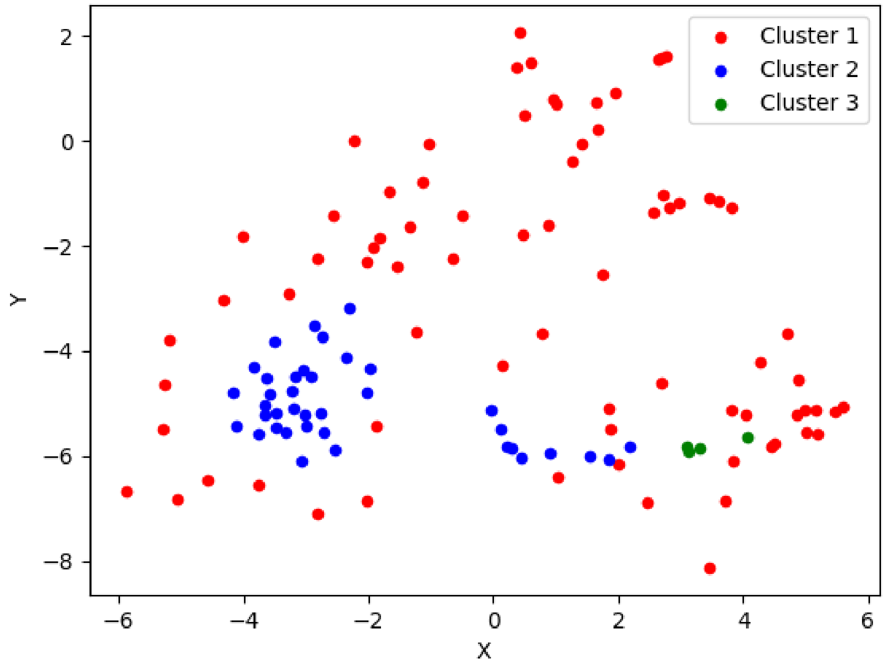


Figure 41: Clusters after approach eight

Approach 9

Then the same algorithm were tried out after dimension reduction of the dense data set using TSNE and PCA as the dimension reduction methods. From these two approaches applying PCA gave better clusters compared to applying TSNE method as clusters with PCA dimension reduction got higher average silhouette score value compared to the other one (Figure 42). The clusters gained through this method is as follows (Refer appendix I) (Figure 43),

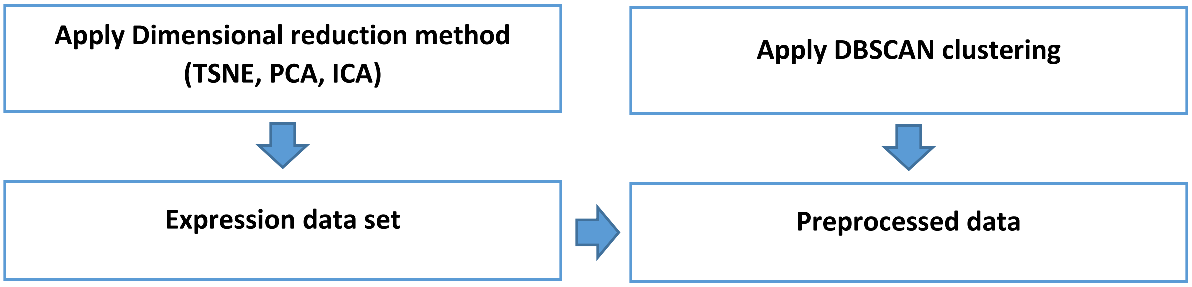


Figure 42: Approach nine

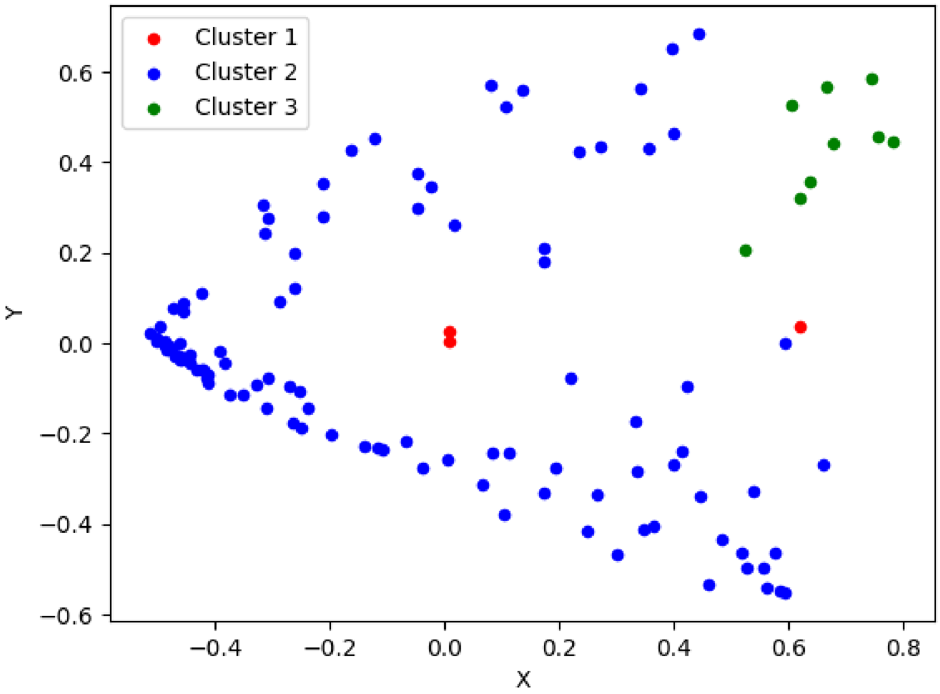


Figure 43: Clusters after approach nine

Approach 10

Since we are trying to compare the clustering methods and options as much as possible, then find a better way to perform better clusters, this DBSCAN method was applied also on the CNA data set which were preprocessed using one hot method. Here PCA method was used as the dimension reduction method (Figure 44). Using this method three clusters were performed (Refer appendix J) (Figure 45) with the average silhouette score is of -0.13301 and a better distribution of clusters.

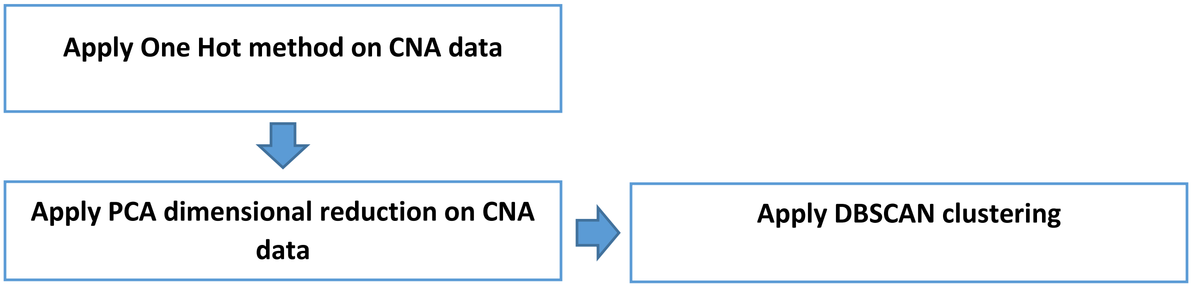

Apply One Hot method on CNA data
Apply PCA dimensional reduction on CNA data
Apply DBSCAN clustering

Figure 44: Approach ten

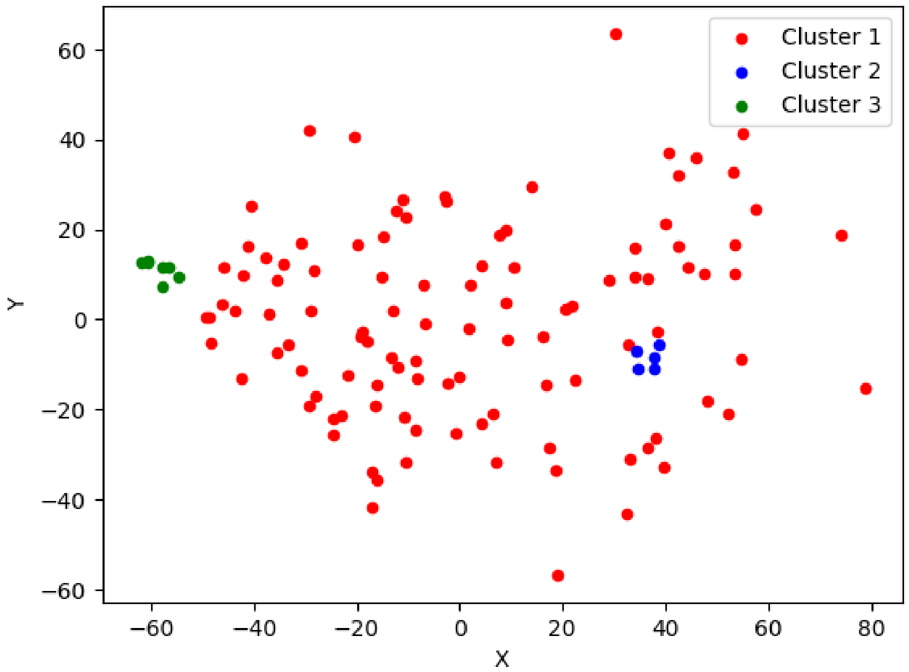


Figure 45: Clusters after approach ten

Approach 11

The next clustering approach will be carried out using both the data sets, CNA and expression data set at once, in order to explore the distribution and the similarities of the clusters which will be gain using both the features. For this method CNA data which were preprocessed using one hot method and the expression data will be used. PCA dimension reduction method was applied on both the data sets (Figure 46). Clusters were performed with the average silhouette score of 0.1553 and distribution as below (Figure 47) (Refer appendix K).

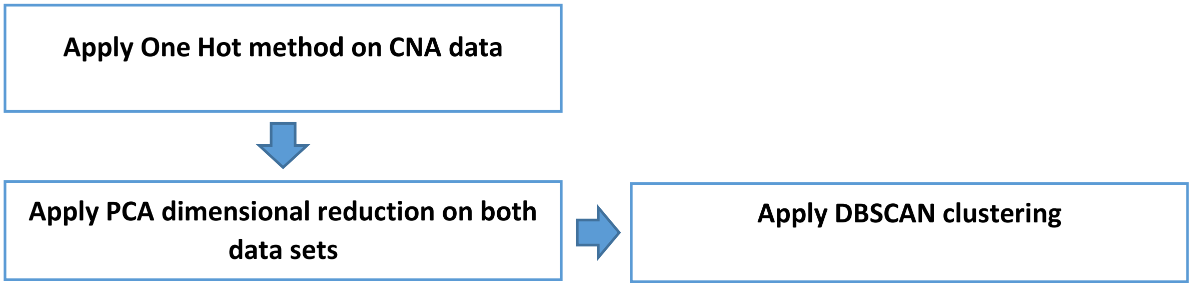

Apply One Hot method on CNA data
Apply PCA dimensional reduction on both data sets
Apply DBSCAN clustering

