Study on Integrative Clustering of Multiple Genomic Data to Discover Cancer Subtypes

SANKHA SUBHRA MULLICK

ROLL NUMBER: MTC-1212

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENT FOR THE DEGREE OF MASTER OF TECHNOLOGY
IN COMPUTER SCIENCE IN INDIAN STATISTICAL INSTITUTE 2014

INDIAN STATISTICAL INSTITUTE

TO WHOM IT MAY CONCERN

I hereby recommended that the thesis entitled "Study on Integrative Clustering of Multiple Genomic Data to Discover Cancer Subtypes "prepared under my supervision by Sankha Subhra Mullick (Roll No. MTC-1212), may be accepted in partial fulfillment for degree of Master of Technology in Computer Science in Indian Statistical Institute.

_____-

Dr. Pradipta Maji
Associate Professor
Machine Intelligence Unit
Indian Statistical Institute, Kolkata

Abstract

With the advancement of technology, different sources of genetic information become available with a low cost. In the research for finding cancer subtypes, what will help to proceed with a targeted treatment, this opened up a new dimension. However, the basic problem is how to reach towards a proper integration scheme such that both the personal significance and interactive information is conserved, because only then it will be possible to utilize the data resource and obtain richer information about subtypes. On the other hand, as the subtypes are not always properly defined or even known, thus any solution should be unsupervised in nature. This study presented an integration scheme based on the concept of iCluster method, to address these issues. With its many merits, however the crisp nature of clusters obtained by iCluster is not always natural in the case of overlapping and incomplete nature of the data, thus a rough-fuzzy clustering approach will be more suitable, where an addition of intelligent initial center selection algorithm is most desired. A variety of cluster validation index are used to support the claims and present the findings on two different cancer data.

Acknowledgement

I wish to express my deep sense of gratitude to Dr. Pradipta Maji, for his guidance, encouragement and facilities extended without which this project would not have taken this present shape. I would also like to thank Dr. Sushmita Paul, for her support and valuable advices regarding the topic.

I would like to thank my parents and all my friends and everyone else who supported me with this project. I am very much grateful to all those who was there to help me in understanding the topics, by useful discussions. Your help will always be remembered and appreciated.

_____-

Sankha Subhra Mullick

List of Figures

Contents

Chapter 1

Introduction

With the advance of technology not only we get to know about new sources of genetic informations originated from the description of different structural and functional features of a genetic body such as mRNA, miRNA or the DNA itself, but also obtain them from samples at a low cost. This huge supply of information gave a new dimension to genetic research what aims to integrate these different sources with the intuition of obtaining richer insights on their objectives. This helps form the addition of interactive effects between sources, over their personal significance that is different for the sources. Now it is established that data like gene expression, miRNA expression, copy number variation (CNV) of genes and DNA methylation of cytosine residues at CpG di-nucleotides are very much correlated and coveys very useful description of a cell's health and functional behaviour. But the exact pathway of interaction is quiet vague and hard to isolate from their own modifications. As for example gene expression can be altered in disease state, for mutation or corruption in a very personal modification, but miRNA, methylation pattern and copy number regulates the expression pattern of genes, or different copy of a gene can be methylated differently or targeted by different miRNA resulting in different functional activity.

Here we will use four major types of genetic data, both mRNA and

miRNA expression data is similar in nature. They are both obtained through microarray experiments. They forms a expression tables where we usually take each column to be a gene or miRNA and each row a sample. Each entry is the expression of a sample, which is either a normalized log2 ratio of the intensity of the two channels in a microarray probe which will be mapped to a gene, or a single intensity value for a single channel microarray depending on the make (For multiple probes corresponds to a gene, an average is taken over the final values). DNA methylation data indicates the methylation level of a gene, methylation of gene acts as a switch regulating the expression. This data set is similar in nature where only, instead of genes in the column we have probes and each entry is a β value ranged between 0 and 1. This data set can be transformed into a set, similar to expression table by mapping the probes to corresponding genes and taking average over the case of multiple correspondence[4]. The β value is the ratio of the methylated probe intensity and sum of overall i.e. methylated and unmethylated probe intensity, where a probe corresponds to a CpG site of a gene. So a β value equals to 0 means a completely unmethylated CpG site, and 1 the exact opposite. By default each of the gene occurs in pairs in the two sister chromosomes, except in gamet cells. For reasons this copy number changes by alteration (duplication or deletion in nature) of a part or the entire gene. Being an intrinsic part of evolution CNV or series of single nucleotide polymorphism (SNP) over a large region is observed almost regularly. However as it leads to a complete change of the gene, an unfortunate or corrupted one is very sure to cause genetic disorder and diseases[21]. CNV data is obtained these days through a SNP array which is a variation of microarray.

Cancer is known to effect the entire genetic system and alter their structure and functions differently, thus both the interaction pattern and private feature goes through drastic changes. Like copy number changes of genes which changes the coding and the number of gene and irregular methylation pattern leading to faulty expression of oncogenes and tumor suppressors,

are common. Methylation acts as a switch, the more methylated a gene, the expression lowers down in reverse. Also miRNA shows a very unpredictable behaviour in target mRNA selection and directly regulates a gene's functions taking a major role in different cancers[2]. Thus not only gene expression which in most of the time the result of a deeper alteration, but all these causal factors comes into play when a cancer takes place and the minute variation of their interaction and alteration, results into chancer subtypes. Thus the research of subtype discovery can be highly benefited by this integrative analysis approach. Defining proper subtypes mentioning their individual characteristics is very much necessary for diagnosis and targeted treatment, unfortunately for most of the cases none exists. Thus a type of integration that captures the inter-source association keeping their heterogeneity and specificity alive will obtain more useful and rich information. But such an integration is always a challenge as the data sources are from different domain, scale and interprets differently. Also as the subtypes are not properly defined thus obtaining a labeled data set is impossible, so classification of unknown samples after training is beyond question, however a solution can be obtained through clustering.

In the growing interest and importance, many people trying to attack this integrative clustering problem. Mainly they followed one of the two paths, either the clustering took place on individual data sets followed by an post hoc method of integration, or combine the data set and "jointly" cluster them. In both of this direction there are many notable works. In the first approach the level of agreement between clusters can be found by adjusted rand index or by the help of a consensus clustering (also known as ensemble clustering), an example can be the integration proposed by *Cancer Genome Atlas Network*[12]. However the approach may be attractive due to its simplicity but unfortunately looses shared information by the process. The second approach on the other hand always faces a problem to conserve the source specific information alongside the shared. For example *Qin et*

al.[14] performed a hierarchical clustering on the correlation matrix of gene and miRNA expression. *Lee et. al.*[8] applied a biclustering on the correlation matrix of CNV and gene expression data. While both these methods captured the shared information but failed to accumulate the individual. *Lock et. al.*[9] developed a statistical model known as JIVE, where a decomposition was proposed to capture the individual, joint and residual noise separately. The method was inspired by the PCA and provides a dimension reduction. *Zhang et. al.*[22] used a non-negative matrix factorization method for integrative analysis of different genomic data.

One of the popular method for integrative clustering is iCluster proposed by *Shen et. al. (2009)*[16], which was inspired by probabilistic PCA (PPCA)[18], where the tumor subtypes are modeled with a Gaussian latent variable. However with the many benefits of the algorithm, for a high dimension data the time complexity will be large enough due to a matrix inversion, what will take place in the iterative phase of the algorithm and only a crisp clustering is produced, we will highlight these points later, at the time of presenting the details. iCluster method can be seen as a two stage process, where a feature extraction is followed by a clustering. There can be various types of clustering methods available based on the nature of the data and shape of clusters. Here, in genetic data the inherent possible incompleteness and overlapping character opposes a crisp partition of the samples, we intended to tackle this situation by a rough fuzzy clustering[10] with an intelligent starting center selection[11].

Chapter 2

iCluster: A Brief Discussion

The method, iCluster is one of the most interesting and popular methods of genomic data integration and clustering. The algorithm is notable for its many merits such as a strong mathematical basis, which also opens up opportunities for further development or extension by adding modules supporting different new data type or type specific treatment. The algorithm provides internal feature selection process by making the insignificant features to have less effect on the result and outputs a transformed data in a low-dimensional space that can be used for further analysis, acting similar to a feature extraction algorithm. However the iCluster used a crisp natured clustering algorithm like k-means to cluster the samples into subtypes, which is not always supportive to the nature of real life data which contains vagueness or incompleteness and overlapping clusters. Also, it has a high computational complexity, which is discussed in this chapter.

