COHERENT KNOWLEDGE DISCOVERY IN HEALTH CARE USING GENOMIC K-MEANS CLUSTERING

Rosy Mishra  
Asst.Proff in CSE, VIT,BARGARH

Seema Baghel  
B.Tech 3rd yr CSE

Rosan Patel  
B.Tech 3rd yr CSE

Monali Patel  
B.Tech 3rd yr CSE

Computer Science and Engineering, 
Vikash Institute of Technology,Bargarh, India

Abstract: The clustering of dataset has become a challenging issue in the field of health analytics. Existing k-means clustering algorithms requires scalable solutions to manage large datasets. The goal of the present work is to detect the number of non-informative or meaningless variable present in a data set but cluster select those variable which have some meaning as well as important variable. It is based on genetic algorithm that help us to make the data easy to understand. We use these procedure in K-means by using Silhouette plot as the function of genetic algorithm. The main problem with it is when we use these method in the real data and re-apply at that time do not know a prior what the real cluster structure is that means whether it is a number and composition of the groups. These work leads to the evolution of the silhouette values that is built by the application of k-means any meaningless or non-informative variable is added to the original data set in both literature as well as high dimension data. These procedure is also added to the real data set. The methodology is applied to different genomic data set like lung cancer, brain tumor, normal tumor, diabetes and kidney gene expression data set.

Keywords:— clustering, k-means, silhouette plot, PCA, Genomic Datasets

I. INTRODUCTION

In k-means we have to give n set of data point in D-dimension space R and an integer K and we face problem like we cannot determine K points in Rd called center so that we can minimize the mean squared distance from each data point to its nearest center. Popular algorithm for k-means is Lloyd’s which is also called as filtering algorithm. Here Kd-tree is used as major data structure. We can increase the efficiency in two ways. First we have to show that the algorithm run faster as the separation between clusters increases by presenting the data sensitive algorithm at the running time. Second For the applications in color quantization, data compression and image segmentation we can present a number of pure logical studies both on logically generated data on the real data sets.

Here we introduce silhouette k-means clustering, in a new data set integration strategy that combines experimental and clinical high-dimensional data-means clustering identifies sets of genes that stand out in experimental data while at the same time display different expression in clinical data. Different application we report: The integration of clinical high dimensional as well as low dimensional data with (i) genome-wide chromatin and (ii) with cell perturbation.Cluster is generally used for finding groups like biology, psychology and economics field where cluster is use. The main problem of different data set is that it forms different cluster. If we know the cluster number then we can know how many groups are inherited from the clustering algorithm design. One way to control cluster analysis is that first the end user have to input the cluster number in advance that is useful to identify the number of cluster form in the data set. The second way is that it can automatically sense the formation of cluster. There are various ways to determine a cluster but we usually use k-means which is a simple way. K-means mainly have 6 approach i.e. 1) by rule of thumb, 2) Elbow method, 3) Information Criterion Approach, 4) An information theoretic approach, 5) Choosing k using the silhouette, 6) cross-validation.

Medical expert generally interested in independent decision. Data mining provides oriented approach to the novel and hidden data pattern. It is difficult to discover hidden relationship. It is used to extract data mining pattern. There are many techniques to extract data pattern like classification, clustering, association rule, decision tree and ANN. Clustering is used to group similar type of data and each group is called as cluster. This paper to analyze the different clustering technique in diagnosis the medical disease. Data mining health organization is one of the motivating area for research. It is a new powerful technology which is of high interest in computer world. The data generated by health organization is very vast and complex due to which it is difficult to analyze the data in order to make important decision regarding patient health, hospital related data, patients, medical claims, treatment cost etc. and keep all the record. Data mining is responsible to analyze various factor such as types of food, different working environment, education level, living conditions, availability of pure water, health care services, and cultural, environmental and agricultural factors.

II. Related Work:

Medical diagnosis is subjective it depend on available data and physician experience. Diagnosis differ from physician to physician at various time. Data mining is a most challenging field because of its huge dataset, complex, diverse, hierarchical, time series and verifying quality. Dataset is in divide and diffused state so it is very challenging task to integrate data. For example we take a small subset from the medical dataset from which we get the accurate result but from that in some dataset we have to apply algorithm also because if we apply algorithm to all dataset it is difficult to extract the information. Clustering is one of the important field not only for computer scientist but also for statistician and also for patterns recognition experts.
Fradkin and Madigan developed a variation of k-means clustering to compress the image called “divisive k-means” which require a powerful machine to get the accurate results.