Figure 46: Approach eleven

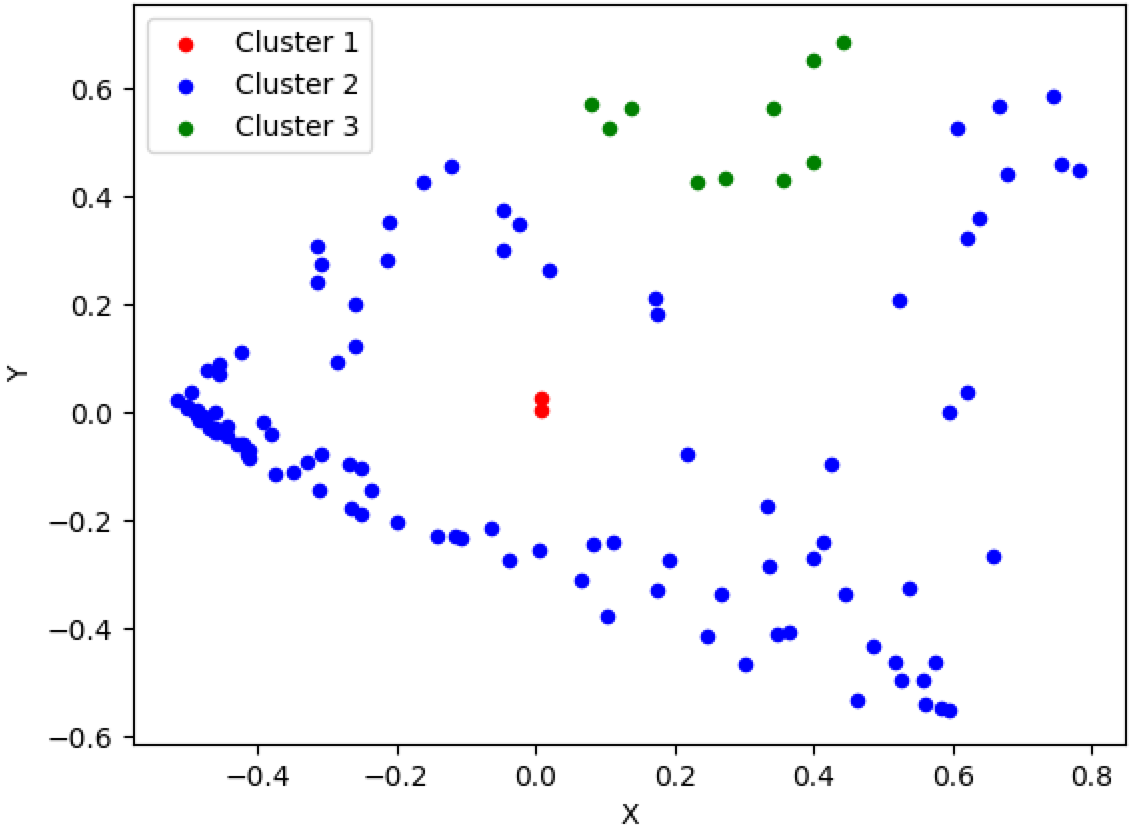


Figure 47: Clusters after approach eleven

Approach 12

As the next phase of applying DBSCAN on expression data set, outliers which were identified in the phase of applying K-means clustering on CNA data set, were removed from the expression data set. Then again apply DBSCAN clustering on the filtered data set (Figure 48).

Since the data set has filter out from the outliers, the values for the parameters (eps, min_samples) should be calculated again. Most suitable parameter values were identified considering number of clusters and the average silhouette score. Parameter values which performed three clusters were identified first and then the silhouette score was considered.

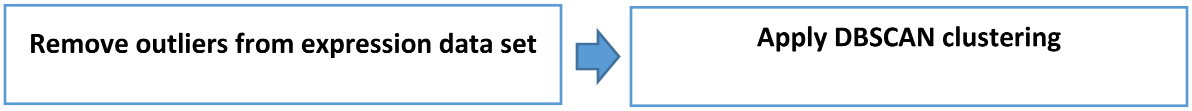


Figure 48: Approach twelve

eps=0.31818181818181823, min_samples=2.0 values were selected as the parameter values for applying the DBSCAN clustering on the expression data set without outlier. Three clusters were performed (Figure 49) (Refer appendix L) with above parameter values and the average silhouette score of 0.101227.

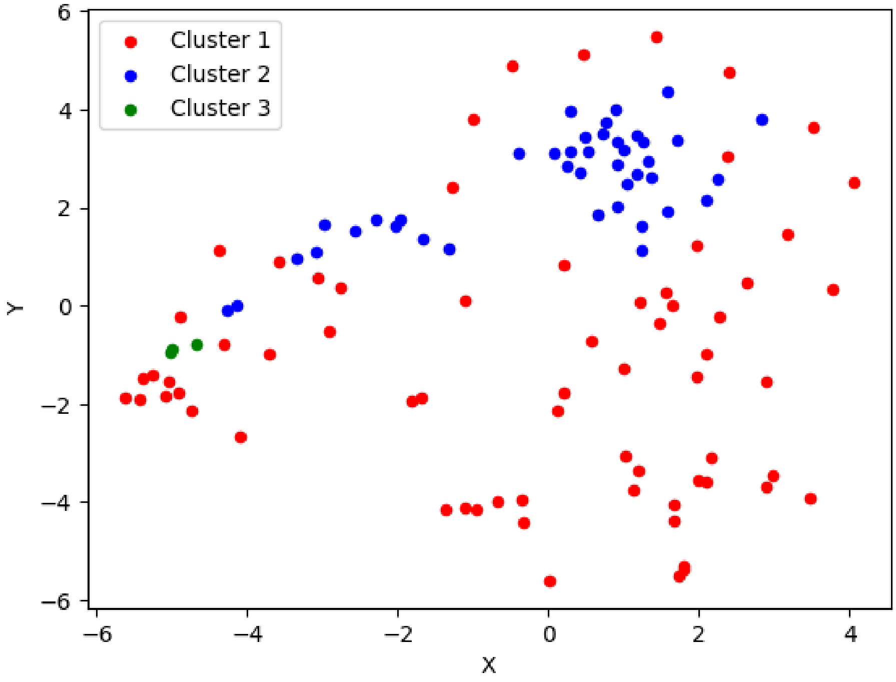


Figure 49: Clusters after approach twelve

Approach 13

As the next step of applying DBSCAN clustering method both the data set CNA and expression data sets were used at once as the inputs to the DBSCAN clustering method. CNA data set and the expression data set were store in two different 2D arrays. The value of the particular gene for particular sample is considered with the value of the expression data of that same gene for the same sample (Figure 50).

Considering both the points, DBSCAN method was applied on both the data sets. Below is the clusters which were performed using this method with the average silhouette score of 0.04748047. Even though there is a very little gain on silhouette score compared to the clusters which were performed using only expression data, the distribution (Figure 51) of the clusters are different from each other.

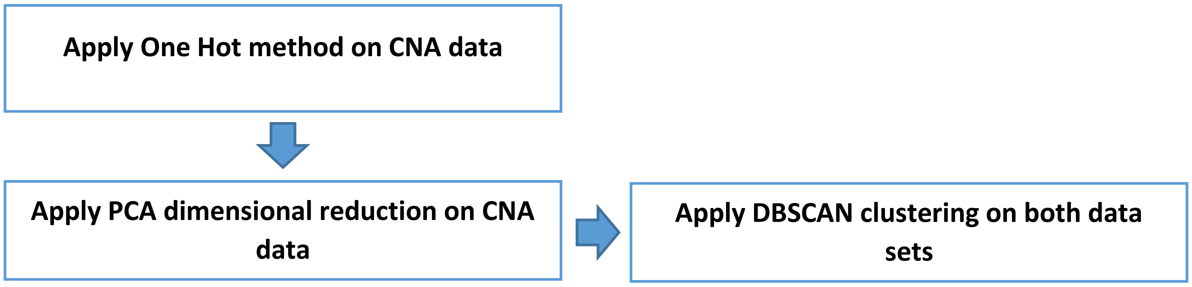

Apply One Hot method on CNA data
Apply PCA dimensional reduction on CNA data
Apply DBSCAN clustering on both data sets

Figure 50: Approach Thirteen

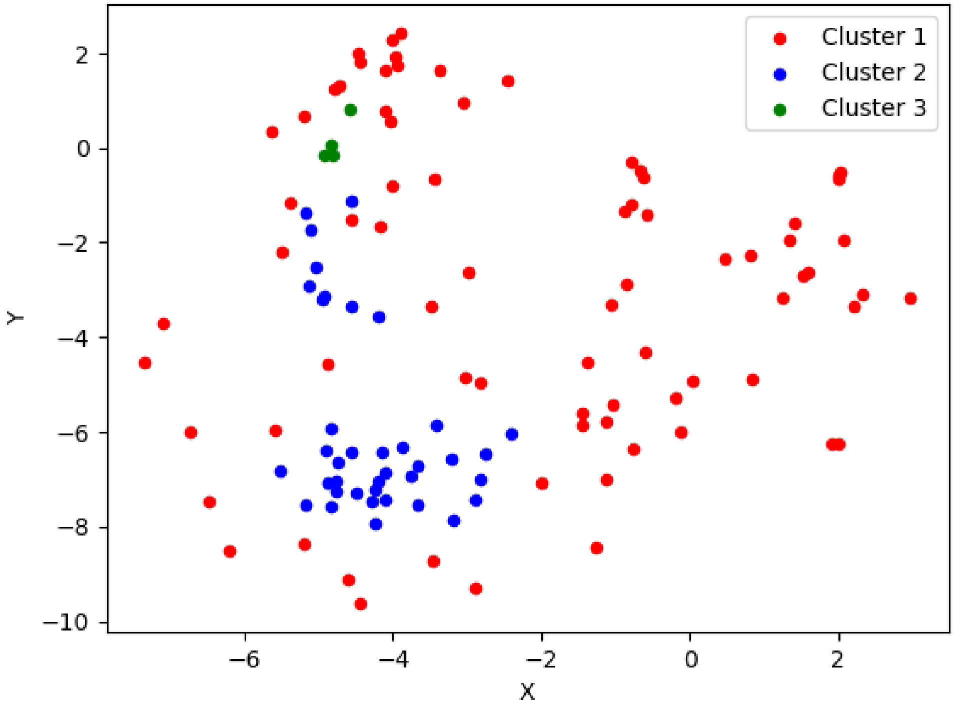


Figure 51: Clusters after approach thirteen

Approach 14

As the final approach of performing clusters of this research, outliers which were identified in K-means application phase, were removed from both the data sets, CNA and expression data set. Here CNA data set was preprocessed using one hot method and dimension reduction was done by using PCA method. Then apply the DBSCAN clustering for both the data set using same steps as the above process (Figure 52). This method performed three clusters (Refer appendix M) with the average silhouette score of 0.30222757, highest score from all the procedures that were carried out earlier without dimension reduction (Figure 53).