Being one of the simplest clustering algorithm k-means is used widely. It iteratively minimizes an objective function that is actually the total within cluster distance. However a demerit of the algorithm is, it suffers from a local minima convergence problem. The k-means algorithm is very much sensitive to the initial cluster centers, which are usually chosen at random, and reaches to a local minima of the objective function rather than the global one. *Zha et.*

al.[20] showed that the objective function of k-means can be expressed in the form of a matrix trace maximization problem. This type of formulation of the problem is appreciated because, now the continuous solution of the cluster indicator matrix will be represented by the k Eigen vector corresponding to the top k Eigen values of the sample gram matrix. A k-means applied on this continuous solution can recover the interpretability of the cluster indicator matrix, or the samples can be labeled properly. *He et. al.*[3] pointed out that among this k Eigen vectors one is a linear combination of others, thus only $k-1$ EIgen vector is sufficient. Now the first $k-1$ Eigen vector actually represents the directions along which, if the dataset is projected, maximum variance will be achieved in that low dimensional space. As a result any distinct subgroup will be best identified. This approach is similar to the principal component analysis (PCA), here only the top $(k-1)$ principal component are considered.

But PCA has the problem of failing to distinguish between the variance and co-variance. This is very important drawback in our topic as the co-variances and variances actually interprets the interaction and specific significance of the datasets. This problem was solved by the introduction of probabilistic PCA[18] or PPCA. PPCA used a Gaussian latent variable model which for a mean centered data is like

$$X = WZ + \epsilon \tag{2.1}$$

inspired by the factor analysis model which inherently takes care of the above problem by representing the correlations by the latent variable Z and the individual variances by ϵ. To be more specific about the model, $Z \sim N(0,1)$ and $\epsilon \sim N(0,\psi)$ thus from the properties of normal distribution $X \sim N(WW' + \psi)$. However, the space spanned by the PCA and PPCA is not similar, such equality is possible under the condition of isotropic error terms, where $\psi = \sigma^2 I$, which is very unlikely in real life cases.

By the virtue of this model an direct extension to multiple dataset can

be observed. Let the datasets be X_i, where $i = 1, \cdots p$, each having n samples and m_i features, and Z is the common latent space, $Z = [z_{kj}], k = 1, \cdots c-1, j = 1, \cdots n$, here we take c to be the number of cancer subtypes. We also take $\mathcal{M} = \sum_{i=1}^{p} m_i$. for a single dataset, let the i^{th} model be

$$X_i = W_i Z + \epsilon_i$$

The p set of models are connected by the latent component which represent the interaction, while $\epsilon = (\epsilon_1, \cdots \epsilon_p)$, where each ϵ_i having mean zero and diagonal covariance matrix say ψ_i represents the residual variance specific to each data types. $W = (W_1, \cdots, Wp)$ is a coefficient or projection matrix. To be precise each W_{ij} actually denotes the contribution of feature i in determining the cluster j. This property allow us to introduce a lasso type error on W which reduces the effect of insignificant features towards zero. Lasso error is a L_1 norm penalty that in many cases actually reduces to a soft thresholding operation defined as follows, S is a soft threshold operation on x under a penalty term λ:

$$S(x, \lambda) = \begin{cases} x - \lambda & \text{if } x > \lambda \\ 0 & \text{if } |x| \leq \lambda \\ x + \lambda & \text{if } x < -\lambda \end{cases}$$

However, to match our model with the model and solution framework of the PPCA we need to assume a continuous parameterization of Z, naming Z^*, such that $Z^* \sim N(0, 1)$. We define $\epsilon \sim N(0, \psi)$, such that $\psi = diag(\psi_1, \cdots \psi_{\mathcal{M}})$. We take $X = (X_1, \cdots X_p)$, and obtain the marginal distribution to be following multivariate normal with mean zero, and variance Σ, where $\Sigma = WW' + \psi$. Now our model resembles the PPCA model defined in equation (2.1). Now to obtain the maximum likelihood estimate of W and ψ to determine $E(Z^*|X)$, we need to utilize the EM algorithm. Before doing

so we need to write the complete data log likelihood, as defined below:

$$\begin{aligned} l_c(W,\psi) = -\frac{n}{2}\left\{\sum_{i=1}^{p} m_i ln(2\pi) + ln(det|\psi|)\right\} \\ -\frac{1}{2}\left\{tr((X - WZ^*)'\Sigma^{-1}(X - WZ^*)) + tr(Z^{*'}Z^*)\right\} \end{aligned} \tag{2.2}$$

We will now use the E and M step of the EM algorithm, which will directly follow up from the theorems of Gaussian distribution. In E step we will calculate the mean and variance of Z^* given X, and in the M step will use those results to update the value of W and ψ. To introduce the effect of the lasso error, we can actually soft threshold the updated W at each iteration. At convergence we will find the $E(Z^*|X)$, which will be our transformed data. The main steps of the algorithm are:

Input: Let us take n samples and p datasets, X_i having dimension m_i, where $i = 1, \cdots, p$ and $\mathcal{M} = \sum_{i=1}^{p} m_i$. The lasso error L, which may be a vector of length p, if we want to use different error for different dataset. We want to cluster the samples into c clusters.

Output: A transformed dataset $E(Z^*|X) = [z_{ij}]$, where $i = 1, \cdots, c-1$ and $j = 1, \cdots, n$.

1. Stack the datasets to form a dataset X, having n sample and $\mathcal{M}$ dimensions. Mean center each dimension (here imagined as the the rows).

2. Find the covariance matrix by calculating XX'. Find the Eigen vectors a_i and Eigen values λ_i, where $i = 1, \cdots, n$ of the covariance matrix. (As $n < \mathcal{M}$, thus the bottom $\mathcal{M} - n$ Eigen values will be 0.)

3. Initialize as follows:

$$sigmaError = \frac{1}{n - c + 1}\sum_{i=c}^{n}\sqrt{\lambda_i} \tag{2.3}$$

$$A = [a_1 \quad a_2 \cdots a_{c-1}]$$

$$b_i = \sqrt{\lambda_i - sigmaError}, \quad i = 1, \cdots c-1.$$

$$B = diag(b_1, \cdots b_{c-1})$$

$$W = AB \tag{2.4}$$

$$\Sigma = WW' + I * sigmaError \tag{2.5}$$

Initialize *convergenceRate=1.0, iteration=0, acceptableError, maxIteration* as per choice and $E(Z^*|X) = 0$. Initialize *psi* as a $\mathcal{M}$ dimensional vector with all elements being *sigmaError*.

4. **E Step:**

$$E(Z^*|X)_{old} = E(Z^*|X)$$

$$E(Z^*|X) = W'\Sigma^{-1}X \tag{2.6}$$

$$E(Z^*Z^{*'}|X) = I - W'\Sigma^{-1}W + E(Z^*|X)E(Z^*|X)' \tag{2.7}$$

M Step: Let $psi^{(t)}$ and $W^{(t)}$ be the *psi* and W at the t^{th} iteration.

$$psi^{(t+1)} = \frac{1}{n} diag\left\{XX' - W^{(t)}E(Z^*|X)X'\right\} \tag{2.8}$$

$$W^{(t+1)} = XE(Z^*|X)'E(Z^*Z^{*'}|X)^{-1} \tag{2.9}$$

The lasso error application will be a soft threshold type operation as follows:

$$W_{lasso}^{(t+1)} = sign(W^{(t+1)}) \left(|W^{(t+1)}| - L\right)_+ \tag{2.10}$$

For the L being a vector, the lasso error corresponding to the dataset will be used for the rows of W which correspond to the same dataset. The normalization step will be done by dividing each element, by the L_2 norm of the corresponding column vector. Calculate convergenceRate as

$$E = |E(Z^*|X)_{old} - E(Z^*|X|$$

$$if \quad E = [e_{ij}]_{(c-1)*n}, \quad then \quad convergenceRate = \max_{\substack{i=1,\cdots,(c-1) \\ j=1,\cdots,n}} e_{ij} \tag{2.11}$$

5. Repeat step 4 till *convergenceRate¡acceptableError* and *iteration* $\leq$ *maxIteration*.

The analysis of the time complexity starts from the finding of Eigen vectors and Eigen values. For finding the covariance matrix we will need a time of $\Theta(\mathcal{M}^2 n)$, which is followed by the Jacobi Eigen value finding algorithm involving Jacobi rotation, which takes a $O(\mathcal{M}^2)$ time. However being an iterative algorithm which usually converges at a high number of iteration for large matrices, increase the overall time greatly. So, Jacobian Eigen value finding takes $\Theta(k\mathcal{M}^2)$ time where $k \gg \mathcal{M}$, being a "large constant" value. The initialization step takes the calculation Σ, which involves a multiplication done in $\Theta(\mathcal{M}^2(c-1))$ time and an addition of $\Theta(\mathcal{M})$ time. The E step takes a time of $O(\mathcal{M}^3)$ dominated by the finding of the inverse of Σ, to update the

value of $E(Z^*|X)$ in equation (2.6). The multiplication of $E(Z^*|X)E(Z^*|X)'$ takes a time of $\Theta((c-1)^2 n)$ at the time of update of $E(Z^* Z^{*'}|X)$ in equation (2.7), while the multiplication $W'\Sigma^{-1}W$ takes a time of $O(\mathcal{M}^2(c-1))$, dominating the time needed for the step. Clearly update of ψ in the M step is overruled by the calculation of covariance matrix, which may be used from the previous, thus limiting the actual time to be $O(\mathcal{M}^2(c-1))$ corresponding to the the remaining multiplication. The update of W takes $O(\mathcal{M}n(c-1))$, followed by an $O(\mathcal{M}(c-1))$ time for the lasso error. The normalization will take $O(\mathcal{M}(c-1))$ time. So from the explanation it is observed that the iterative steps re dominated by an $O(\mathcal{M}^3)$ time, which is taken by the matrix inversion. If we assume the number of iteration to be k again, it will usually be $k \ll \mathcal{M}$, thus the complexity of the iterative steps remains as before.

