Enliven Archive
Intensifying Thoughts

Research Article www.enlivenarchive.org Enliven: Journal of Genetic, Molecular and Cellular Biology

ISSN: 2379-5700

Enhancement of K-Mean Clustering for Genomics of Drugs

Swathi Muppalaneni

Master of Science in Medical Sciences, Long Island University, Brooklyn & Brookville, New York, USA

*Corresponding author: Swathi Muppalaneni,Master of Science in Medical Sciences, Long Island University, Brooklyn & Brookville, New York, USA, E-mail: Smuppalaneni@gmail.com

Received Date: 6th January 2018
Accepted Date: 10th February 2018
Published Date: 17th February 2018

Citation: Swathi M (2018) Enhancement of K-Mean Clustering for Genomics of Drugs. Enliven: J Genet Mol Cell Biol 5(1): 001.

Abstract

Cancer is known as one of the dangerous diseases in this world in the current time frame. The medical world is trying to be as optimal as possible in order to cure a cancer patient. Cancer have different kind of stages and hence have different levels of treatment as well. Drugs for such kind of diseases have a complex architecture system and they are not very easy to understand for a common man. This research paper aims to create clusters for drugs of cancer disease so that they can be even identified by common man. The architecture of drugs are often referred as Genomics and it has several key features like IC 50 value. The presented architecture has enhanced traditional K-Means algorithm by adding inner groups in the outer cluster. The evaluation has been made on the basis of Cluster Accuracy and error rate.Any clustering technique works with the threshold segmentation and value analysis and without any approval from external component, they can't be termed as accurate and precise. The proposed architecture understands this requirement and hence it has utilized support vector machine for the cluster classification in order to understand the preciseness of the clustering.

Keywords: Lung cancer; Genomics of disease; K means; Clustering

Introduction

Cancer is one of the most critical diseases of this world. Treatment and drug architecture is as complicated as the disease. Although with similar clinical symptoms, different patients may have different responses to the same medication or therapy. Therefore, personalized medicine that makes medical decisions based on the patient's genetic content becomes the main direction of medical science in the future. In order to develop and obtain targeted treatments for individuals, one must resort to the lengthy and costly process of drug development and validation in clinical trials, the most direct way to assess drug efficacy and toxicity. However, the scarcity of resources has limited the practical application of this program. A possible solution to this problem is to directly measure the sensitivity of the patient's tumor cells to the drug of interest in a two-dimensional / three-dimensional in vitro culture [1] or in vivo models like mouse xenograft and genetically engineered mouse models [2]. This method has the potential to capture most of the relevant biological characteristics of a patient's tumor and provides a better model for testing drug sensitivity. Though, this method is expensive, time-consuming and difficult to scale up to simultaneously screen dozens or hundreds of drugs. As high-throughput technologies have evolved over the past decades, several groups have proposed alternative approaches to establish the genomic predictor of drug response in large cancer cell lines [3-8]. Most of these methods are based on gene expression profiling.This article focuses on the use of data mining architecture to cluster the genome. Data mining is the procedure of classifying through the large set of data to classify patterns and set relationships to resolve problems by data analysis [1]. Data mining tools empower enterprises to predict expected trends.The principal steps involved in a data mining method are:

1. To extract, transform and upload data into a data shed
2. To save and control data in multidimensional databases
3. To implement data access to business analysts by using application software
4. To display analyzed data in easily readable forms e.g. graphs (Figure1).

Clustering is one of the most general utilized fields of data mining [2]. A cluster is a group of subset of objects that are similar. A subset of object is considering the one that have minimum distance among some two objects in the cluster.

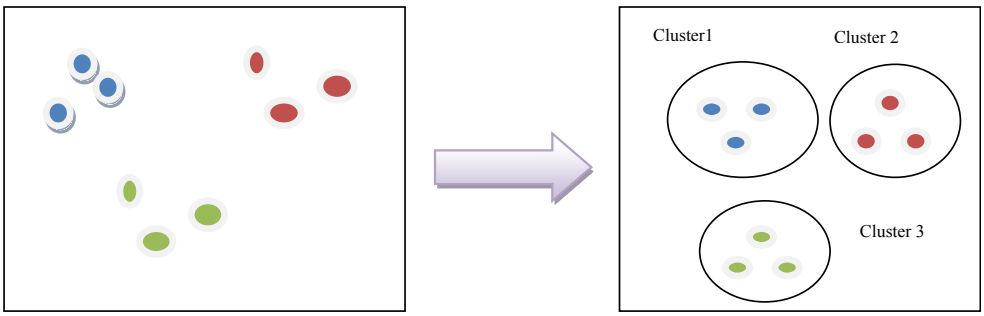

Figure 1: Scattered and clustered document

K-Mean Clustering Algorithm

It is an unsupervised learning algorithm which is used to resolve the clustering problem. By using this algorithms K number of different clusters are grouped. The main aim of k-mean clustering is to reduce the cluster as per the features of data [5-7] (Figure 2).