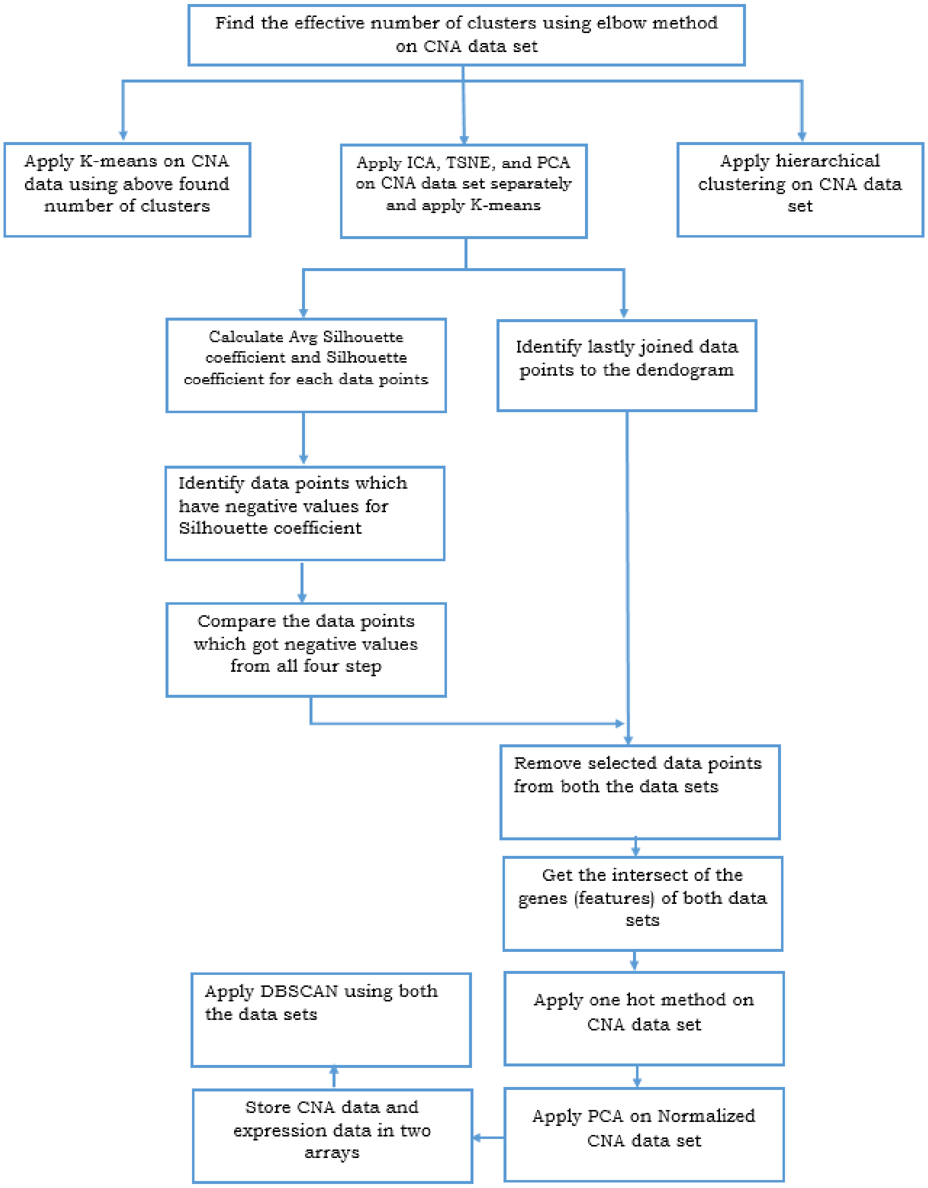

Find the effective number of clusters using elbow method on CNA data set
Apply K-means on CNA data using above found number of clusters
Apply ICA, TSNE, and PCA on CNA data set separately and apply K-means
Apply hierarchical clustering on CNA data set
Calculate Avg Silhouette coefficient and Silhouette coefficient for each data points
Identify lastly joined data points to the dendogram
Identify data points which have negative values for Silhouette coefficient
Compare the data points which got negative values from all four step
Remove selected data points from both the data sets
Get the intersect of the genes (features) of both data sets
Apply DBSCAN using both the data sets
Apply one hot method on CNA data set
Store CNA data and expression data in two arrays
Apply PCA on Normalized CNA data set

Figure 52: Approach fourteen

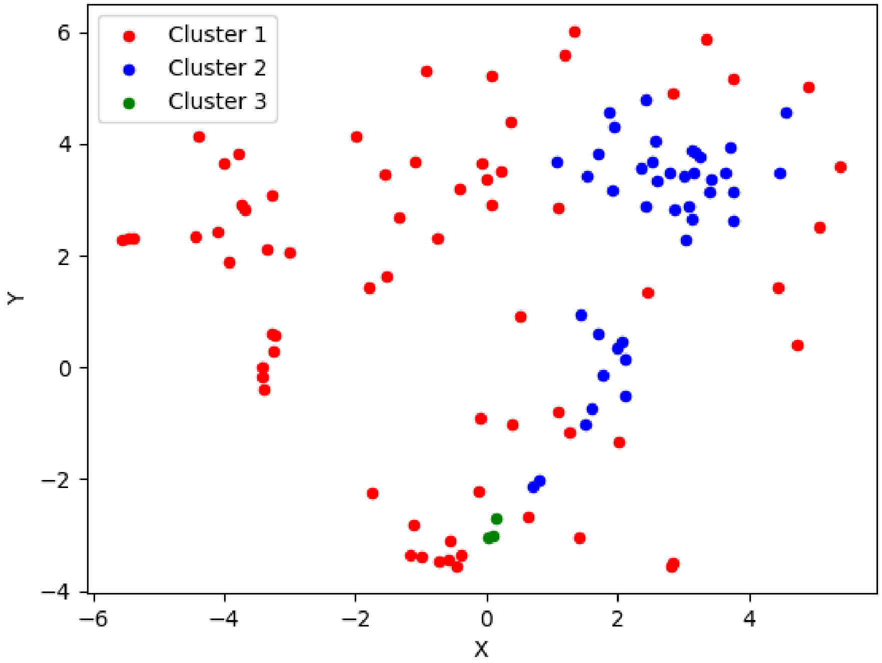


Figure 53: Clusters after approach fourteen

As the another step preprocessed CNA data set with one hot and the dense data set was filter out from the outliers which were identified in the earlier steps. Then dimension reduction was done using PCA technique in the both data sets. Then DBSCAN clustering method was carried out for both the data sets. But there were no any significant difference in the clusters when comparing to the step which followed the same above steps with the outliers. Average silhouette score was (-0.1553) same as the above mention step and the distribution of the data points in the clusters were almost same.

Chapter 6 Evaluation

6.1 Overview

In this chapter, we present the evaluation that was carried out to assess the validity of the method designed and explored to identify a better clustering method for high dimensional unlabeled genomic data. There are mainly two approaches to evaluate this hybrid clustering method.

6.2 External Evaluation

In external evaluation, clustering results are evaluated based on data that was not used for clustering, such as known class labels and external benchmarks. Such benchmarks consist of a set of pre-classified items, and these sets are often created by (expert) humans [7]. So in the evaluation process of this research it will be utilized another data set from the same data source which is again not labeled and high dimensional genomic data. This data set is utilized in order to validate the clusters that will be the result of novel hybrid clustering method.

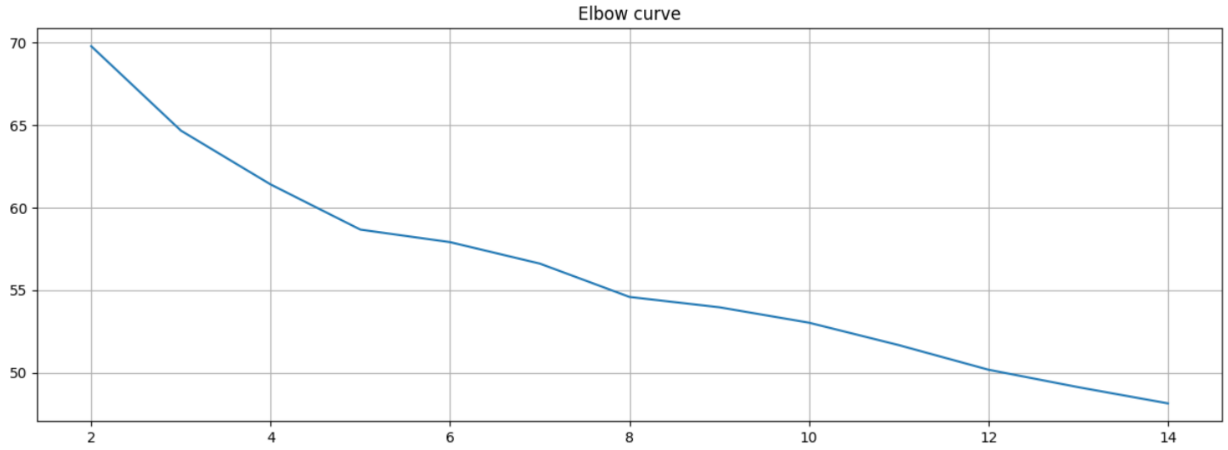


Figure 54: Elbow curve for the test data set

Selected full data set has CNA data set and expression data set separately. This CNA data set has 107 samples and expression data set has 49 samples. Both have 18609 features which are different genes. All the approaches which were carried out throughout this research were again applied on this new test data set and compared the results in order to validate this hybrid method, showing that better clusters with better scores for measurements indexes can be performed using this hybrid method on another data set which was not used for the experimental purpose.

Elbow method was used on new CNA data set in order to find the efficient number of clusters (Figure 54). Most appropriate number of clusters were five. Rest of the approaches were carried out considering that the number of clusters is five.

The results are shown in the below table with the approach carried out. As explained above 1st four methods was used to identify outliers with the dendogram which were performed by hierarchical clustering method (Figure 55). The dendogram which was performed for this data set is shown below.

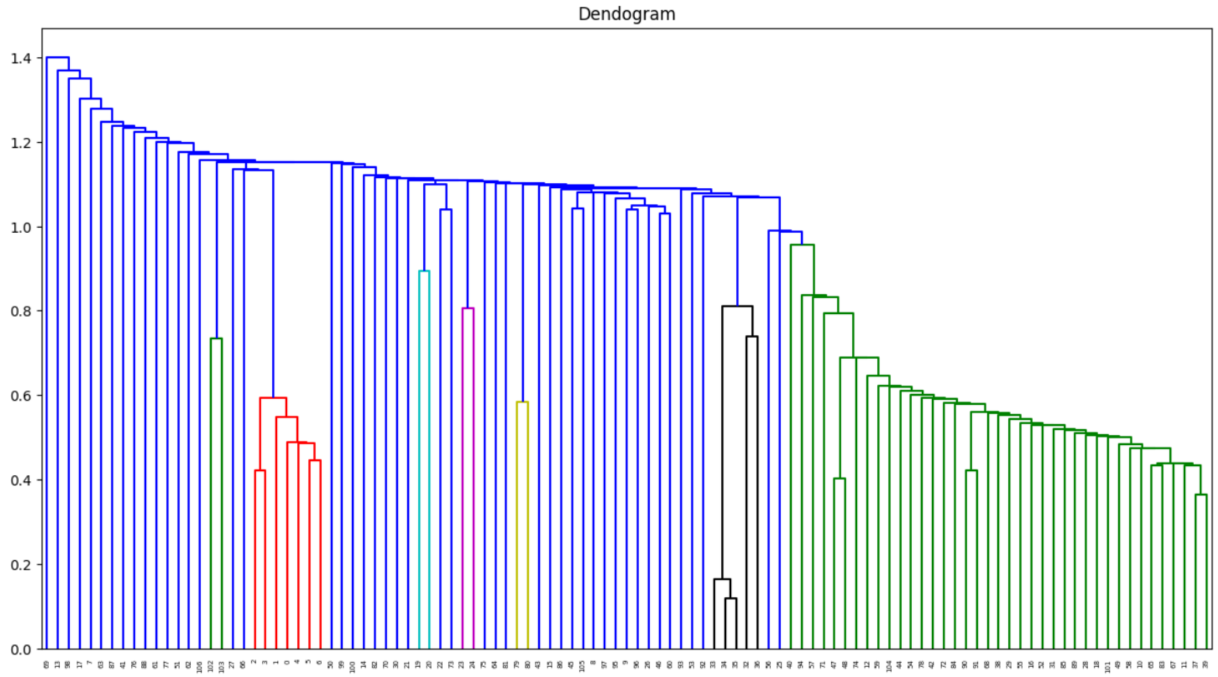
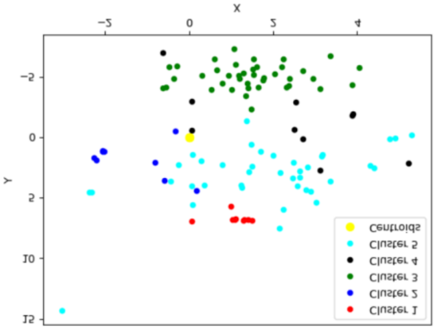
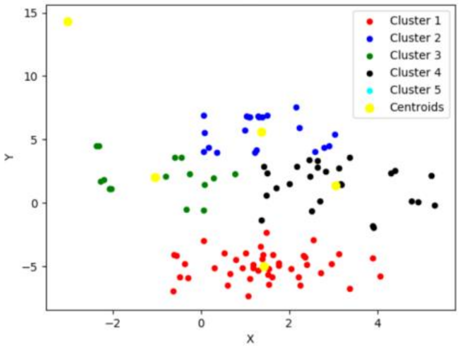
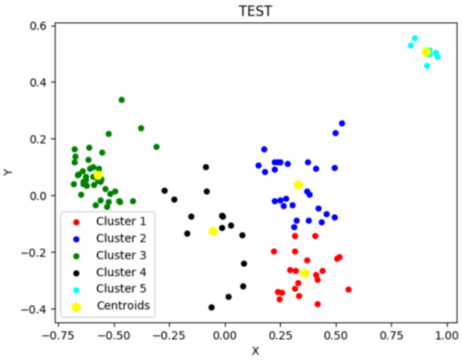
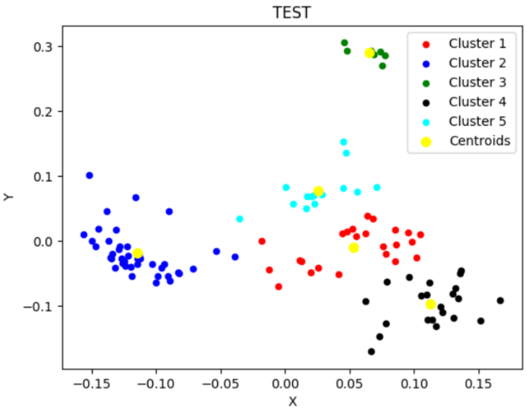
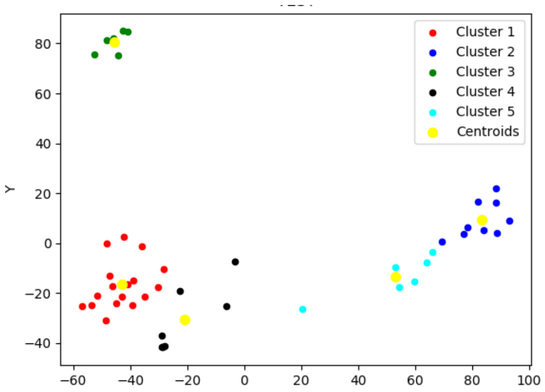
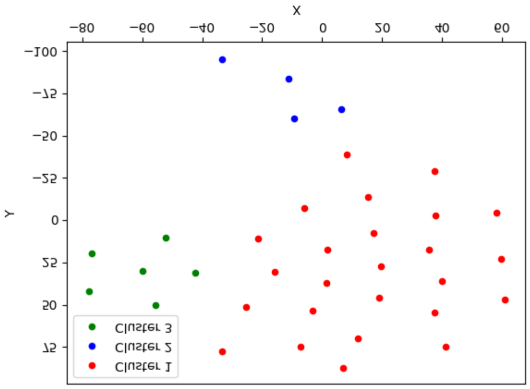