**K-means clustering:**

K-mean clustering is a well know algorithm that solve number of clustering problem by one of the simplest unsupervised method. It has an easy way to classify a given data set that following a simple rule having a certain criteria through a number of clusters fixed a priori. One for each cluster the main idea is to define k centroids. Its centroid should be placed in different location causes different results, so the better option is to place them far from each other as much as possible. In associated to the centroid by taking the next step where each point belonging to a given data set. After the first step an early grouping is done when there is no point is pending. Resulting from the previous step at that point of centroid we have to re-calculate the K new centroids as barycenter’s of the cluster. We have to do a new binding between the same points of the data set and the nearest new centroid is formed, the a loop will be generated which will result to the k centroid change their location step by step until no more changes are done. In other words centroids do not move from one location to another. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

\[
J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left( \|x_i - v_j\|^2 \right)
\]

where,

- \(\|x_i - v_j\|^2\) is the Euclidean distance between \(x_i\) and \(v_j\).
- \(c_i\) is the number of data points in \(i^{th}\) cluster.
- \(c\) is the number of cluster centers.

**Silhouette plot:**

The separation distance between the resulting clusters can be study by the silhouette analysis. It shows the closest distance by measuring from one of its point in a cluster to its neighboring point and its give us a way to access the number of cluster visually. It ranges from [-1, 1].

To check whether each point of one cluster is nearer to the neighboring cluster or not, we can detect it by examining whether the silhouette coefficient is near +1 or 0 or a negative value. If it is near +1 then it can point out that the sample is far away from the neighboring clusters and if the value of silhouette coefficient is detected as 0 then the sample is on or very close to the decision boundary between the two neighboring clusters and If the silhouette coefficient is a negative value than those samples might have been assigned to the wrong cluster.

In this example we have to choose a most favorable value for n_clusters after the silhouette analysis. The silhouette plot displays that n_cluster value of 3, 5, and 6 are a bad pick for the given data due to the existing of clusters with lower level of average silhouette scores also due to the vast fluctuations in the size of the silhouette plots. This analysis is more indecisive in deciding between 2 and 4.

The cluster size can be imagine from the thickness of the silhouette plot. When the silhouette plot is equal to 2 then the silhouette cluster is 0 and owing to its bigger size the grouping of 3 sub clusters is converted into one big cluster. However all the plots are more or less of likely to have same thickness if n_clusters is equal to 4. Hence it can be substantiated for the labeled scatter plot on the right because its similar size.

### III. Proposed Model:
[Fig1: Steps to cluster on Brain Tumor data]

- Brain Tumor Microarray Data (90*5921), GLIOMAS (1:25)
- Random Feature Selection
- Feature Selection by finding Principal Component (PCA)
- Reduced Matrix by PCA (90*255)
- Clustering

- GLIOM
- MENINGIO
- NORMAL
- PITUITAR ADENOM
- NERVE SHEATH
Dataset Used For Analysis Without the proper data any research area is incomplete. To train a machine by a machine learning algorithm right dataset is more essence. Microarray technology has given a platform to generate genome data by using intensity color depth measurement from DNA chip.

<table>
<thead>
<tr>
<th>Tumor</th>
<th>Size</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Cancer</td>
<td>203*126 01</td>
<td>ADENO (1:139),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NORMAL (140:156),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMCL (157:162),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SQUAMOUS (163:183)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>COID (185:203)</td>
</tr>
<tr>
<td>Brain Tumor</td>
<td>90*5921 0</td>
<td>GLIOMAS (1:25),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Meningiomas (26:48)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NORMAL (49:57)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NERVE SHEATH TUMOR (58:72)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PITUITARADENOMAL (73:90)</td>
</tr>
<tr>
<td>Normal Tumor</td>
<td>4*2000</td>
<td>PT0: NORMAL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PT1: PRIMARY TUMOR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PT2: TUNICA VAGINALIS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PT3: SPERMATIC CORD</td>
</tr>
<tr>
<td>Diabetes</td>
<td>3*300</td>
<td>COMPENSATION</td>
</tr>
<tr>
<td></td>
<td></td>
<td>STABLE ADAPTATION</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UNSTABLE EARLY DECOMPENSATION</td>
</tr>
<tr>
<td></td>
<td></td>
<td>STABLE DECOMPENSATION</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SEVERE DECOMPENSATION</td>
</tr>
</tbody>
</table>