The above figure shows the K-mean clustering algorithm in which we are considering the n number of data and three types of clusters that are represented by red, black and blue colors [8]. Here, yellow circle represents the centroid of each cluster that shows the mean value of group. The clinical observations have immense use in pharmacy store so to decrease forgery and prevent drug abuse. Documents clustering are used mostly for clinical notes to research for categorizing them into significant clusters. This is done principally to position important patterns. This is proved by increasing speed, efficiency and accuracy of managing information in the medical diagnosis field. K-means clustering is well known extensive clustering technique. It is mainly used to cluster the numerical data. By using this technique, we can cluster text data corpus in unstructured and semi-structured formats. This is completed for a variety of reasons [9].

Related Work

Fuhai et al., has introduced network-based computational methods for the collaborative drug combination prediction.Tao et.al, presented a new, integrated approach that combines ontology reasoning with network-assisted gene sequencing predicts new drug targets. The authors has used colorectal cancer (CRC) as a proof-of-concept case to illustrate the method Iwata et.al, constructed prediction models for sparse-induction classifiers by learning a combination of known drugs based on the Orange Book and KEGG DRUG databases. Russ et.al, Put forward the key point is the drug sensitivity of the cancer cell and synergistic effect of the drug, highlighting the advantages and disadvantages of the current method.

Proposed Clustering Architecture

The enhancement of K Means algorithm has to be applied over a given set of data for cancer. There are several kind of cancers in the sheet data and different drug representation for the same. The following table illustrate the drug architecture (Table 1).

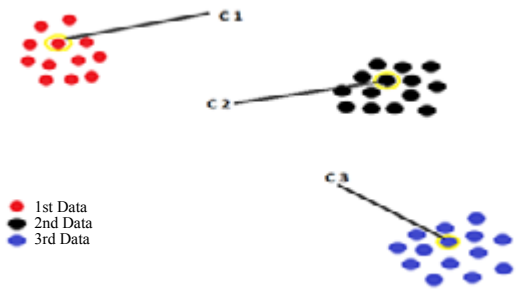

Figure 2: K-mean clustering algorithm

Cell line	TCGA classification	Tissue	Tissue sub-type	IC50	AUC
EoL-1-cell	UNCLASSIFIED	blood	haematopoietic_neoplasm_other	0.0094	0.0237
IGR-37	SKCM	skin	melanoma	0.0109	0.0323
MV-4-11	UNCLASSIFIED	blood	leukemia	0.0121	0.0216
SH-4	SKCM	skin	melanoma	0.0159	0.119
G-MEL	SKCM	skin	melanoma	0.0159	0.0591
K5	THCA	thyroid	thyroid	0.0198	0.0661
A375	SKCM	skin	melanoma	0.0206	0.0425
MEL-HO	SKCM	skin	melanoma	0.0211	0.0786
MOLM-13	LAML	blood	acute_myeloid_leukaemia	0.0264	0.0812
OCI-AML2	LAML	blood	acute_myeloid_leukaemia	0.0293	0.0401
451Lu	SKCM	skin	melanoma	0.034	0.104
M14	SKCM	skin	melanoma	0.0346	0.121
WM35	SKCM	skin	melanoma	0.0348	0.115
BHT-101	THCA	thyroid	thyroid	0.0359	0.121
MMAC-SF	SKCM	skin	melanoma	0.0365	0.124

Table 1: Unstructured dataset

The given data is unstructured and hence to make the data structured in order to study the disease the following algorithm architecture has been implemented.

Algorithm Enhance ClusterKmeanarch=function create innercluster(Actual_ data)

[row,cols]=Size (Actual_data) // Actual Data is the data set utilized in the clustering scenario

Group= [ ]; // Initializing the group parameters

Groupname= [ ];

Grpcount=0; Traindata= [ ]; record_count=1;

Group[0]=ord_data(1).issuenumber

Groupname[0]=org_data(1).issuename

Currentgroup=groupname[0];

For i=1:rows

If (ord_data (i).Tissue name==Currentgroup)

Traindata[recordcount,0:cols]=Org_data(i).Allrecords;

Record_count=record_count+1;Group(record_count)=grpcount;

Else

Groupcount=groupcount+1;

Traindata[recordcount,0:cols]=org_data(i).Allrecords

Record_count=record:count+1;

Group(record_count)=grpcount;

Apply K mean(Traindata);

End if

After the successful implementation of the proposed algorithm, the following structured set sample is obtained. (Table 2)

The evaluated parameters are presented in the next section.