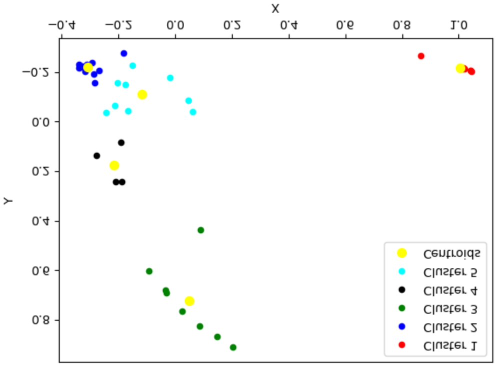
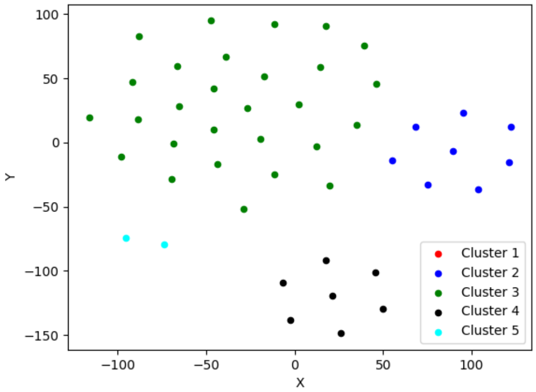
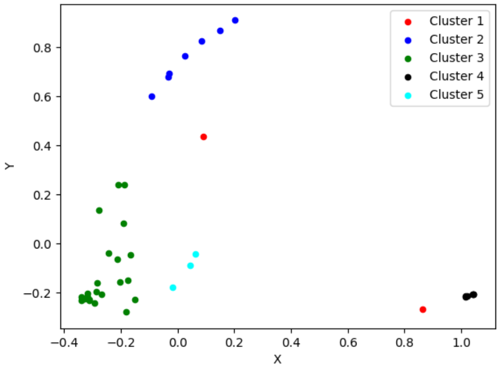
Figure 55: Dendogram for the test data set

The outliers for this data set was identified by comparing the silhouette score which was gained for each data points for above all five methods. Identified outliers are shown in the below table (Table 7).

Method Followed	Avg. Silhouette Score	Avg. Davies–Bouldin index	Davies–Bouldin index for each cluster				
	0.18	2.6	1.85	2.44	1.96	3.53	3.53

Raw CNA data with K-means clustering							
Apply TSNE on CNA data then K-means clustering	0.18	2.6	1.85	2.44	1.96	3.53	3.53
							
Apply PCA on CNA data then K-means clustering	0.16	2.6	1.85	2.44	1.96	3.53	3.53
							
	0.19	2.6	1.85	2.44	1.96	3.53	3.53

Apply ICA on CNA data then K-means clustering							
Apply one hot method on CNA data then PCA dimention reduction then K-means clustering	0.5	23.17	14.3	49.68	1.21	49.68	0.902
							
Apply one hot method on CNA data then K-means clustering	0.16	1.6	1.42	1.84	1.29	1.88	1.88
							
Apply one hot method on CNA data,	0.6	0.5	0.1	0.6	0.3	0.66	0.69

then PCA for both the data sets then K-means clustering							
Apply DBSCAN for Expression data set	0.42	0.60	1.61	0.69	0.39	0.61	0.61
							
Apply PCA on expression data then DBSCAN clustering	0.5	0.9	1.2	0.8	0.8	0.8	1.2
							
	0.2	0.77	0.23	0.89	0.84	0.78	0.62

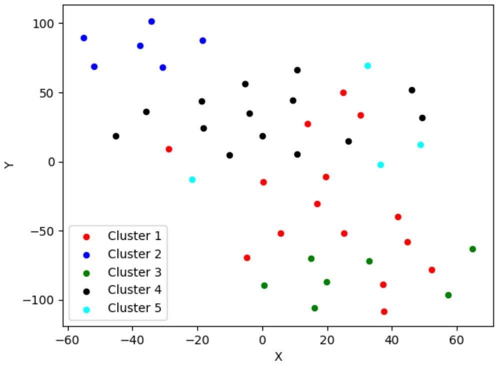
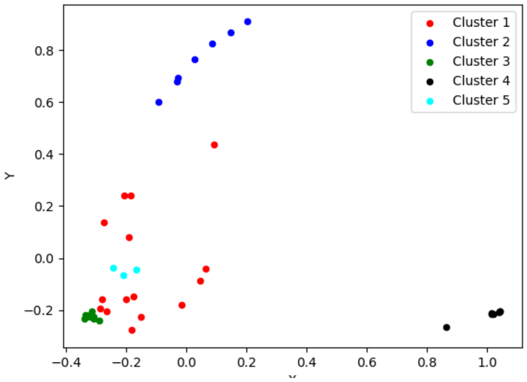
Apply one hot on CNA data then PCA for CNA data sets then DBSCAN clustering							
Remove outliers apply one hot method on CNA data and then PCA only on CNA data, then apply DBSCAN on both the data sets	0.5	0.57	1.4	0.21	0.27	0.42	0.57
							

Table 7: Comparison of evaluation measurements for test data set

According to the results of the explained procedure it is clear that presented novel approach has given the best scores for the measurement index values.

6.2 Internal Evaluation

If we are going to use clustering data itself to evaluate the result it is called as internal evaluation. Having high intra cluster similarity and low inter cluster similarity is one of the characteristics of an effective clustering methods. Hence, the clustering method that will be the final outcome of this research will be evaluated by using these measurements.

There are three main measures which are used for internal evaluation in cluster evaluation. They are,

- Davies–Bouldin index
- Dunn index

- Silhouette coefficient

These messures can be used for evaluate clusters which were performed by particular clustering method. These messurements gives a score for each cluster so that the clusters can be evaluated internally [23].

Since this reseach was an explorative research each and every time clusters was performed internal evaluation was carried out [25]. Hence the most effective clusters can be identified in the every step. Silhouette coefficient was used in every step inorder to do the internal evaluation [24]. Hence internal evaluation has been carried out thrugh out the research.

Since the result of Silhouette coefficient is used for the exploration part of the research, another messurment technique need to be used in the evaluation phase [39]. Hence Davies–Bouldin index was used to evaluate the clusters in each step and the final cluster set.

Davies–Bouldin index

Following formula can be used for calculate the Davies–Bouldin index

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

n = Number of clusters

Cx = the centroid of cluster x

σ x = the average distance of all elements in cluster x

d(Ci, cj) = the distance between centroids

A better clustering algorithm will always perform clusters with low intra cluster distance which should have high intra cluster similarity [40]. When considering inter cluster distance good cluster set may have high inter cluster distance and low inter cluster similarity. A cluster with these qualities will have a low Davies–Bouldin index, the clustering algorithm that produces a collection of clusters with the smallest Davies–Bouldin index is considered the best algorithm based on this criterion [40].

The Davies–Bouldin index values get for the each step which have been done in the methodology section to perform clusters are as follows with the corresponding value for silhouette score in the each step.

Method Followed	**Avg. Silhouette Score**	**Avg. Davies–Bouldin index**	**Davies–Bouldin index for each cluster**		
Raw CNA data with K-means clustering	0.320	1.563	1.63	1.42	1.63
Apply TSNE on CNA data then K-means clustering	0.212	1.563	1.63	1.42	1.63
Apply PCA on CNA data then K-means clustering	0.311	1.563	1.63	1.42	1.63
Apply ICA on CNA data then K-means clustering	0.307	1.563	1.65	1.42	1.63
Apply one hot method on CNA data then PCA dimention reduction then K-means clustering	0.233	1.13	1.02	1.33	1.01
Apply one hot method on CNA data then K-means clustering	-0.09	167845171	Large value	Large value	Large value
Apply one hot method on CNA data, then PCA for both the data sets then K-means clustering	0.5	0.6	0.5	0.6	0.6
Apply DBSCAN for Expression data set	0.047	1.56	1.61	1.61	1.4
Apply DBSCAN for both CNA and Expression data	0.047	1.56	1.61	1.61	1.4
Apply PCA on expression data then DBSCAN clustering	0.03	1.8	2.5	2.5	0.6
Apply one hot on CNA data then PCA for CNA data sets then DBSCAN clustering	-0.13	0.9	1.06	1.06	0.6
Apply one hot on CNA data then PCA for both the data sets then DBSCAN clustering	-0.15	4.5	6.3	6.3	0.9
Remove outliers apply DBSCAN for both data sets	-0.1553	4.53	6.34	6.34	0.92
Remove outliers apply one hot method on CNA data and then PCA only on CNA data, then apply DBSCAN on both the data sets	0.464	0.5643	0.577	0.584	0.532

Table 8: Comparison of evaluation measurements

When consider the above table (Table 8), the clustering approaches which have got nearly same value for the average silhouette score, have got nealy same values also for the average Davies–Bouldin index. So the approaches which used for perform clusters can be evaluated using these messurements. The procedure which follows to evaluate the clusters can be confirmed using the score values of the Davies–Bouldin index for each cluster.

Chapter 8 Results & Analysis

8.1 Overview

This chapter presents the overall findings of this research study and the analysis and discussion of them.

8.2 Results and analysis on each clustering approaches

There are several clustering approach which were followed in order to find a better clustering approach for high dimensional unlabeled genomic data. Here the main focus was on main three clustering algorithms which were used heavily in the literature. They are k-means clustering algorithm, Hierarchical clustering algorithm and DBSCAN clustering algorithm.

These algorithm was applied in different ways on two main data sets which are CNA data set and expression data set using different preprocessing methods. The clusters which were produced using these different methods were evaluated using different measurements like silhouette score and Davies–Bouldin index.