After obtaining the transformed data by the above process, a clustering of samples is used to find the subtypes of cancer. A work flow is given in figure: 2.1.

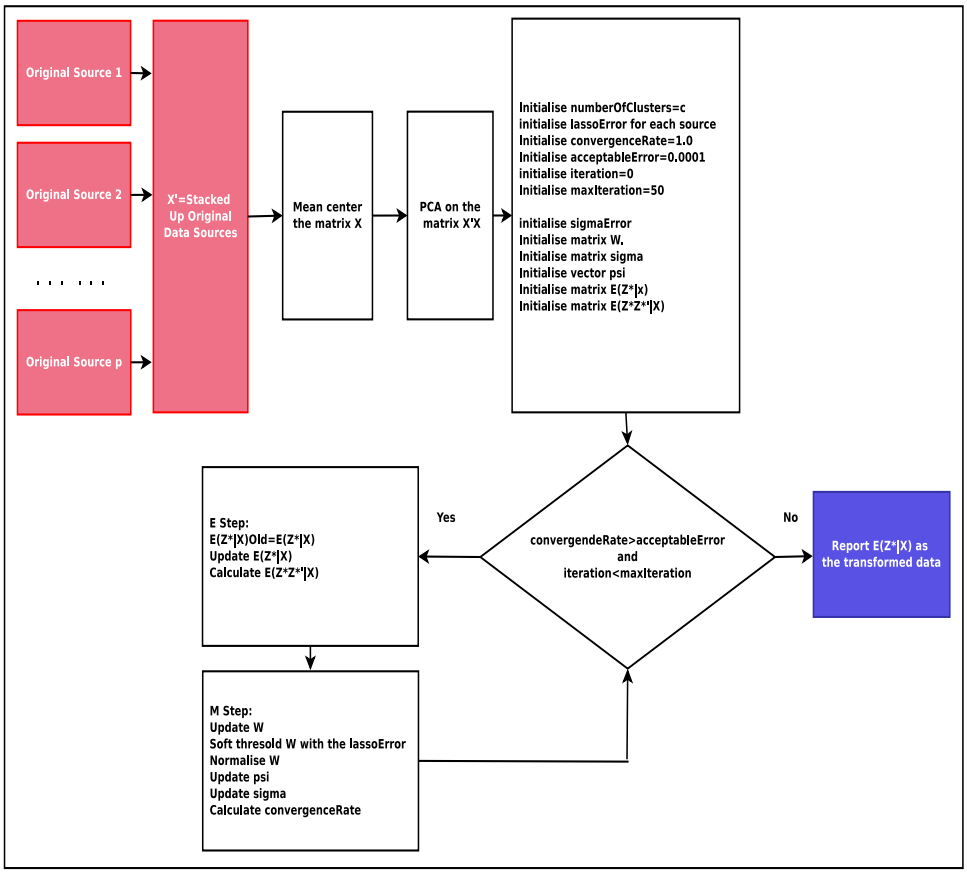

Figure 2.1: The work flow of the iCluster method

Chapter 3

Proposed Algorithm

The proposed method starts with the collection of different data sources for the same set of samples having a similar cancer. After feeding these datasets to iCluster alongside an appropiate L value (The choice of L value is important as one or more dimension of the transformed dataset may lead down to 0, for an unsuitable one, as demonstrated in the experments chapter). The iCluster ouput, in the form of the transformed data, is then fed to a initial center selection algorithm, followed by a RFCM clustering with the obtained initial centers. The evaluation of clustering solutions and optimisation of different parameters will follow next. The proposed algorithm follows a workflow as shown in figure: 3.1.

3.1 Initial Center Selection

A limitation of any c-means algorithm is that it can only achieve a local optimum solution that depends on the initial choice of the cluster prototypes. Consequently, computing resources may be wasted in that some initial centers get stuck in regions of the input space with a scarcity of data points and may therefore never have the chance to move to new locations where they are needed.

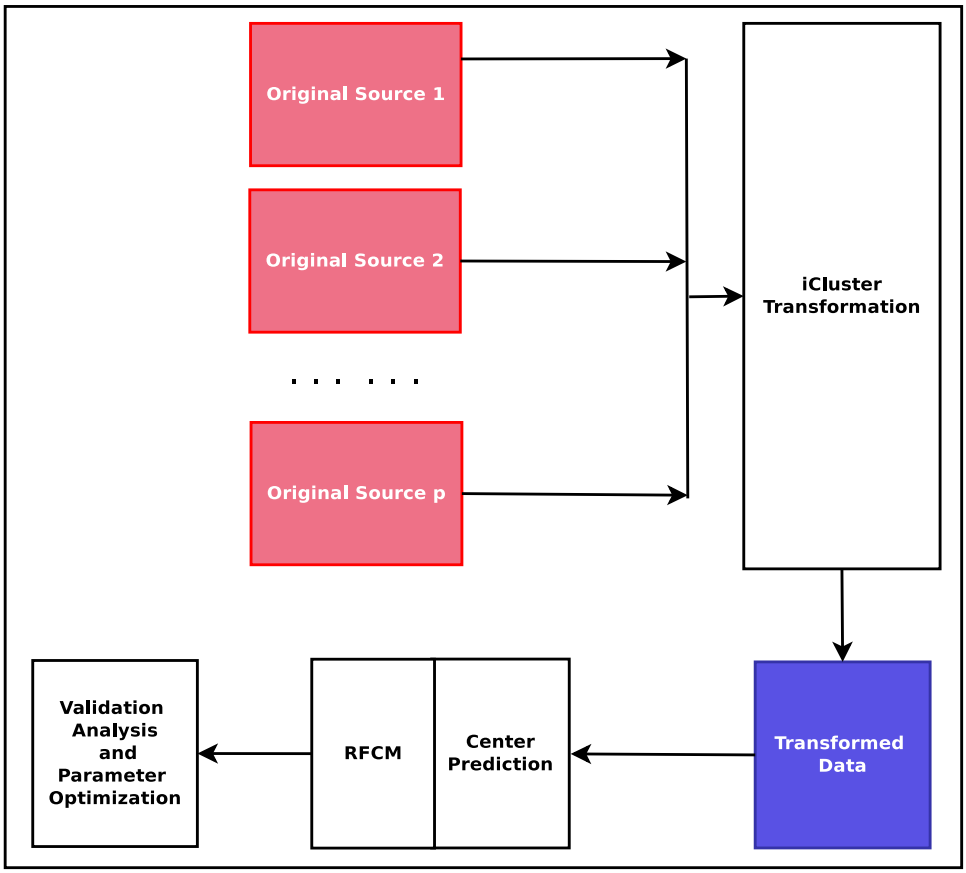

Figure 3.1: The workflow of the proposed method

Input: The dataset X of n samples and m features. The number of clusters c.

Output: The set of centers $V = \{v_1, v_2, \cdots v_c\}$.

1. For each sample x_i, calculate $d(x_i, x_j)$ between itself and the sample x_j, where $i, j = 1, 2, \cdots n$. where $d(x_i, x_j)$ is the distance measure calculated as follows.

$$d(x_i, x_j) = \sqrt{\left[\sum_{r=1}^{m} \left(\frac{x_{ir} - x_{jr}}{max_r - min_r}\right)^2\right]}$$

2. Calculate similarity score between two samples x_i and x_j as follows:

$$S(x_i, x_j) = \begin{cases} 1 & \text{if } d(x_i, x_j) \leq \gamma \\ 0 & \text{Otherwise} \end{cases}$$

3. For each sample x_i, calculate total number of similar samples of x_i as

$$N(x_i) = \sum_{j=1}^{n} S(x_i, x_j).$$

4. Sort n samples according to their values of $N(x_i)$ such that $N(x_1) > N(x_2) > \cdots > N(x_n)$.

5. If $N(x_i) > N(x_j)$ and $d(x_i, x_j) \leq \gamma$, then x_j cannot be considered as an initial cluster center, resulting in a reduced set of samples to be considered for c initial cluster centers v_i, $i = 1, 2, \cdots, c$.