Cell line	**TCGA classification**	**Tissue**	**Tissue sub-type**	**IC50**	**AUC**
Group 1					
IST_MELI	SKCM	Skin	Melanoma	.0042	.1550
C32	SKCM	Skin	Melanoma	.0060	.1760
RPM1	SKCM	Skin	Melanoma	.0100	.2230
Group 2					
A549	LUAD	Lung	Lung_NSCL	.0045	.1540

Table 2: Structured formatted data

Result Analysis

As the clustering has been done by means of the proposed architecture system, (Figure 3) represents different drug responses and IC 50 value after the implementation of Enhanced K means based on the clustered feature set, the following parameters have been evaluated.

1. Precision= (True Drug IC 50 Value + False Drug IC 50 Value) / Total IC 50 Value

2. Recall = True Drug IC 50 Value / True Drug IC 50 Value+ False Drug IC 50 Value (Table 3, Figure 4).

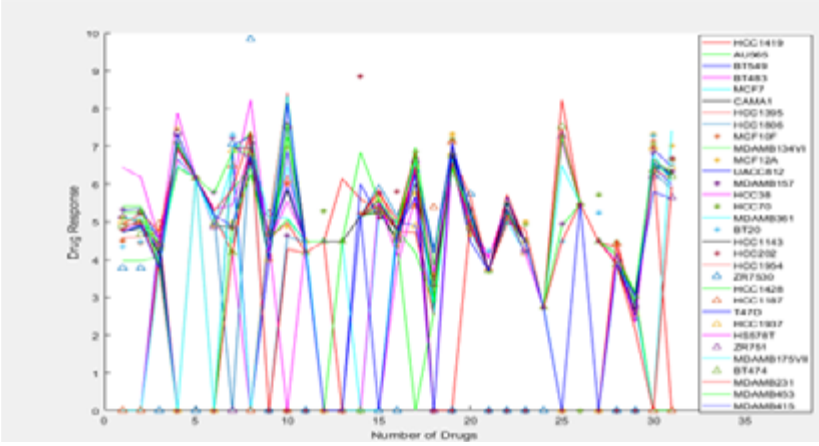

Figure 3: Different drug responses

Drug Group	Precision	Recall
HCC 1419	.74	.68
BZR7530	.76	.72
HCC1187	.784	.73
BT474	.79	.75

Table 3: Precision and Recall of Created Cluster Groups

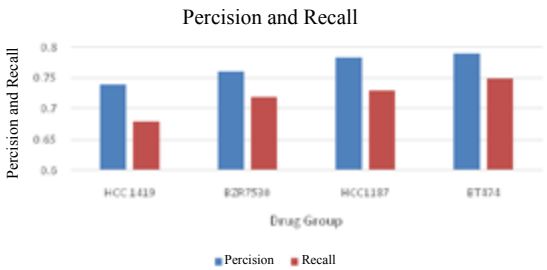

Figure 4: Precision and Recall of Clustered Groups

The clustered group has been passed to the evaluation parameters and maximum precision of .79 has been attained. High precision value will lead to a high recall value as well as high classification rate. The precision as well as recall value represents that the presented clustering architecture is significant enough to take the work to the next level of classification.The evaluation is still pending as discussed in the abstract that any clustering algorithm can not be termed as efficient till it does not prove its worth with any classification algorithm. The proposed work utilizes Support Vector Machines (SVM) to solve this problem. There are several reasons due to which SVM has been opted in this contrast.

a) It is a binary classifier and hence it is simple to understand as compared to other architectures

b) The proposed architecture divides the entire data into two segments and hence binary classification suits the proposed architecture very well.

c) There are several kernel option in Support Vector Machine and hence it performs the work more significantly

The architecture of SVM can be written algorithmically as follows

1. Function SVM_ Class_Bifurcation(Clusters , Clusters_Threshold)
2. Grouping=[ ];
3. Svm.Kernel.Function=”Linear/Polynomila”;
4. Foreachelem in Clusters.Count
5. GroupValue(elem)=Data.Clustervalue(elm);
6. If Processing.Data.Cluster==1
7. Group(gcount)=1
8. Gcount=gcount+1;
9. Else
10. Group(gcount)=2;
11. Gcount=1;
12. End if
13. SVM_Struct=Train_SVM(GroupValue,Group,Kernel_Function);
14. Classification_SVM_Vector=SVM_Classify(Test_set,SVM_Struct)
15. Evaluate Classification Error( )
16. End Function

The Function takes the data of both the clusters one by one. The inputs to the functions are the Clustered elements of both the groups. Here the Support Vector Machine will also be getting Group elements as labels of 1 and 2. Both the labels will be trained against their respective class values and the following training results will be obtained. (Figure 5(a) and 5(b)) represents different training architecture with SVM functioning with different kernel function and different boundary dimensions. The figure clearly depicts the if the kernel function changes then the architecture of the entire set completely changes. The proposed architecture has attempted to use both the algorithm. The classification results are listed in (Figure 6(a) and Figure6(b)) represents the classification architecture of the development architecture. Both types of kernels have their advantages and disadvantages. The proposed architecture have utilized both the models and notices that the cluster efficiency different by 5 to 6 % by changing the kernel.