This analysis will be carried out using these meassurements and analysing the clusters comparing the data points in each clusters which were perormed using different methods.

Approches with K-means clustering

When consider the values for the measurement indexes in the t able 8 it shows that there is no any considerable impact when using dimension reduction methods such as ICA, TNSE and PCA on CNA data and the apply K-means clustering compared to the approach where K-means clustering apply directly on the CNA data set.

Method Followed	**Avg. Silhouette Score**	**Avg. Davies–Bouldin index**	**Davies–Bouldin index for each cluster**		
Row CNA data with K-means clustering	0.320	1.563	1.63	1.42	1.63
Apply TSNE on CNA data then K-means clustering	0.212	1.563	1.63	1.42	1.63
Apply PCA on CNA data then K-means clustering	0.311	1.563	1.63	1.42	1.63
Apply ICA on CNA data then K-means clustering	0.307	1.563	1.65	1.42	1.63

Table 9: Comparison of evaluation measurements

There are small differences of the average silhouette score but average Davies–Bouldin index is almost same for all the approaches (Table 9).

Converting the categorical data which is in CNA data set to numerical data using one hot method and then apply K-means clustering without using dimention reduction is the most failed approach among the approaches which were carried out through this reseach. The meassurement indexes have got very large values for this clustering approach (Table 10).

Method Followed	**Avg. Silhouette Score**	**Avg. Davies–Bouldin index**	**Davies–Bouldin index for each cluster**		
Apply one hot method on CNA data then K-means clustering	-0.09	167845171	Large value	Large value	Large value

Table 10: Comparison of evaluation measurements

Use one hot method on the CNA data set and then apply PCA method to dimension reduction and apply k-means clustering on these preprocessed data shows considerable improvement when consider the values for the measurement indexes for that method (Table 11).

Method Followed	**Avg. Silhouette Score**	**Avg. Davies–Bouldin index**	**Davies–Bouldin index for each cluster**		
Apply one hot method on CNA data then PCA dimention reduction then K-means clustering	0.233	1.13	1.02	1.33	1.01

Table 11: Comparison of evaluation measurements

Using one hot for CNA data and then apply PCA for both the data set which are preprocessed CNA and expression data set and apply k-means for perform clusters has the best score among the approaches which uses k-means as the main clustering algorithm (Table 12).

Method Followed	**Avg. Silhouette Score**	**Avg. Davies–Bouldin index**	**Davies–Bouldin index for each cluster**		
Apply one hot method on CNA data, then PCA for both the data sets then K-means clustering	0.5	0.6	0.5	0.6	0.6

Table 12: Comparison of evaluation measurements

Approches with DBSCAN clustering

There are several approaches carried out using DBSCAN clustering method on both the data sets, separately and as a combined data set. The measurement indexes were calculated for these clustering approaches as well.

When considering these measurement indexes it shows that there's no any considerable impact when clustering only expression data using DBSCAN method compared to clustering the data set using both CNA and expression data without using any conversion method such as one hot method or any dimension reduction method such as PCA, on any of the data set (Table 13).

Method Followed	Avg. Silhouette Score	Avg. Davies–Bouldin index	Davies–Bouldin index for each cluster		
Apply DBSCAN for Expression data set	0.047	1.56	1.61	1.61	1.4
Apply DBSCAN for both CNA and Expression data	0.047	1.56	1.61	1.61	1.4

Table 13: Comparison of evaluation measurements

When analyzing the other approach which were carried out with DBSCAN clustering method. There are three approaches which got almost the same results. They are (Table 14),

- Apply DBSCAN clustering only on the CNA data which were converted from categorical data to numerical data using one hot method and apply PCA for dimension reduction.
- Apply DBSCAN clustering on both the CNA data which were converted from categorical data to numerical data using one hot method and expression data set after applying PCA for dimension reduction.
- Remove outliers from both the data sets and then use DBSCAN for both the data sets without using any preprocessing methods.

Method Followed	Avg. Silhouette Score	Avg. Davies–Bouldin index	Davies–Bouldin index for each cluster		
Apply one hot on CNA data then PCA for CNA data sets then DBSCAN clustering	-0.13	0.9	1.06	1.06	0.6
Apply one hot on CNA data then PCA for both the data sets then DBSCAN clustering	-0.15	4.5	6.3	6.3	0.9

Remove outliers apply DBSCAN for both data sets	-0.1553	4.53	6.34	6.34	0.92

Table 14: Comparison of evaluation measurements

Last approach with the DBSCAN clustering method got the highest values for all the measurement indexes. In this approach CNA data set was clustered using k-means clustering in three steps,

- Without any dimension reduction
- With TNSE dimension reduction
- With ICA dimension reduction
- With PCA dimension reduction

Then the CNA data set was clustered using hierarchical clustering method and drew the dendogram so that how and in which order the data points get joined with the clusters will be cleared.

By analyzing above four methods outliers were identified. Then outliers were removed from both CNA and expression data sets. Since CNA data has categorical data, this data set was converted to numerical data using one hot method. But any preprocesses method was not carried out on the expression data set. Then DBSCAN clustering was applied on both the data sets. The measurement values for this method is as follows (Table 15),

Method Followed	**Avg. Silhouette Score**	**Avg. Davies–Bouldin index**	**Davies–Bouldin index for each cluster**		
Remove outliers apply one hot method on CNA data and then PCA only on CNA data, then apply DBSCAN on both the data sets	0.464	0.5643	0.577	0.584	0.532

Table 15: Comparison of evaluation measurements

As the next phase of the analysis, best cluster of each approach will be compare with the corresponding cluster which were performed by the found hybrid method. Apply one hot method on

CNA data, then PCA for both the data sets then applying K-means clustering approach has the lowest which is the best Davies–Bouldin index for the 1st and 2nd clusters. Those two clusters were compared with the 1st and 2nd cluster of the final results which got form the novel hybrid method.

The 1st cluster of above method got 41 data points, and the 1st cluster of the hybrid method got 32 data points, out of those 32 data points 28 data points are also include in the 1st cluster which were performed by above mentioned method.

Also in the 2nd cluster of above method has 67 data points, and the 2nd cluster of the hybrid method has 60 points, out of that 60 points 52 points are also include in the 1st cluster of above metioned method.

When considering about the 3rd cluster, Apply PCA on expression data then Apply DBSCAN method and Apply one hot method on CNA data then apply PCA on both the data sets and apply DBSCAN clustering method have the best Davies–Bouldin index for 3rd cluster. Hence these two clusters were compared with the 3rd cluster which were performed by the hybrid method.

In the above mentioned two clusters there are 3 data point in one cluster and 5 data points in the other clusters respectively. The corresponding cluster which performed by the novel method has 3 data points. Three out of these three data points are include in the cluster which performed by the above mentioned 2nd method. Two out of three data points are also include in the cluster which was performed by the above mentioned 1st method.

Chapter 9 Conclusion

9.1 Overview

As is the case with any research, there are several limitations inherent in this study which were unavoidable. This chapter presents a summary of the research, limitations and following that, the recommendations for future work are provided.

9.2 Summary

This paper presents a novel approach which we propose for the clustering high dimensional genomic data. It includes a method which can be considere as a hybrid method as it is an approach which includes some of the known clustering methods, K-means, hierarchicla and DBSCAN. Several approches were carried out using different steps of each in order to perform a better cluster set. Internal evaluation phase carried out in each approach to identify better clusters in order to proceed futher.

The summary details of the evaluation messurements such as silhouette score and Davies–Bouldin index were used to assess the novel approach. Internal evaluation as well as the external evaluation shows that the novel approach has a better results when comparing to the other approaches which were carried out through out the research. Overall performance measures of the novel clustering approach as well as the similarity measure of the each cluster seperatly were carried out to assess the quality of the novel method. The values achieved as discussed previously proved that this method can be used in performing better clusters of high dimensional genomic data. Hence we can state that our novel method is successful in performing clusters with better similarities among the data points.

9.3 Limitations

Due to time constraint we are unable to evaluate each clusters we gain through different method using bio medical knowledge of a domain expert. Since we have tried out nearly sixteen different method of clustering approaches. Each approach we got three clusters, all together there are 48 clusters to be evaluated considering how they are related in bio medical domain. Since this is a time consuming task we decided to evaluate these clusters using computational methods such as inter cluster similarity and intra cluster similarity. While doing the research in order to decide the next step we used silhouette score as one of such methods.

No labels are attached to the performed clusters as they are not thoroughly evaluated using biomedical knowledge is another drawback of this research. Even though we performed clusters with high values for the measurement indexes there is no real meaning of those clusters when considering the biological aspect.

This method is built using main two data sets which is CNA data and expression data. The genomic data which is not under these categories are not tested with this method. CNA data set is used for the 1st phase of the method, for the 2nd phase both the data sets were used.

Another limitation that we noticed is, when preparing the data set we got intersect of the features of both the data sets. Hence we had to remove some of the genes from the expression data set. There might be some important data is been removed because of this step.

Due to time limitation we have used only three main clustering algorithms for this research which were frequently used in literature. But there can be some different clustering algorithms which will give better outputs when combining them as did in this research.

This method has only tested with genomic data hence we can’t recommend this method for any other data categories. For the methodology of the research we used one genomic data set and for the evaluation phase of the research used another genomic data set. Hence we can’t ensure that this method will work for any other type of data.

9.4 Suggestions for Future Work

While this thesis has presented the potential of clustering the unlabeled genomic data and can be lead to identify some important information on those data, many opportunities for extending the scope of this thesis remain. Hence, in this section we are suggesting some future research directions that could improve the mutation recognition and analysis process.

The most important next step will be evaluate the clustering results using bio medical knowledge of a domain expert and tune the method. As we are focusing on genomic data it is important to evaluate these clusters using bio medical knowledge so that we can decide the correctness of the clusters. Using those results we can tune the method more to give more accurate clusters.

We can test this method on data which is not genomic data. Because we have used this method only on genomic data with two data sets. We can try this method on some other data set and evaluate the output so that we can generalize this method to be used on any data set.

When preprocessing the data sets we had to remove some of the genes which we considered as the features of the data set as CNA and the expression data set does not contain the same gene set. When removing some of these data, some of the important data might have been removed. Tune this method to use the data set as it is without removing any of them can be considered as another future work.

To explore this method we have only used three main clustering methods, K-means, hierarchical and DBSCAN. As a future work we can expand the number of selected clustering methods. We can try out different clustering method and combine the results with this method to tune the method in order to achieve better outputs.

List of References:

[1]C. Sotiriou, S. Neo, L. McShane, E. Korn, P. Long, A. Jazaeri, P. Martiat, S. Fox, A. Harris and E. Liu, "Breast cancer classification and prognosis based on gene expression profiles from a population-based study", *Proceedings of the National Academy of Sciences*, vol. 100, no. 18, pp. 10393-10398, 2003.

[2]T. Golub, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", *Science*, vol. 286, no. 5439, pp. 531-537, 1999.

[3]B. Feldman, "Genomics and the role of big data in personalizing the healthcare experience", *O'Reilly Media*, 2018. [Online]. Available: https://www.oreilly.com/ideas/genomics-and-the-role-of-big-data-in-personalizing-the-healthcare-experience. [Accessed: 09- Jul- 2018].

[4]J. Lee, P. Williams and S. Cheon, "Data Mining in Genomics", 2018.

[5]T. Manolio, "Genomewide Association Studies and Assessment of the Risk of Disease | NEJM", *New England Journal of Medicine*, 2018. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMra0905980. [Accessed: 09- Jul- 2018].

[6]E. Koonin and M. Galperin, "Information Sources for Genomics", *Ncbi.nlm.nih.gov*, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK20256/. [Accessed: 09- Jul- 2018].

[7]A. Thalamuthu, I. Mukhopadhyay, X. Zheng and G. Tseng, "Evaluation and comparison of gene clustering methods in microarray analysis", *Bioinformatics*, vol. 22, no. 19, pp. 2405-2412, 2006.