The initial centers should be in the dense most region of the data, as the process of clustering actually tries to find those dense regions and return us the representatives for those regions. The initial center selection process helps this activity as it tries to identify those regions at the start, by selecting the sample having most number of close neighbours. Now none of it's neighbours

can be a good center as they belong to the same dense region, thus the algorithm discards them and looks for another point from the remaining, in the same way till c centers are found. Hence, the initialization method[11] helps to identify different dense regions present in the data set. The identified dense regions ultimately lead to discovering natural groups present in the data set. The whole approach is, therefore, data dependent. The value γ is an user given parameter and $0.51 < \gamma < 0.99$, whose value takes a major role over the performance. However an optimal value may be chosen by comparing cluster validity indexes for different values of γ.

3.2 Fuzzy C-Means and Rough Sets

This section presents the basic notions of fuzzy c-means and rough c-means. The rough-fuzzy c-means algorithm is developed based on these algorithms[10].

3.2.1 Fuzzy C-Means

Let $X = \{x_1, \cdots, x_j, \cdots, x_n\}$ be the set of n objects and the set of c centroids $V = \{v_1, \cdots, v_i, \cdots, v_c\}$, where $x_j \in \Re^m$ and $v_i \in \Re^m$. The fuzzy c-means provides a fuzzification of the hard c-means [1, 5]. It partitions X into c clusters by minimizing the objective function

$$J = \sum_{j=1}^{n} \sum_{i=1}^{c} (\mu_{ij})^{\acute{m}_1} ||x_j - v_i||^2 \tag{3.1}$$

where $1 \leq \acute{m}_1 < \infty$ is the fuzzifier, v_i is the ith centroid corresponding to cluster β_i, $\mu_{ij} \in [0, 1]$ is the probabilistic membership of the pattern x_j to cluster β_i, and $||.||$ is the distance norm, such that

$$v_i = \frac{1}{n_i} \sum_{j=1}^{n} (\mu_{ij})^{\acute{m}_1} x_j; \text{ where } n_i = \sum_{j=1}^{n} (\mu_{ij})^{\acute{m}_1} \tag{3.2}$$

$$\mu_{ij} = (\sum_{k=1}^{c} (\frac{d_{ij}}{d_{kj}})^{\frac{2}{\acute{m}_1 - 1}})^{-1};\ d_{ij}^2 = ||x_j - v_i||^2;\ \text{subject to}$$
$$\sum_{i=1}^{c} \mu_{ij} = 1, \forall j, 0 < \sum_{j=1}^{n} \mu_{ij} < n, \forall i. \tag{3.3}$$

The process begins by randomly choosing c objects as the centroids (means) of the c clusters. The memberships are calculated based on the relative distance of the object x_j to the centroids $\{v_i\}$ by Equation 3.3. After computing memberships of all the objects, the new centroids of the clusters are calculated as per Equation 3.2. The process stops when the centroids stabilize. That is, the centroids from the previous iteration are identical to those generated in the current iteration. The basic steps are outlined as follows:

1. Assign initial means v_i, $i = 1, 2, \cdots, c$. Choose values for $\acute{m}_1$ and threshold ϵ. Set iteration counter $t = 1$.

2. Compute memberships μ_{ij} by Equation 3.3 for c clusters and n objects.

3. Update mean (centroid) v_i by Equation 3.2.

4. Repeat steps 2 to 4, by incrementing t, until $|\mu_{ij}(t) - \mu_{ij}(t-1)| > \epsilon$.

In fuzzy c-means, the memberships of an object are inversely related to the relative distance of the object to the cluster centroids. In effect, it is very sensitive to noise and outliers. Also, from the standpoint of "compatibility with the centroid", the memberships of an object x_j in a cluster β_i should be determined solely by how close it is to the mean (centroid) v_i of the class, and should not be coupled with its similarity with respect to other classes.

To alleviate this problem, Krishnapuram and Keller [6, 7] introduced possibilistic c-means algorithm, where the objective function can be formulated as

$$J = \sum_{i=1}^{c} \sum_{j=1}^{n} (\nu_{ij})^{\acute{m}_2} ||x_j - v_i||^2 + \sum_{i=1}^{c} \eta_i \sum_{j=1}^{n} (1 - \nu_{ij})^{\acute{m}_2} \tag{3.4}$$

where $1 \leq \acute{m}_2 \leq \infty$ is the fuzzifier and η_i represents the scale parameter. The membership matrix ν generated by the possibilistic c-means is not a partition matrix in the sense that it does not satisfy the constraint

$$\sum_{i=1}^{c} \nu_{ij} = 1 \tag{3.5}$$

The update equation of ν_{ij} is given by

$$\nu_{ij} = \frac{1}{1+\mathrm{D}}; \quad \text{where} \quad \mathrm{D} = \left\{ \frac{||x_j - v_i||^2}{\eta_i} \right\}^{1/(\acute{m}_2 - 1)} \tag{3.6}$$

$$\text{subject to} \quad \nu_{ij} \in [0,1], \forall i, j; 0 < \sum_{j=1}^{n} \nu_{ij} \leq n, \forall i; \text{ and } \max_i \nu_{ij} > 0, \forall j.$$

The scale parameter η_i represents the zone of influence of the cluster β_i. The update equation for η_i is

$$\eta_i = \mathrm{K} \cdot \frac{\mathrm{P}}{\mathrm{Q}}; \text{ where } \mathrm{P} = \sum_{j=1}^{n} (\nu_{ij})^{\acute{m}_2} ||x_j - v_i||^2; \text{ and } \mathrm{Q} = \sum_{j=1}^{n} (\nu_{ij})^{\acute{m}_2} \tag{3.7}$$

Typically K is chosen to be 1. In each iteration, the updated value of ν_{ij} depends only on the similarity between the object x_j and the centroid v_i. The resulting partition of the data can be interpreted as a possibilistic partition, and the membership values may be interpreted as degrees of possibility of the objects belonging to the classes, i.e., the compatibilities of the objects with the means (centroids). The updating of the means proceeds exactly the same way as in the case of the fuzzy c-means algorithm.

3.2.2 Rough Sets

The theory of rough sets begins with the notion of an approximation space, which is a pair $< U, R >$, where U be a non-empty set (the universe of discourse) and R an equivalence relation on U, i.e., R is reflexive, symmetric, and transitive. The relation R decomposes the set U into disjoint classes in such a way that two elements x, y are in the same class iff $(x, y) \in R$. Let denote by U/R the quotient set of U by the relation R, and

$$U/R = \{X_1, X_2, \cdots, X_m\}$$

where X_i is an equivalence class of R, $i = 1, 2, \cdots, m$. If two elements x, y in U belong to the same equivalence class $X_i \in U/R$, we say that x and y are indistinguishable. The equivalence classes of R and the empty set $\emptyset$ are the elementary sets in the approximation space $< U, R >$. Given an arbitrary set $X \in 2^U$, in general it may not be possible to describe X precisely in $< U, R >$. One may characterize X by a pair of lower and upper approximations defined as follows [13]:

$$\underline{R}(X) = \bigcup_{X_i \subseteq X} X_i; \quad \overline{R}(X) = \bigcup_{X_i \cap X \neq \emptyset} X_i$$

That is, the lower approximation $\underline{R}(X)$ is the union of all the elementary sets which are subsets of X, and the upper approximation $\overline{R}(X)$ is the union of all the elementary sets which have a non-empty intersection with X. The interval $[\underline{R}(X), \overline{R}(X)]$ is the representation of an ordinary set X in the approximation space $< U, R >$ or simply called the rough set of X. The lower (resp., upper) approximation $\underline{R}(X)$ (resp., $\overline{R}(X)$) is interpreted as the collection of those elements of U that definitely (resp., possibly) belong to X. Further, we can define:

- a set $X \in 2^U$ is said to be definable (or exact) in $< U, R >$ iff $\underline{R}(X) = \overline{R}(X)$.