Table1: genome data with its categories

IV. Result Analysis:
Experimental work was designed to compare the performance of proposed K-mean algorithm. Number of data elements selected was 155 in case of lung cancer PCA based reduction. And for the sake of experiment, 5 numbers of clusters (k) were entered at run time for lung cancer and 2 numbers of clusters (k) for colon cancer. The process was repeated 10 times for different data sets generated by MATLAB. The proposed K-mean algorithm is efficient because of less number of iterations and improved cluster quality, as well as reduced elapsed time. In Figure 4, Basic and proposed K-mean clustering algorithms are compared in terms of different data sets. For each run different data sets are generated by MATLAB and entered, to observe the number of iterations. In Figure, basic and proposed K-mean clustering algorithms are compared in terms of same data set. For each run same data set is entered, to observe that at each time numbers of iterations are different in basic K-mean clustering algorithm. The numbers of iterations are fixed in proposed K-mean clustering algorithm because initial centroids are not selected randomly. Basic K-mean clustering algorithm gives different clusters, as well as clusters size differs in different runs. Table 1 shows different results for same data set as well elapsed time. Proposed K-mean clustering algorithm gives same clusters, as well as clusters size is same in different runs. Comparison with Other K-Mean Clustering Algorithms: Comparison of proposed K-mean clustering algorithm with basic and other enhanced algorithm is given. Comparison of Basic and Proposed K-Mean Clustering Algorithm: Proposed K-mean algorithm is efficient from basic K-mean algorithm in terms of iterations, cluster quality as well as elapsed time. As in basic K-mean algorithm, initial centroids are selected randomly from the input data, so clusters vary from one another, because of which the number of iterations and total elapsed time also changes in each run of the same data.

![Fig2:K-Means plot and Silhouette plot for clustering of Brain tumor original data having 5 classes](image-url)
Fig3: K-Means plot and Silhouette plot for clustering of Normal tumor original data having 5 classes

Fig4: K-Means plot and Silhouette plot for clustering of Normal tumor original data having 5 classes

Fig5: K-Means plot and Silhouette plot for clustering of Diabetis original data having 5 classes
V. Conclusion

Many research tried to devote same matrices that is based on clustering but till now there is no algorithm where greedy approach is used to find out the same and nearly result as we get form k-means. In this paper we have a techniques where we can get cluster of large data set which is not possible by k-means. In future by using G-mean algorithm we can form the distributed version where data and computational power can be spread on different remote machine. Another method is there that we can handle variable by taking k instead of a constant as this is very helpful in this real world examples where data is always changing and k cannot be fixed. We also plan to work on clustering algorithm and data taken should be preprocessed using specific rules so that they can enhance the performance and the results of the clustering algorithm.

VI. Reference:
[5]. YoupingDeng ,DheerajKayarat, Mohamed O. Elasri, Susan J. Brown,” Microarray Data Clustering Using Particle Swarm Optimization K-means Algorithm”,
[8]Rosy Mishra(Asst. Professor in CSE) &Er. Plaban Dash (Lecturer in EEE) “Feature Reduction and Robust clustering on Coloncancer data using principal component analysis "in the International Journalof Recent Engineering Research and Development (IJRERD) ,ISSN: 2455-8761,Volume No. 02 - Issue No. 07, 2017

Rosy Mishra: B.Tech in Information Technology from Trident Academy of Technology Bhubaneswar. M.Tech in computer science engineering From Gandhi Institute Technology Bhubaneswar. Currently She is working as Asst. Professor in the department of Computer science & Engineering, in Vikash Institute of Technology, Bragarh.

SeemaBaghel: B.Tech in Computer Science Engineering from Vikash Institute of TechnologyBargarh.

Rosan Patel: B.Tech in Computer Science Engineering from Vikash Institute of TechnologyBargarh.
Monali Patel: B.Tech in Computer Science Engineering from Vikash Institute of Technology Bargarh.