[8]T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lønning, P. O. B. A.-L. Børresen-Dale, and D. Botstein, “Repeated observation of breast tumor subtypes in independent gene expression data sets,” PNAS, 08-Jul-2003. [Online]. Available: http://www.pnas.org/content/100/14/8418.full. [Accessed: 09-Jul-2018].

[9]“FAQ,” cBioPortal for Cancer Genomics. [Online]. Available: http://www.cbioportal.org/faq.jsp. [Accessed: 09-Jul-2018].

[10] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer Statistics, 2017.,” Advances in pediatrics., Jan-2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28055103. [Accessed: 09-Jul-2018].

[11] [7]N. Zhang, "Hierarchical Latent Class Models for Cluster Analysis", *Jmlr.org*, 2018. [Online]. Available: http://www.jmlr.org/papers/volume5/zhang04a/zhang04a.pdf. [Accessed: 09- Jul- 2018].

[12]R. Fakoor, F. Ladhak, A. Nazi and M. Huber, "Using deep learning to enhance cancer diagnosis and classification", 2017.

[13] Genetics Home Reference, 'What is DNA?', 2015. [Online]. Available: http://ghr.nlm.nih.gov/handbook/basics/dna.[Accessed: 15- April-2018].

[14] Genomenewsnetwork.org, 'What's a Genome?', 2015. [Online]. Available: http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp2_2.shtml. [Accessed: 15-April- 2018].

[15] Udel.edu, 'How does Sanger sequencing work?', 2015. [Online]. Available: http://www.udel.edu/dnasequence/Site/How_does_DNA_sequencing_work.html. [Accessed: 18- April- 2018].

[16] E. Help, 'Entrez Programming Utilities Help', National Center for Biotechnology Information (US), 2010.

[17] Genome.gov, 'Genetic Variation Program', 2015. [Online]. Available: http://www.genome.gov/10001551.[Accessed: 18- April-2018].

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

[19] 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.

[20] Anil K Jain, “Data clustering: 50 years beyond Kmeans”,Journal Pattern Recognition Letters, Volume no.- 31, Issue no.- 8,pp- 651-666, 2010.

[21] R. Karim, A. Zappa, R. Sahay and D. Rebholz-Schuhmann, "A Deep Learning Approach to Genomics Data for Population Scale Clustering and Ethnicity Prediction", *Insight-centre.org*, 2018. [Online]. Available: https://www.insight-centre.org/sites/default/files/publications/lncs_final.pdf. [Accessed: 27- Oct- 2018].

[22] D. Arthur and S. Vassilvitskii. k-means++ the advantages of careful seeding. In Symposium on Discrete Algorithms, 2007.

[23] J. M. Neuhaus and J. D. Kalbfleisch. Between- and within-cluster covariate effects in the analysis of clustered data. Biometrics, 54(2):638–645, Jun. 1998.

[24] Kuncheva, Ludmila I., Hadjitodorov, Stefan T.: Using Diversity in Cluster Ensembles. IEEE SMC International Conference on Systems, Man and Cybernetics, 2004.

[25] Meila, Marina: Comparing Clusterings. COLT 2003.

[26] Strehl, Alexander, Ghosh, Joydeep: Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. Journal of Machine Learning Research, 3:583–617, 2002.

[27] Li, Tao, Ogihara, Mitsunori, Ma, Sheng: On Combining Multiple Clusterings. Proceedings of the ACMConference on Information and Knowledge Management, (13):294-303, 2004.

[28]Davies, D. L., and Bouldin, D. W. 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 2: 224–27.

[29]H. B. Zhou and J. T. Gao, "Automatic Method for Determining Cluster Number Based on Silhouette Coefficient", Advanced Materials Research, Vol. 951, pp. 227-230, 2014

[30]Agrawal, Rakesh, Gehrke, Johannes, Gunopulos, Dimitrios, Raghavan, Prabhakar, 1998. Automatic subspace clustering of high dimensional data for data mining applications. In: Proc. ACM SIGMOD, pp. 94–105.

[31] Ball, G., Hall, D., 1965. ISODATA, a novel method of data anlysis and pattern classification. Technical report NTIS AD 699616. Stanford Research Institute, Stanford, CA.

[32] Bradley, P.S., Fayyad, U., Reina, C., 1998. Scaling clustering algorithms to large databases. In: Proc. 4th KDD.

[33] Cheng, Yizong, Church, George M., 2000. Biclustering of expression data. In: Proc. Eighth Internat. Conf. on Intelligent Systems for Molecular Biology, AAAI Press, pp. 93–103.

[34] Charrad M., Ghazzali N., Boiteau V., Niknafs A. (2014). NbClust: An R Packagefor Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 61(6), 1-36.

[35] A. Kassambara and A. Kassambara, Practical guide to cluster analysis in R. United States: STHDA, 2017.

[36] V. Kotte, S. Rajavelu and E. Rajsingh, "A Similarity Function for Feature Pattern Clustering and High Dimensional Text Document Classification", Foundations of Science, 2019. Available: 10.1007/s10699-019-09592-w.

[37]A. Jain, "Data clustering: 50 years beyond K-means", Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666, 2010. Available: 10.1016/j.patrec.2009.09.011.

[38]A. Torrente, M. Kapushesky and A. Brazma, "A new algorithm for comparing and visualizing relationships between hierarchical and flat gene expression data clusterings", Bioinformatics, vol. 21, no. 21, pp. 3993-3999, 2005. Available: 10.1093/bioinformatics/bti644.

[39] Bolshakova N. and Azuaje F., ”Cluster Validation Techniques for Genome Expression Data”,
Signal Processing, 83, 2003, pp. 825-833.

[40] G¨unter S. and Bunke H., ”Validation Indices for Graph Clustering”, J. Jolion, W. Kropatsch,
M. Vento (Eds.) Proceedings of the 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, CUEN Ed., Italy, 2001, pp. 229-238.

Appendix A: Clusters when apply K-means on CNA data set

0	1	2
'1115244'	'1115156'	'1115153'
'6115121'	'6115115'	'1115154'
'6115227'	'6115117’	'1115157'
'6115250'	'6115118'	'1115161'
'MO_1114'	'6115122'	'1115183'
'MO_1179'	'6115123'	'1115202'
'MO_1184'	'6115224'	'6115114'
'MO_1219'	'6115233'	'6115219'
'SC_9007'	'6115234'	'6115237'
'SC_9017'	'6115247'	'6115242'
'SC_9022'	'6115251'	'MO_1013'
'SC_9038'	'MO_1020'	'MO_1014'
'SC_9057'	'MO_1040'	'MO_1071'
'SC_9059'	'MO_1054'	'MO_1095'
'SC_9063'	'MO_1084'	'MO_1128'
'SC_9086'	'MO_1094'	'MO_1161'
'SC_9099'	'MO_1118'	'MO_1202'
'TP_2054'	'MO_1124'	'MO_1221'
'TP_2061'	'MO_1176'	'MO_1232'
	'MO_1192'	'MO_1249'
	'MO_1215'	'SC_9008'
	'MO_1241'	'SC_9009’
	'MO_1244'	'SC_9010'
	'MO_1262'	'SC_9016'
	'MO_1277'	'SC_9018'
	'MO_1316'	'SC_9023'
	'MO_1336'	'SC_9028'
	'MO_1337'	'SC_9030'
	'MO_1339'	'SC_9034'
	'SC_9001'	'SC_9037'
	'SC_9012'	'SC_9047'
	'SC_9019'	'SC_9049'
	'SC_9026'	'SC_9060'
	'SC_9029'	'SC_9061'
	'SC_9031'	'SC_9062'
	'SC_9032'	'SC_9068'
	'SC_9036'	'SC_9071'

	'SC_9043'	'SC_9073'
	'SC_9046'	'SC_9083'
	'SC_9048'	'SC_9091'
	'SC_9050'	'TP_2009'
	'SC_9054'	'TP_2060'
	'SC_9055'	
	'SC_9058'	
	'SC_9072'	
	'SC_9080'	
	'SC_9081'	
	'SC_9092'	
	'SC_9093'	
	'SC_9094'	
	'SC_9097'	
	'TP_2001'	
	'TP_2010'	
	'TP_2020'	
	'TP_2032'	
	'TP_2034'	
	'TP_2064'	

Appendix B: Clusters when apply K-means on CNA data set with TNSE dimensional reduction

0	1	2
'1115154'	'6115115'	'1115153'
'1115157'	'6115117’	'1115156'
'1115161'	'6115122'	'1115183'
'1115244'	'6115224'	'1115202'
'6115118'	'6115233'	'6115114'
'6115219'	'6115234'	'6115121'
'6115227'	'6115237'	'6115123'
'6115242'	'6115247'	'MO_1013'
'6115250'	'MO_1020'	'MO_1095'
'6115251'	'MO_1054'	'MO_1161'
'MO_1014'	‘MO_1071’	'MO_1174'
'MO_1040'	'MO_1084'	'MO_1249'
'MO_1114'	'MO_1094'	'MO_1262'
'MO_1128'	'MO_1118'	'SC_9007'
'MO_1179'	'MO_1124'	'SC_9009’
'MO_1184'	'MO_1192'	'SC_9016'
'MO_1202'	'MO_1215'	'SC_9018'
'MO_1219'	'MO_1241'	'SC_9028'
'MO_1221'	'MO_1244'	'SC_9034'
'MO_1231'	'MO_1277'	'SC_9036'
'MO_1339'	'MO_1316'	'SC_9037'
'SC_9008'	'MO_1336'	'SC_9046'
'SC_9010'	'MO_1337'	'SC_9047'
'SC_9017'	'SC_9001'	'SC_9054'
'SC_9022'	'SC_9012'	'SC_9060'
'SC_9023'	'SC_9019'	'SC_9062'
'SC_9030'	'SC_9026'	'SC_9063'
'SC_9038'	'SC_9027'	'SC_9068'
'SC_9049'	'SC_9031'	'SC_9071'
'SC_9055'	'SC_9032'	'SC_9072'
'SC_9057'	'SC_9043'	'SC_9073'
'SC_9059'	'SC_9048'	'SC_9080'
'SC_9061'	'SC_9050'	'SC_9097'
'SC_9072'	'SC_9058'	'TP_2037'
'SC_9083'	'SC_9081'	'TP_2064'

'SC_9086'	'SC_9091'	
'SC_9093'	'TP_2001'	
‘SC_9094’	'TP_2009'	
'SC_9099'	'TP_2010'	
'TP_2054'	'TP_2020'	
	'TP_2032'	
	'TP_2060'	
	'TP_2061'	

Appendix C: Clusters when apply K-means on CNA data set with PCA dimensional reduction

0	**1**	**2**
'1115154'	'6115117’	'1115153'
'1115157'	'6115118'	'1115154'
'1115161'	'6115224'	'1115156'
'1115244'	'6115233'	'1115183'
'6115121'	'6115234'	'1115202'
'6115219'	'6115247'	'6115114'
'6115227'	'6115251'	'6115115'
'6115250'	'MO_1020'	'6115122'
'MO_1013'	'MO_1040'	'6115123'
'MO_1114'	'MO_1084'	'6115237'
'MO_1179'	'MO_1094'	'MO_1014'
'MO_1184'	'MO_1118'	'MO_1054'
'MO_1202'	'MO_1124'	'MO_1071'
'MO_1219'	'MO_1192'	'MO_1095'
'MO_1221'	'MO_1215'	'MO_1128'
'MO_1232'	'MO_1244'	'MO_1161'
'MO_1249'	'MO_1262'	'MO_1174'
'MO_1339'	'MO_1277'	'MO_1241'
'SC_9007'	'MO_1316'	'SC_9008'
'SC_9010'	'MO_1336'	'SC_9009’
'SC_9016'	'MO_1337'	'SC_9018'
'SC_9017'	'SC_9001'	'SC_9023'
'SC_9022'	'SC_9012'	'SC_9023'
'SC_9023'	'SC_9019'	'SC_9026'
'SC_9034'	'SC_9027'	'SC_9028'
'SC_9038'	'SC_9031'	'SC_9036'
'SC_9047'	'SC_9032'	'SC_9037'
'SC_9055'	'SC_9043'	'SC_9048'
'SC_9057'	'SC_9046'	'SC_9049'
'SC_9059'	'SC_9050'	'SC_9061'
'SC_9060'	'SC_9054'	'SC_9062'
'SC_9063'	'SC_9058'	'SC_9068'
'SC_9071'	'SC_9080'	'SC_9083'
'SC_9072'	'SC_9081'	'SC_9091'