- for any $X, Y \in 2^U$, X is said to be roughly included in Y, denoted by $X \tilde{\subset} Y$, iff $\underline{R}(X) \subseteq \underline{R}(Y)$ and $\overline{R}(X) \subseteq \overline{R}(Y)$.
- X and Y is said to be roughly equal, denoted by $X \simeq_R Y$, in $< U, R >$ iff $\underline{R}(X) = \underline{R}(Y)$ and $\overline{R}(X) = \overline{R}(Y)$.

In [13], Pawlak discusses two numerical characterizations of imprecision of a subset X in the approximation space $< U, R >$: accuracy and roughness. Accuracy of X, denoted by $\alpha_R(X)$, is simply the ratio of the number of objects in its lower approximation to that in its upper approximation; namely

$$\alpha_R(X) = \frac{|\underline{R}(X)|}{|\overline{R}(X)|}$$

The roughness of X, denoted by $\rho_R(X)$, is defined by subtracting the accuracy from 1:

$$\rho_R(X) = 1 - \alpha_R(X) = 1 - \frac{|\underline{R}(X)|}{|\overline{R}(X)|}$$

Note that the lower the roughness of a subset, the better is its approximation. Further, the following observations are easily obtained:

1. As $\underline{R}(X) \subseteq X \subseteq \overline{R}(X)$, $0 \leq \rho_R(X) \leq 1$.
2. By convention, when $X = \emptyset$, $\underline{R}(X) = \overline{R}(X) = \emptyset$ and $\rho_R(X) = 0$.
3. $\rho_R(X) = 0$ if and only if X is definable in $< U, R >$.

3.3 Rough-Fuzzy C-Means Algorithm

Incorporating both fuzzy and rough sets, rough-fuzzy c-means or RFCM method, adds the concept of fuzzy membership of fuzzy sets, and lower and upper approximations of rough sets into c-means algorithm. While the membership of fuzzy sets enables efficient handling of overlapping partitions, the

rough sets deal with uncertainty, vagueness, and incompleteness in class definition.

3.3.1 Objective Function

RFCM partitions a set of n objects into c clusters by minimizing the objective function

$$J_{\mathrm{RF}} = \begin{cases} w \times \mathcal{A}_1 + \tilde{w} \times \mathcal{B}_1 & \text{if } \underline{A}(\beta_i) \neq \emptyset,\ B(\beta_i) \neq \emptyset \\ \mathcal{A}_1 & \text{if } \underline{A}(\beta_i) \neq \emptyset,\ B(\beta_i) = \emptyset \\ \mathcal{B}_1 & \text{if } \underline{A}(\beta_i) = \emptyset,\ B(\beta_i) \neq \emptyset \end{cases} \tag{3.8}$$

$$\mathcal{A}_1 = \sum_{i=1}^{c} \sum_{x_j \in \underline{A}(\beta_i)} (\mu_{ij})^{\acute{m}_1} ||x_j - v_i||^2; \quad \text{and } \mathcal{B}_1 = \sum_{i=1}^{c} \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{\acute{m}_1} ||x_j - v_i||^2$$

where the parameters w and $\tilde{w}$ $(= 1 - w)$ correspond to the relative importance of lower and boundary region. Note that, μ_{ij} has the same meaning of membership as that in fuzzy c-means.

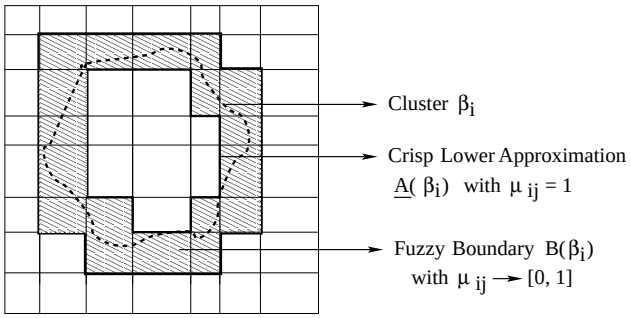

Figure 3.2: Rough-fuzzy c-means: cluster β_i is represented by crisp lower bound and fuzzy boundary

In RFCM, each cluster is represented by a centroid, a crisp lower approximation, and a fuzzy boundary (Figure: 3.2). The lower approximation influences the fuzziness of final partition. According to the definitions of lower

approximations and boundary of rough sets, if an object $x_j \in \underline{A}(\beta_i)$, then $x_j \notin \underline{A}(\beta_k), \forall k \neq i$, and $x_j \notin B(\beta_i), \forall i$. That is, the object x_j is contained in β_i definitely. Thus, the weights of the objects in lower approximation of a cluster should be independent of other centroids and clusters, and should not be coupled with their similarity with respect to other centroids. Also, the objects in lower approximation of a cluster should have similar influence on the corresponding centroid and cluster. Whereas, if $x_j \in B(\beta_i)$, then the object x_j possibly belongs to β_i and potentially belongs to another cluster. Hence, the objects in boundary regions should have different influence on the centroids and clusters. So, in RFCM, the membership values of objects in lower approximation are $\mu_{ij} = 1$, while those in boundary region are the same as fuzzy c-means (Equation 3.3). In other word, the RFCM first partitions the data into two classes - lower approximation and boundary. Only the objects in boundary are fuzzified. Thus, $\mathcal{A}_1$ reduces to

$$\mathcal{A}_1 = \sum_{i=1}^{c} \sum_{x_j \in \underline{A}(\beta_i)} ||x_j - v_i||^2$$

and $\mathcal{B}_1$ has the same expression as that in Equation 3.8.

3.3.2 Cluster Prototypes

The new centroid is calculated based on the weighting average of the crisp lower approximation and fuzzy boundary. Computation of the centroid is modified to include the effects of both fuzzy memberships and lower and upper bounds. The modified centroid calculation for RFCM is obtained by solving Equation 3.8 with respect to v_i:

$$v_i^{\mathrm{RF}} = \begin{cases} w \times \mathcal{C}_1 + \tilde{w} \times \mathcal{D}_1 & \text{if } \underline{A}(\beta_i) \neq \emptyset,\ B(\beta_i) \neq \emptyset \\ \mathcal{C}_1 & \text{if } \underline{A}(\beta_i) \neq \emptyset,\ B(\beta_i) = \emptyset \\ \mathcal{D}_1 & \text{if } \underline{A}(\beta_i) = \emptyset,\ B(\beta_i) \neq \emptyset \end{cases} \tag{3.9}$$

$$\mathcal{C}_1 = \frac{1}{|\underline{A}(\beta_i)|} \sum_{x_j \in \underline{A}(\beta_i)} x_j; \quad \text{and } \mathcal{D}_1 = \frac{1}{n_i} \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{\acute{m}_1} x_j;$$

$$\text{where } \; n_i = \sum_{x_j \in B(\beta_i)} (\mu_{ij})^{\acute{m}_1}$$

$|\underline{A}(\beta_i)|$ represents the cardinality of $\underline{A}(\beta_i)$.

Thus, the cluster prototypes (centroids) depend on the parameters w and $\tilde{w}$, and fuzzifier $\acute{m}_1$ rule their relative influence. The correlated influence of these parameters and fuzzifier, makes it somewhat difficult to determine their optimal values. Since the objects lying in lower approximation definitely belong to a cluster, they are assigned a higher weight w compared to $\tilde{w}$ of the objects lying in boundary region. Hence, for RFCM, the values are given by $0 < \tilde{w} < w < 1$.

Chapter 4

Experiments and Results

We will first start by giving a brief descriptions of the datasets used and definition of the cluster validity indexes. We have used three parameters, L as the lasso error in iCluster, γ in the center selection and ω in the RFCM, alongside the number of clusters. We will vary those parameter values in the allowable range, and present the results obtained from the different clustering algorithms (k-means, FCM and RFCM) for the different parameter combination, and show RFCM to be working better. We will also compare the result by the integration of different data sources combination and study their effects. We will follow up with a detailed analysis of the nature of variation of the indexes with the variation of parameter values. We will also demonstrate the variation of performance of individual and different combinations of the datasets.

4.1 Description of Datasets

We used two cancer datasets, both available freely with the R package "iCluster".

Breast Cancer (BC): This dataset contains the copy number variation (CNV) and gene expression (GE) having 354 features each, for 41 samples.

Among those 41 samples 37 are known to be from tumor while the remaining 4 are cell lines. From the experiments of *Shen et. al.*[16] we know the dataset performs best clustering with number of clusters to be 4 and L to be 0.2.

Glioblastoma Multiforme (GBM): This dataset contains three data sources, copy number variation, gene expression and DNA methylation (MET), having 1599, 1740 and 1515 number of features respectively for 55 samples. Though the number of subtypes for GBM is found to be 4[19], another experiments from *Shen et. al.*[15] found only 3 of them.