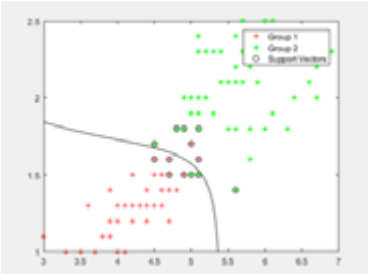

Figure 5(a): Training with Polynomial Kernel

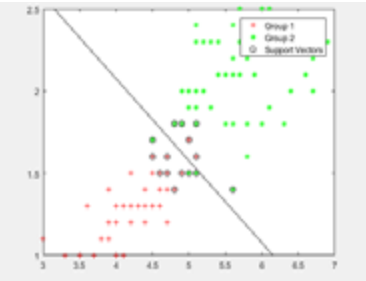

Figure 5(b): Training with Linear Kernel

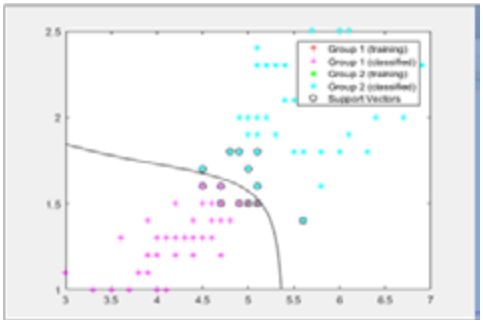

Figure 6(a): Polynomial ClassificationModel

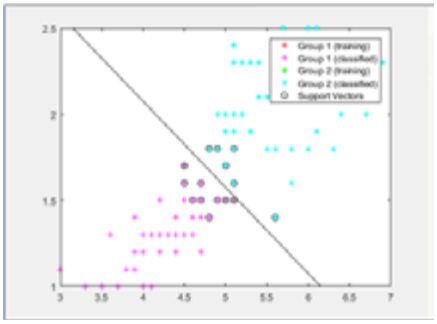

Figure 6(b): Linear Classification Model

Conclusion and Future Scope

The presented work has enhanced the clustering approach for the drug identification and grouping so that the classification process becomes easy. The presented architecture takes a precision value of .79 and the recall value of .75 at max. The presented algorithm opens a lot of opportunities for the future researchers. The future research work may utilize the proposed architecture in order to enhance the classification mechanism. The clustered group may also be passed to Machine Learning algorithm for predicting the accurate drug to accurate type of disease.The classification mechanism is performed by Support Vector Machine which is a binary classifier. The reason behind choosing this model as classifier is because the proposed architecture divides the data into two clusters only and they are sufficient for SVM. The proposed model has utilized different kernel function and has noticed that the efficiency of clustering varies by 5 to 6 % if the kernel function is changed. The proposed architecture has opened a lot of future gates. The future research workers may try their hand in optimizing the similarity indexes and also can try to optimize the kernel function by varying Gaama and Lemma values.

References

1. Jia J, Zhu F, Ma X, Cao Z, Cao ZW, et al. (2009) Mechanisms of drug combinations: interaction and network perspectives. Nat Rev Drug Discov 8: 111-128.
2. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45: 1113-1120.
3. Huang L, Li F, Sheng J, Xia X, Ma J, et al. (2014) DrugComboRanker: drug combination discovery based on target network analysis. Bioinformatics 30: 228- 236.
4. Hofree M, Shen JP, Carter H, Gross A, Ideker T (2013) Network-based stratification of tumor mutations. Nat Methods 10: 1108-1115.
5. Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, et al. (2013) Integrated genomic characterization of endometrial carcinoma. Nature 497: 67-73.
6. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, et al. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483: 603-607.

7. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483: 570-575.
8. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, et al. (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res 41: 955-961.
9. Li F, Huang L, Sheng J, Wong S (2016) Network-based Computational Drug Combination Prediction, IEEE.
10. Tao C, Sun J, Zheng WJ, Chen J, Xu H (2015) Colorectal cancer drug target prediction using ontology-based inference and network analysis, IEEE.