'SC_9073'	'SC_9092'	'TP_2009'
'SC_9086'	'SC_9093'	'TP_2020'
'SC_9094'	'SC_9097'	'TP_2034'
'SC_9099'	'TP_2001'	'TP_2060'
'TP_2054'	'TP_2010'	
'TP_2061'	'TP_2032'	
	'TP_2064'	

Appendix D: Silhouette Score for each point

Data point	Score
'MO_1013'	4.38E-03
'MO_1014'	-4.39E-03
'MO_1020'	1.86E-02
'MO_1040'	1.67E-03
'MO_1054'	3.55E-02
'MO_1071'	3.83E-02
'MO_1084'	4.43E-02
'MO_1094'	6.61E-02
'MO_1095'	1.88E-02
'MO_1114'	3.38E-02
'MO_1118'	2.43E-02
'MO_1124'	1.32E-02
'MO_1128'	2.50E-02
'MO_1161'	2.33E-02
'MO_1176'	9.45E-03
'MO_1179'	2.88E-02
'MO_1184'	3.46E-02
'MO_1192'	6.46E-02
'MO_1202'	4.49E-03
'MO_1215'	3.35E-02
'MO_1219'	9.01E-02
'MO_1221'	3.93E-04
'MO_1232'	4.11E-03
'MO_1241'	9.49E-03
'MO_1244'	2.01E-02
'MO_1249'	8.45E-03
'MO_1262'	9.33E-03
'MO_1277'	4.13E-02
'MO_1316'	7.02E-02
'MO_1336'	2.08E-02
'MO_1337'	5.34E-02
'MO_1339'	4.57E-03
'1115161'	1.51E-02
'1115183'	3.13E-02
'1115202'	3.14E-02
'1115244'	3.75E-02
'6115219'	5.30E-04

'6115224'	3.09E-02
'6115227'	4.07E-02
'6115233'	2.81E-02
'6115234'	5.20E-03
'6115237'	8.09E-03
'6115242'	1.65E-02
'6115247'	7.03E-03
'6115250'	2.94E-02
'6115251'	7.08E-03
'1115153'	1.23E-02
'1115154'	-1.43E-03
'1115156'	1.37E-02
'1115157'	-4.69E-03
'6115114'	2.68E-02
'6115115'	1.39E-02
'6115117'	2.70E-02
'6115118'	-5.84E-04
'6115121'	2.30E-02
'6115122'	7.52E-03
'6115123'	1.47E-02
'SC_9001'	8.89E-02
'SC_9007'	7.35E-03
'SC_9008'	3.07E-02
'SC_9009'	3.51E-02
'SC_9010'	4.69E-03
'SC_9012'	3.83E-02
'SC_9016'	5.80E-03
'SC_9017'	6.59E-02
'SC_9018'	8.31E-03
'SC_9019'	2.79E-02
'SC_9022'	3.61E-02
'SC_9023'	1.38E-02
'SC_9026'	3.47E-02
'SC_9028'	6.76822255e-02 2
'SC_9029'	9.93E-03
'SC_9030'	3.29E-02
'SC_9031'	4.00E-02
'SC_9032'	6.27E-03
'SC_9034'	-6.54E-03
'SC_9036'	3.97E-02
'SC_9037'	3.67E-02

'SC_9038'	4.20E-02
'SC_9043'	1.63E-02
'SC_9046'	1.66E-02
'SC_9047'	1.10E-02
'SC_9048'	9.85E-03
'SC_9049'	2.83E-02
'SC_9050'	3.99E-02
'SC_9054'	4.00E-04
'SC_9055'	1.64E-02
'SC_9057'	1.17E-02
'SC_9058'	4.02E-02
'SC_9059'	4.99E-02
'SC_9060'	7.69E-04
'SC_9061'	1.00E-02
'SC_9062'	2.45E-02
'SC_9063'	1.52E-03
'SC_9068'	2.19E-02
'SC_9071'	1.89E-02
'SC_9072'	1.94E-03
'SC_9073'	2.31E-02
'SC_9080'	7.13E-02
'SC_9081'	1.52E-02
'SC_9083'	1.62E-03
'SC_9086'	6.38E-03
'SC_9091'	2.59E-02
'SC_9092'	-4.02E-03
'SC_9093'	5.43E-05
'SC_9094'	3.04E-02
'SC_9097'	-5.93E-03
'SC_9099'	6.49E-02
'TP_2001'	4.73E-03
'TP_2009'	3.29E-02
'TP_2010'	1.45E-02
'TP_2020'	2.16E-02
'TP_2032'	1.80E-02
'TP_2034'	1.69E-02
'TP_2054'	2.63E-02
'TP_2060'	7.41E-02
'TP_2061'	9.25E-03
'TP_2064'	2.68E-04

Appendix E: Clusters when apply K-means on CNA data set with ICA dimensional reduction

0	1	2
‘6115242’	'6115117’	‘1115157’
'1115161'	'6115118'	'1115153'
'1115244'	'6115224'	'1115154'
'6115121'	'6115233'	'1115156'
'6115219'	'6115234'	'1115183'
'6115227'	'6115247'	'1115202'
'6115250'	'6115251'	'6115114'
'MO_1013'	'MO_1020'	'6115115'
'MO_1114'	'MO_1040'	'6115122'
'MO_1179'	'MO_1084'	'6115123'
'MO_1184'	'MO_1094'	'6115237'
'MO_1202'	'MO_1118'	'MO_1014'
'MO_1219'	'MO_1124'	'MO_1054'
'MO_1232'	'MO_1192'	'MO_1071'
'MO_1249'	'MO_1215'	'MO_1095'
'MO_1339'	'MO_1244'	'MO_1128'
'SC_9007'	'MO_1262'	'MO_1161'
'SC_9010'	'MO_1277'	'MO_1174'
'SC_9016'	'MO_1316'	'MO_1221'
'SC_9017'	'MO_1336'	'MO_1241'
'SC_9022'	'MO_1337'	'SC_9008'
'SC_9023'	'SC_9001'	'SC_9009’
'SC_9034'	'SC_9012'	'SC_9018'
'SC_9038'	'SC_9019'	'SC_9026'
'SC_9047'	'SC_9029'	'SC_9028'
'SC_9055'	'SC_9031'	'SC_9030'
'SC_9057'	'SC_9032'	'SC_9036'
'SC_9059'	'SC_9043'	'SC_9037'
'SC_9060'	'SC_9046'	'SC_9048'
'SC_9063'	'SC_9050'	'SC_9049'
'SC_9071'	'SC_9054'	'SC_9061'
'SC_9072'	'SC_9058'	'SC_9062'
'SC_9073'	'SC_9080'	'SC_9068'
'SC_9086'	'SC_9081'	'SC_9083'

'SC_9094'	'SC_9092'	'SC_9091'
'SC_9099'	'SC_9093'	'SC_9097'
'TP_2054'	'TP_2001'	'TP_2009'
'TP_2061'	'TP_2010'	'TP_2020'
	'TP_2032'	'TP_2034'
	'TP_2064'	'TP_2060'

Appendix F: Clusters performed by approach three

0	1	2
1115153	1115154	6115117
1115156	1115157	6115123
1115161	1115183	6115224
1115244	1115202	6115233
6115114	6115115	6115234
6115227	6115118	MO_1020
6115247	6115121	MO_1084
6115250	6115122	MO_1094
6115251	6115219	MO_1124
MO_1040	6115237	MO_1176
MO_1054	6115242	MO_1192
MO_1071	MO_1013	MO_1215
MO_1095	MO_1014	MO_1244
MO_1118	MO_1114	MO_1277
MO_1128	MO_1161	MO_1316
MO_1179	MO_1202	MO_1336
MO_1184	MO_1221	MO_1337
MO_1219	MO_1232	SC_9001
MO_1262	MO_1241	SC_9022
SC_9007	MO_1249	SC_9029
SC_9008	MO_1339	SC_9031
SC_9012	SC_9009	SC_9032
SC_9018	SC_9010	SC_9036
SC_9019	SC_9016	SC_9043
SC_9028	SC_9017	SC_9050
SC_9034	SC_9023	SC_9060
SC_9037	SC_9026	SC_9061
SC_9046	SC_9030	SC_9080
SC_9047	SC_9038	SC_9086
SC_9048	SC_9054	SC_9092
SC_9049	SC_9055	TP_2001
SC_9057	SC_9058	TP_2009
SC_9063	SC_9059	TP_2010
SC_9071	SC_9062	TP_2034
SC_9073	SC_9068	
SC_9081	SC_9072	
SC_9091	SC_9083	
TP_2020	SC_9093	

TP_2032	SC_9094	
TP_2054	SC_9097	
TP_2060	SC_9099	
	TP_2061	
	TP_2064	

Appendix G: Clusters performed by approach four

0	1	2
1115153	1115156	1115161
1115154	6115115	1115202
1115157	6115121	1115244
1115183	6115123	6115117
6115114	6115224	6115118
6115122	6115250	6115227
6115219	MO_1071	6115233
MO_1013	MO_1084	6115234
MO_1014	MO_1094	6115237
MO_1020	MO_1179	6115242
MO_1054	MO_1184	6115247
MO_1095	MO_1219	6115251
MO_1114	MO_1336	MO_1040
MO_1124	SC_9001	MO_1118
MO_1128	SC_9016	MO_1176
MO_1161	SC_9017	MO_1232
MO_1192	SC_9018	MO_1244
MO_1202	SC_9023	MO_1262
MO_1215	SC_9026	MO_1316
MO_1221	SC_9029	SC_9007
MO_1241	SC_9031	SC_9036
MO_1249	SC_9034	SC_9038
MO_1277	SC_9037	SC_9058
MO_1337	SC_9046	SC_9081
MO_1339	SC_9048	SC_9092
SC_9008	SC_9071	SC_9097
SC_9009	SC_9072	
SC_9010	SC_9091	
SC_9012	SC_9094	
SC_9019	TP_2009	
SC_9022	TP_2020	
SC_9028	TP_2054	
SC_9030		
SC_9032		
SC_9043		
SC_9047		
SC_9049		
SC_9050		

SC_9054		
SC_9055		
SC_9057		
SC_9059		
SC_9060		
SC_9061		
SC_9062		
SC_9063		
SC_9068		
SC_9073		
SC_9080		
SC_9083		
SC_9086		
SC_9093		
SC_9099		
TP_2001		
TP_2010		
TP_2032		
TP_2034		
TP_2060		
TP_2061		
TP_2064		

Appendix H: Clusters performed by approach five

0	1	2
'SC_9091'	'TP_2061'	'MO_1176'
'MO_1336'	'SC_9086'	'SC_9097'
'SC_9031'	'MO_1339'	'6115237'
'SC_9081'	'MO_1337'	
'SC_9080'	'SC_9068'	
'MO_1316'	'SC_9062'	
'TP_2054'	'SC_9061'	
'SC_9073'	'SC_9060'	
'SC_9072'	'SC_9057'	
'SC_9071'	'MO_1221'	
'MO_1277'	'SC_9055'	
'MO_1202'	'SC_9038'	
'SC_9063'	'SC_9050'	
'MO_1262'	'MO_1192'	
'MO_1249'	'SC_9049'	
'MO_1244'	'SC_9047'	
'MO_1241'	'SC_9043'	
'SC_9059'	'SC_9030'	
'SC_9058'	'TP_2032'	
'MO_1232'	'SC_9028'	
'MO_1219'	'SC_9022'	
'MO_1215'	'SC_9019'	
'SC_9054'	'MO_1128'	
'SC_9048'	'MO_1114'	
'SC_9046'	'SC_9010'	
'MO_1184'	'SC_9009'	
'MO_1179'	'SC_9008'	
'SC_9018'	'TP_2001'	
'SC_9037'	'MO_1054'	
'SC_9036'	'MO_1014'	
'SC_9034'	'MO_1013'	
'SC_9032'	'TP_2064'	
'TP_2034'	'SC_9093'	
'MO_1161'	'SC_9099'	
'SC_9029'	'6115219'	
'SC_9026'	'6115114'	
'TP_2020'	'1115183'	
'SC_9023'	'6115118'	