4.2 Cluster Validation

We will evaluate our obtained clusters by the help of four cluster validity indices, namely POD[16], Dunn index (DNI)[17], Davies Bouldin index (DBI)[17] and silhouette index (SHI)[17].

POD Index: Let us arrange the sample point of the dataset X, in such a way that samples from same clusters resides in following rows of the data matrix. We now calculate $B = XX^T$ and standardize the elements of B by having $B_{ij} = B_{ij}/\sqrt{B_{ii}B_{jj}}$, where $i, j = 1, \cdots n$, we also impose a non-negativity by making the negative terms 0. We assume the points in the clusters to most related and least related with other cluster members. Thus B should ideally be a block diagonal matrix, let us take an ideal block diagonal matrix $\bar{B}$, where if x_i and x_j belongs to same cluster then $\bar{B}_{ij} = 1$ or it is 0 otherwise. We calculate the deviation by $\mathcal{D} = \sum_{i=1}^{n}\sum_{j=1}^{n}|B_{ij} - \bar{B}_{ij}|$. We define $POD = \mathcal{D}/n^2$. Clearly $0 < POD < 1$. The less deviation from the diagonal block structure, indicates better clustering' thus a low value of POD indicates a good cluster solution.

For the following indexes the distance between two m dimensional point

x and y are calculated as

$$D(x,y) = \frac{1}{m}\sqrt{\sum_{j=1}^{m}(x_j - y_j)^2}$$

We also assume there are c clusters, $C_1, C_2, \cdots C_c$ having means as $centroid_i$ for the i^{th} cluster. The distance between two clusters say C_i and C_j is defined as:

$$D(C_i,C_j) = \frac{1}{m}\sqrt{\sum_{k=1}^{m}(centroid_{ik} - centroid_{jk})^2}$$

The width of a cluster C_i having n_i points is defined as:

$$W(C_i) = \frac{1}{n_i}\sum_{x \in C_i} D(x, centroid_i)$$

Dunn Index: Let us call the maximum width between all the clusters as W_{max}=max$_{i=1,2,\cdots c}\, W(C_i)$. We define Dunn index as

$$Dunn = \min_{i=1,\cdots c}\left\{\min_{j=i+1,\cdots c}\left(\frac{D(C_i,C_j)}{W_{max}}\right)\right\}$$

Clearly as we indicate a good cluster as separate from others as possible and as "compact" as possible, thus the width of a good cluster goes to be small while the distance from other cluster being greater. This indicates the more the value of Dunn index the better clustering has been obtained.

Davies Bouldin Index: The Davies Bouldin or DB index is defined as:

$$DB = \frac{1}{c}\sum_{i=1}^{c}\max_{i \neq j}\left\{\frac{W(C_i)+W(C_j)}{D(C_i,C_j)}\right\}$$

DB index minimizes the width of a cluster, and maximizes the distance between two clusters, resulting in a lower value of DB to indicate a good clus-

tering. Thus we try to minimize the DB index to obtain a good cluster solution.

Silhouette Index: The value of γ in prediction of center plays a major role over the performance of the clustering algorithm. The optimum value of γ, say γ^* can be determined by running the clustering algorithm over all the possible values of γ between 0.51 and 0.99 with an increment of 0.01, and compare the result with the help of Silhouette index as shown by *Maji et, al.(2013)*[11]. For each point $x_i \in X$, where $i = 1, \cdots n$ we define a_i as the average distance between itself and all the other points, member of the same cluster. Let the point x_i belongs to cluster C_j, having n_j members.

$$a_i = \frac{1}{n_j - 1} \sum_{\substack{y \in C_j \\ y \neq x_i}} D(x_i, y)$$

We also define b_i for the i^{th} to be the minimum distance between the point and the nearest cluster other than C_j.

$$b_i = \min_{\substack{k=1,\cdots c \\ k \neq i}} \left\{ \frac{1}{n_k} \sum_{y \in C_k} D(x_i, y) \right\}$$

We define the silhouette width of the point x_i to be s_i as:

$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)}$$

Clearly $s_i \in [-1, 1]$, a value of s_i close to 1, means that the point has significantly more distance from other clusters and quiet similar to the other members of its own clusters, or b_i value overwhelms a_i value indicating a good clustering of the point. Similarly s_i value closer to -1 indicates that it is distant from its own cluster member, than other clusters, thus clustered wrongly. For a boundary point s_i is close to 0. For a cluster C_j the silhouette $s(C_j)$ is calculated as the average silhouette width of all its member points.

We define silhouette index as the average of all the silhouette of the clusters, or $SH = \frac{1}{c}\sum_{j=1}^{c} s(C_j)$. Clearly as the greater value of SH indicates a good clustering solution, thus we need to maximize that over the runs of clustering algorithm. We will select the γ^* producing the best SH value.

4.3 Parameter Optimization

For the purpose of parameter optimization, all the parameters were exhaustively searched in their spaces. We used the number of clusters to be 3, 4 and 5. The value of L was between 0 and 0.5, incremented by the step of 0.05. The value of ω varied between 0.55 to 0.95, incremented by the step of 0.05, including 0.51 and 0.99. The value of γ varied between 0.51 to 0.99, with the step of increment 0.01. For each possible combination we obtain the clustering solution and four cluster validity index values. We will now present the best values of those indexes and the corresponding parameter combination.

4.3.1 Result on Breast Cancer (BC) Dataset

For the BC dataset we only had the chance of exploring the capacity of identifying subtypes for three cases, Two for the individual data sources, and one the combined. We give the summary as follows:

From the above table we can see that most of the times a higher value of ω, to be specific 0.99 is giving the best result. While for the POD index only once we obtained the best at the minimum ω. For the CNV dataset, all the indexes reached the best with number of clusters 3, while for the other two datasets we have a confusion. For the integrated dataset, two indexes became best with number of clusters being 4, as the previous studies were made with only POD index, and our result does match for that index, thus the confusion can be due to the nature of quantifying the cluster "goodness". POD uses a different measure of distance, which is similar to the angular

Table 4.1: The best index values for different parameter combination for BC dataset

Dataset	Validity Index	No. of Clusters	L	ω	γ	Value
CNV	DBI	3	0.1	0.99	0.66	0.388084
	DNI	3	0.05	0.99	0.71	3.257866
	POD	3	0.10	0.90	0.71	0.130892
	SHI	3	0.15	0.99	0.71	0.577102
GE	DBI	5	0.1	0.99	0.97	0.389353
	DNI	3	0	0.99	0.72	3.360559
	POD	5	0.25	0.51	0.75	0.139975
	SHI	3	0.2	0.8	0.64	0.559596
CNV, GE	DBI	5	0.3	0.99	0.97	0.362967
	DNI	4	0.05	0.99	0.68	3.22024
	POD	4	0.2	0.99	0.85	0.129092
	SHI	3	0	0.99	0.69	0.576171

distance between two vectors, in different in nature from the other three Euclidean distance based index. Among three cases, the integrated dataset outperforms the other for three indexes, POD, DBI and SHI, while DNI is best for the GE dataset.

4.3.2 Result on GBM Dataset

For GBM dataset we get to explore 7 cases formed by the different combination of data sources. We summarize the result as follows:

The number of clusters is found to be similar in every cases and for every index. The number of clusters for the dataset is matched with the reported. Now for the value of ω, is most of the case found to be 0.99, for the CNV, MET dataset for three indexes it has been 0.95, and for two case for the POD index, it became 0.51, similar to what we observed in the BC dataset. Also unlike the BC dataset, the value of L in most of cases 0 or 0.5 and in once it has been 0.1 for the DNI index. The most interesting fact to note is the performance of CNV dataset, it clearly outperformed every other integrated

Table 4.2: The best index values for different parameter combination for BC dataset

Dataset	Validity Index	No. of Clusters	L	ω	γ	Value
CNV	DBI	3	0	0.99	0.68	0.285623
	DNI	3	0	0.99	0.85	4.968927
	POD	3	0	0.99	0.99	0.088481
	SHI	3	0	0.99	0.83	0.734613
MET	DBI	3	0	0.99	0.64	0.515081
	DNI	3	0	0.99	0.70	2.961014
	POD	3	0	0.51	0.94	0.092258
	SHI	3	0	0.99	0.52	0.586846
GE	DBI	3	0	0.99	0.64	0.572217
	DNI	3	0	0.99	0.75	3.151997
	POD	3	0.05	0.99	0.93	0.117447
	SHI	3	0	0.99	0.64	0.521658
CNV, GE	DBI	3	0	0.99	0.73	0.522468
	DNI	3	0.1	0.99	0.93	2.997192
	POD	3	0.05	0.99	0.93	0.117447
	SHI	3	0	0.99	0.73	0.506221
CNV, MET	DBI	3	0	0.95	0.99	0.572285
	DNI	3	0	0.95	0.99	2.718833
	POD	3	0	0.95	0.98	0.100037
	SHI	3	0	0.99	0.97	0.547987
GE, MET	DBI	3	0	0.99	0.98	0.552817
	DNI	3	0	0.99	0.72	3.125528
	POD	3	0.05	0.51	0.98	0.122524
	SHI	3	0	0.99	0.98	0.529917
CNV, GE, MET	DBI	3	0	0.99	0.82	0.566246
	DNI	3	0.05	0.99	0.66	3.133784
	POD	3	0	0.99	0.68	0.11039
	SHI	3	0	0.99	0.98	0.529649

and individual data sources in all of the four index values. This indicates that integration not always guarantees a better result, the subtypes may be caused by a single data source's variation, in that case integration will only increase the noise rather than improving the subtype classification capacity,

what can be observed here.

4.3.3 Effect of Variation of the Parameters

For the BC dataset, we took the integrated data, i.e. CNV, GE, what performed better than the others. For each index we took the number of clusters what gave the best result i.e. 4 for POD and DNI, 5 for DBI and 3 for SHI, for all our following experiments this was kept fixed. The parameter L and ω was then varied while the index values corresponds to best value obtained over all γ for the combination of L and ω. This is demonstrated in the figure: 4.1.

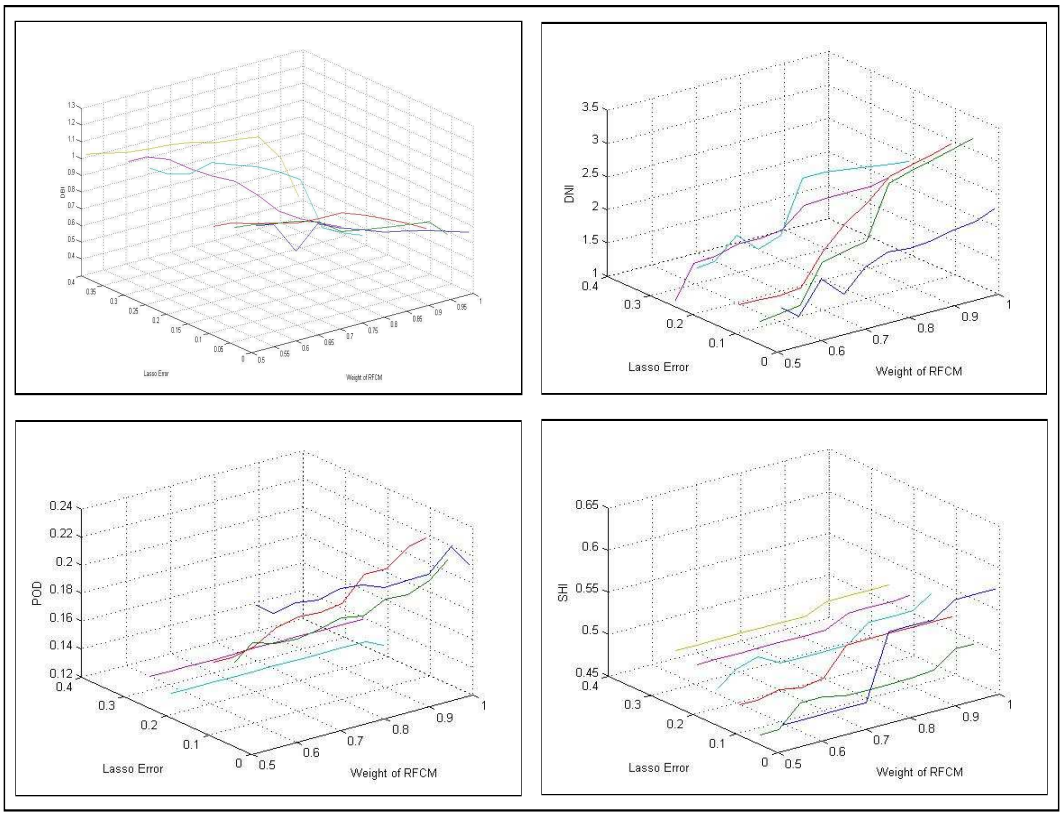

Figure 4.1: Effect of variation of L and ω on different index

To show the effect on indexes due to the variation of L and γ, we keep the ω as 0.99. This is demonstrated in the figure: 4.2.

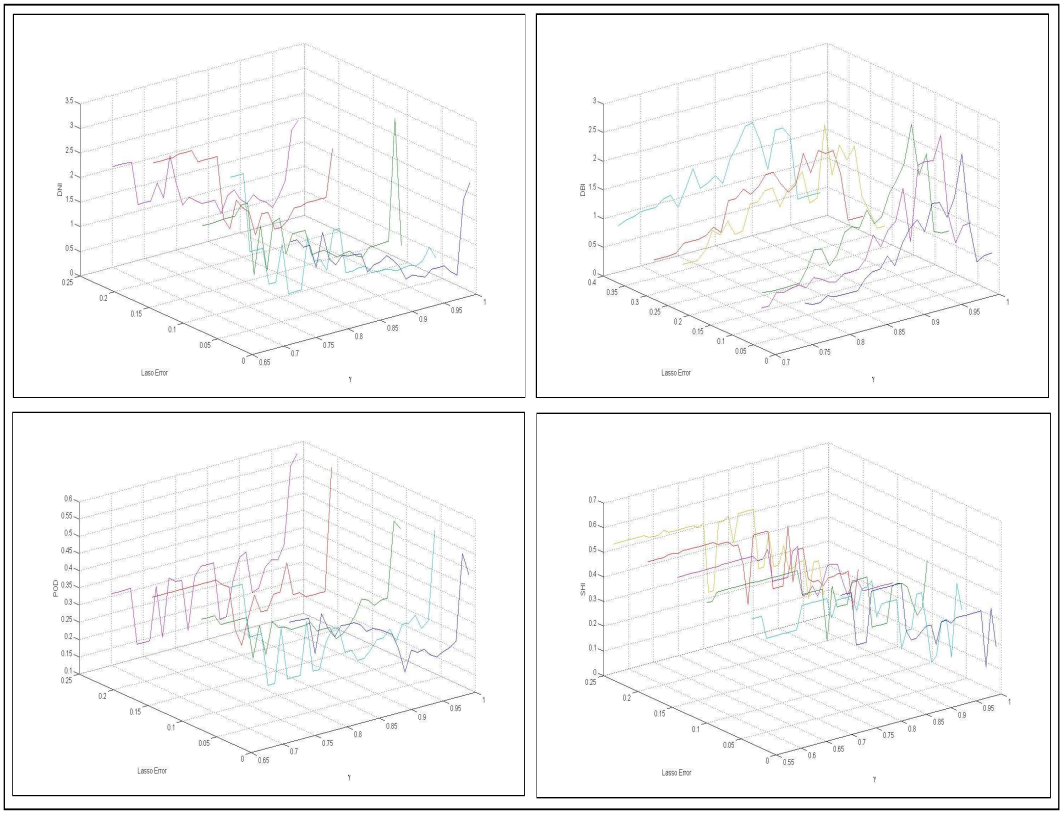

Figure 4.2: Effect of variation of L and γ on different index

To show the effect on indexes due to the variation of ω and γ we keep the number of clusters and L fixed to the optimal value for which that index gave the best result. This is demonstrated in the figure: 4.3.