'SC_9017'		
'SC_9016'		
'MO_1124'		
'MO_1118'		
'SC_9012'		
'SC_9007'		
'TP_2009'		
'TP_2010'		
'MO_1095'		
'SC_9001'		
'MO_1094'		
'MO_1084'		
'MO_1071'		
'MO_1040'		
'MO_1020'		
'TP_2060'		
'SC_9083'		
'SC_9092'		
'SC_9094'		
'6115251'		
'6115247'		
'6115242'		
'6115122'		
'1115202'		
'6115117'		
'6115115'		
'1115161'		
'6115233'		
'6115123'		
'1115153'		
'6115121'		
'1115156'		
'1115154'		
'6115227'		
'6115234'		
'6115224'		
'1115244'		
'1115157'		
'6115250'		

Appendix I: Clusters performed by approach six

0	1	2
1115154	1115161	MO_1176
1115156	6115117	SC_9097
1115157	6115234	SC_9031
1115183	6115251	
1115202	MO_1040	
1115244	MO_1316	
6115114	SC_9036	
6115115	SC_9080	
6115118	SC_9091	
6115121		
6115122		
6115123		
6115219		
6115224		
6115227		
6115233		
6115237		
6115242		
6115247		
6115250		
MO_1013		
MO_1014		
MO_1020		
MO_1054		
MO_1071		
MO_1084		
MO_1094		
MO_1095		
MO_1114		
MO_1118		
MO_1124		
MO_1128		
MO_1161		
1115153		
MO_1179		
MO_1184		
MO_1192		
MO_1202		

MO_1215		
MO_1219		
MO_1221		
MO_1232		
MO_1241		
MO_1244		
MO_1249		
MO_1262		
MO_1277		
MO_1336		
MO_1337		
MO_1339		
SC_9001		
SC_9007		
SC_9008		
SC_9009		
SC_9010		
SC_9016		
SC_9017		
SC_9018		
SC_9019		
SC_9022		
SC_9023		
SC_9026		
SC_9028		
SC_9029		
SC_9030		
SC_9032		
SC_9034		
SC_9037		
SC_9038		
SC_9043		
SC_9046		
SC_9047		
SC_9048		
SC_9049		
SC_9050		
SC_9054		
SC_9055		
SC_9057		
SC_9058		

SC_9059		
SC_9060		
SC_9061		
SC_9062		
SC_9063		
SC_9068		
SC_9071		
SC_9072		
SC_9073		
SC_9081		
SC_9083		
SC_9086		
SC_9092		
SC_9093		
SC_9094		
SC_9012		
SC_9099		
TP_2001		
TP_2009		
TP_2010		
TP_2020		
TP_2032		
TP_2034		
TP_2054		
TP_2060		
TP_2061		
TP_2064		

Appendix J: Clusters performed by approach seven

0	1	2
SC_9007	MO_1202	1115153
SC_9018	MO_1176	1115154
SC_9034	SC_9097	1115156
SC_9047	SC_9062	1115157
SC_9091	6115237	1115161
TP_2060		1115183
		1115202
		1115244
		6115114
		6115115
		6115117
		6115118
		6115121
		6115122
		6115123
		6115224
		6115227
		6115233
		6115234
		6115237
		6115242
		6115247
		6115250
		6115251
		MO_1013
		MO_1014
		MO_1020
		MO_1040
		MO_1054
		MO_1071
		MO_1084
		MO_1094
		MO_1095
		MO_1114
		MO_1118
		MO_1124
		MO_1128
		MO_1161

		MO_1176
		MO_1179
		MO_1184
		MO_1192
		MO_1215
		MO_1219
		MO_1221
		MO_1232
		MO_1244
		MO_1249
		MO_1262
		MO_1277
		MO_1316
		MO_1336
		MO_1337
		MO_1339
		SC_9001
		SC_9008
		SC_9009
		SC_9010
		SC_9012
		SC_9016
		SC_9017
		SC_9019
		SC_9022
		SC_9023
		SC_9028
		SC_9029
		SC_9030
		SC_9031
		SC_9032
		SC_9036
		SC_9037
		SC_9038
		SC_9043
		SC_9046
		SC_9048
		SC_9049
		SC_9050
		SC_9054
		SC_9055

		SC_9057
		SC_9058
		SC_9059
		SC_9060
		SC_9061
		SC_9063
		SC_9068
		SC_9071
		SC_9072
		SC_9073
		SC_9080
		SC_9081
		SC_9083
		SC_9086
		SC_9092
		SC_9093
		SC_9094
		SC_9097
		SC_9099
		TP_2001
		TP_2009
		TP_2010
		TP_2020
		TP_2032
		TP_2034
		TP_2054
		TP_2061
		TP_2064

Appendix K: Clusters performed by approach eight

0	1	2
1115154	6115227	1115153
1115156	6115237	SC_9012
1115157	6115242	
1115161	MO_1118	
1115183	MO_1176	
1115202	MO_1232	
1115244	MO_1244	
6115114	MO_1262	
6115115	SC_9007	
6115117	SC_9097	
6115118		
6115121		
6115122		
6115123		
6115219		
6115224		
6115233		
6115234		
6115247		
6115250		
6115251		
MO_1013		
MO_1014		
MO_1020		
MO_1040		
MO_1054		
MO_1071		
MO_1084		
MO_1094		
MO_1095		
MO_1114		
MO_1124		
MO_1128		
MO_1161		
MO_1179		
MO_1184		
MO_1192		
MO_1202		

MO_1215		
MO_1219		
MO_1221		
MO_1241		
MO_1249		
MO_1277		
MO_1316		
MO_1336		
MO_1337		
MO_1339		
SC_9001		
SC_9008		
SC_9009		
SC_9010		
SC_9016		
SC_9017		
SC_9018		
SC_9019		
SC_9022		
SC_9023		
SC_9026		
SC_9028		
SC_9029		
SC_9030		
SC_9031		
SC_9032		
SC_9034		
SC_9036		
SC_9037		
SC_9038		
SC_9043		
SC_9046		
SC_9047		
SC_9048		
SC_9049		
SC_9050		
SC_9054		
SC_9055		
SC_9057		
SC_9058		
SC_9059		

SC_9060		
SC_9061		
SC_9062		
SC_9063		
SC_9068		
SC_9071		
SC_9072		
SC_9073		
SC_9080		
SC_9081		
SC_9083		
SC_9086		
SC_9091		
SC_9092		
SC_9093		
SC_9094		
SC_9099		
TP_2001		
TP_2009		
TP_2010		
TP_2020		
TP_2032		
TP_2034		
TP_2054		
TP_2060		
TP_2061		
TP_2064		

Appendix L: Clusters performed by approach nine

0	**1**	**2**
'TP_2061'	'SC_9091'	'MO_1244'
'SC_9086'	'MO_1336'	'MO_1176'
'MO_1339'	'SC_9031'	'SC_9097'
'MO_1337'	'SC_9081'	'6115237'
SC_9068'	'SC_9080'	
'SC_9062'	'MO_1316'	
'SC_9061'	'TP_2054'	
'SC_9060'	'SC_9073'	
'SC_9057'	'SC_9072'	
'MO_1221'	'SC_9071'	
'SC_9055'	'MO_1277'	
'SC_9038'	'MO_1202'	
'SC_9050'	'SC_9063'	
'MO_1192'	'MO_1262'	
'SC_9049'	MO_1249'	
'SC_9047'	'MO_1241'	
'SC_9043'	'SC_9059'	
'SC_9030'	'SC_9058'	
'TP_2032'	MO_1232'	
'SC_9028'	'MO_1219'	
'SC_9022'	'MO_1215'	
'SC_9019'	'SC_9054'	
'MO_1128'	'SC_9048'	
'MO_1114'	'SC_9046'	
'SC_9010'	MO_1184'	
'SC_9009'	'MO_1179'	
'SC_9008'	'SC_9018'	
'TP_2001'	'SC_9037'	
'MO_1054'	'SC_9036'	
'MO_1014'	'SC_9034'	
MO_1013'	SC_9032'	
'TP_2064'	'TP_2034'	
'SC_9093'	'MO_1161'	
SC_9099'	'SC_9029'	
'6115219'	'SC_9026'	
6115114'	'TP_2020'	
'1115183'	'SC_9023'	

'6115118'	'SC_9017'	
	'SC_9016'	
	MO_1124'	
	'MO_1118'	
	'SC_9012'	
	'SC_9007'	
	'TP_2009'	
	TP_2010'	
	'MO_1095'	
	'SC_9001'	
	MO_1094'	
	'MO_1084'	
	'MO_1071'	
	'MO_1040'	
	'MO_1020'	
	'TP_2060'	
	'SC_9083'	
	'SC_9092'	
	'SC_9094'	
	'6115251'	
	'6115247'	
	'6115242'	
	'6115122'	
	'1115202'	
	'6115117'	
	'6115115'	
	'1115161'	
	6115233'	
	'6115123'	
	'1115153'	
	'6115121'	
	'1115156'	
	'1115154'	
	6115227'	
	'6115234'	
	'6115224'	
	'1115244'	
	'1115157'	
	'6115250'	

Appendix M: Clusters performed by approach eleven

0	**1**	**2**
'1115153'	'1115183'	'MO_1176'
'1115154'	'6115114'	'SC_9097'
'1115156'	6115219'	'6115237'
'1115157'	'MO_1244'	
'1115161'	'MO_1013'	
'1115202'	'MO_1014'	
1115244'	'MO_1114'	
'6115115'	'MO_1128'	
'6115117'	MO_1161'	
'6115121'	'MO_1192'	
'6115122'	MO_1221'	
'6115123'	'MO_1337'	
'6115224'	'MO_1339'	
'6115227'	'SC_9008'	
'6115233'	SC_9009'	
'6115234'	'SC_9010'	
'6115247'	'SC_9019'	
'6115250'	SC_9022'	
'6115251'	'SC_9028'	
'MO_1020'	'SC_9030'	
'MO_1040'	'SC_9038'	
'MO_1071'	'SC_9043'	
MO_1084'	'SC_9047'	
'MO_1094'	SC_9049'	
'MO_1095'	'SC_9050'	
'MO_1124'	'SC_9055'	
'MO_1179'	'SC_9057'	
'MO_1184'	SC_9060'	
'MO_1202'	'SC_9061'	
'MO_1215'	'SC_9062'	
'MO_1219'	'SC_9063'	
'MO_1232'	'SC_9068'	
'MO_1241'	'SC_9086'	
'MO_1249'	'SC_9093'	
'MO_1262'	'MO_1118'	
'MO_1277'	'SC_9099'	
MO_1316'	'TP_2001'	

'MO_1336'	'TP_2010'	
'SC_9001'	TP_2032'	
'SC_9012'	'TP_2061'	
'SC_9016'	'TP_2064'	
'SC_9017'		
'SC_9018'		
'SC_9023'		
'SC_9026'		
'SC_9029'		
'SC_9031'		
SC_9032'		
'SC_9036'		
'SC_9037'		
'SC_9048'		
'SC_9054'		
'SC_9058'		
'SC_9059'		
'SC_9071'		
'SC_9072'		
SC_9073'		
'SC_9080'		
'SC_9081'		
'SC_9083'		
'SC_9091'		
SC_9094'		
'TP_2009'		
'TP_2020'		
'TP_2034'		
'TP_2054'		
'TP_2060'		