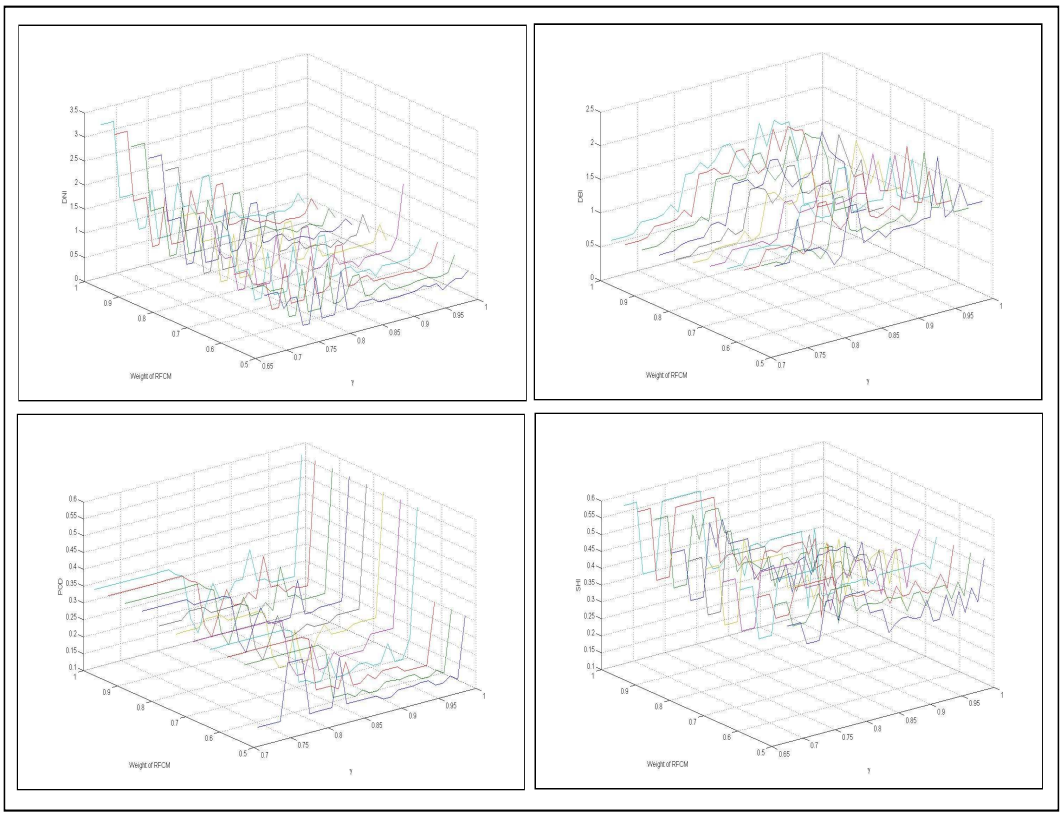

Figure 4.3: Effect of variation of ω and γ on different index

4.4 Comparison with Other Clustering Techniques

To show that the replacement of k-means by RFCM is useful for improving the result, we now present a comparison among hard c-means (HCM), FCM and RFCM. We have the optimal value of number of clusters, L, and γ for which a index gave the best for a dataset from previous tables. We will use that same parameter combination for each of the algorithms and only for RFCM will use the additional ω for which the best result has been obtained. We present the comparison for BC and GBM in the following tables.

Table 4.3: The comparison of algorithms for BC data

Dataset	Validity Index	HCM	FCM	RFCM
CNV	DBI	0.418929	0.65027	0.388084
	DNI	1.569145	1.365277	3.257866
	POD	0.130892	0.130892	0.130892
	SHI	0.563791	0.563791	0.577102
GE	DBI	0.764016	1.227905	0.389353
	DNI	2,618405	1.985616	3.360559
	POD	0.174308	0.146220	0.139975
	SHI	0.559596	0.525333	0.559596
CNV, GE	DBI	0.373214	1.408637	0.362967
	DNI	2.879902	0.967837	3.22024
	POD	0.149589	0.149589	0.129092
	SHI	0.559443	0.441145	0.576171

From the comparison table of BC it can be seen that RFCM outperforms both the HCM and FCM for the optimized parameter combination and similar initial center. Once in GE dataset the SHI is similar to HCM, which again happened for the POD of the CNV dataset. FCM performed worse than HCM in all three cases.

From the comparison table of GBM, RFCM has a significant improvement over the other two algorithms in terms of DNI and DBI, in some cases of POD

Table 4.4: The comparison of algorithms for GBM data

Dataset	Validity Index	HCM	FCM	RFCM
CNV	DBI	0.302500	0.345979	0.285623
	DNI	4.602448	4.570306	4.968927
	POD	0.139568	0.159241	0.088481
	SHI	0.734613	0.734613	0.734613
GE	DBI	0.668243	0.793733	0.572217
	DNI	2.604013	2.207558	3.151997
	POD	0.132512	0.163877	0.117447
	SHI	0.515104	0.451447	0.521658
MET	DBI	0.562117	0.557290	0.515081
	DNI	2.379840	2.755370	2.961014
	POD	0.114521	0.107314	0.092258
	SHI	0.542228	0.556070	0.586846
CNV, GE	DBI	0.621155	0.624608	0.522468
	DNI	2.095124	1.657201	2.997192
	POD	0.132513	0.163877	0.117447
	SHI	0.503357	0.503357	0.506221
CNV, MET	DBI	0.610234	0.613957	0.572285
	DNI	2.326038	2.233771	2.718833
	POD	0.100037	0.104307	0.100037
	SHI	0.547883	0.547883	0.547987
GE, MET	DBI	0.615283	0.621810	0.552817
	DNI	2.661776	2.687317	3.125528
	POD	0.136540	0.139893	0.122524
	SHI	0.521870	0.513956	0.529917
CNV, GE, MET	DBI	0.621281	0.615655	0.566246
	DNI	1.870187	1.869240	3.133784
	POD	0.110389	0.115227	0.11039
	SHI	0.523357	0.523357	0.529649

and SHI it has similar index values with the HCM, beside improvement in most cases. In CNV, GE, MET integrated dataset FCM performed better than HCM in terms of DBI, alongside in CNV, MET and GE, MET, in terms of DNI, in cases it has equal SHI value with the HCM.

Chapter 5

Conclusion and Future Work

Keeping in mind the importance of integrative analysis of genomic data sources, to discover richer information, (in our case classify cancer patients based on disease subtypes, for diagnosis and targeted treatment), this study addressed the multiple issues regarding the topic (like preserving individual significance and incorporating interaction information) and focused on improving performance of a popular integration scheme. The study showed that not only a better result can be obtained by modifying the work flow of the existing method, but also provided results evaluating the importance of integration. The study introduced RFCM clustering algorithm in the scene, what is more suitable for the overlapped, and incomplete natured real data. Alongside it also used an intelligent initial center selection algorithm inspired by the density based clustering approaches. The study used two publicly available datasets, each containing multiple data sources. The result section are targeted to support the claim of improvement of clustering performance by RFCM and demonstrate the effect and classification capacity of different data source combinations.

Analyzing the biological significance of the findings is the first future scope of this study. The relation between the members in the identified clusters can be explained in form of common genetic alterations. These

cluster specific alterations are the cause of generation of subtypes. Both of the dataset selected some genes for each sources based on a variance criteria, a study should support the importance of these selected genes by finding their regular functions, relation with the disease, or how their alteration results into cancer. Secondly, we used a single method of integration in our study, where a joint dataset was used for the clustering. A study can be made by comparing the performance of this method with the performance of clustering the dataset formed by individual dataset's cluster, to obtain the final clustering result. Thirdly the study used only four types of data sources, these days more new sources like somantic mutation or exon information are becoming available, how to integrate those data sources with the existing popluar ones is still under research.

Bibliography

[1] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithm*. New York: Plenum, 1981.

[2] George A. Calin and Carlo M. Croce. MicroRNA signatures in human cancers. *Nature Reviews, Cancer*, 6:857–866, 2006.

[3] Chris Ding and Xiaofeng He. K-means Clustering via Principal Component Analysis. In *ICML, ACM International Conference Proceeding Series*, volume 69, Alberta, Canada, 2004. ACM Bunff.

[4] Pan Du, Xiao Zhang, Chiang-Ching, Huang Nadereh Jafari, Warren A Kibbe1, Lifang Hou, and Simon M Lin. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(587), 2010.

[5] J. C. Dunn. A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact, Well-Separated Clusters. *Journal of Cybernetics*, 3:32–57, 1974.

[6] R. Krishnapuram and J. M. Keller. A Possibilistic Approach to Clustering. *IEEE Transactions on Fuzzy Systems*, 1(2):98–110, 1993.

[7] R. Krishnapuram and J. M. Keller. The Possibilistic C-Means Algorithm: Insights and Recommendations. *IEEE Transactions on Fuzzy Systems*, 4(3):385–393, 1996.

[8] Hyunju Lee, Sek Won Kong, and Peter J. Park. Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. *Bioinformatics*, 24(7):889–896, 2008.

[9] Eric F. Lock, Katherine A. Hoadley, J. S. Marron, and Andrew B. Nobel. Joint and Individual Variation Explained (JIVE) for Integrated Analysis of Multiple Data Types. *The Annals of Applied Statistics*, 7(1):523–542, 2013.

[10] Pradipta Maji and Sankar K. Pal. RFCM: A Hybrid Clustering Algorithm Using Rough and Fuzzy Sets. *Fundamenta Informaticae*, 80:475–496, 2007.

[11] Pradipta Maji and Sushmita Paul. Rough-Fuzzy Clustering for Grouping Functionally Similar Genes from Microarray Data. *IEEE/ACM Transactions On Computational Biology and Bioinformatics*, 10(2):286–299, 2013.

[12] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70, 2012.

[13] Z. Pawlak. *Rough Sets: Theoretical Aspects of Resoning About Data*. Kluwer, Dordrecht, The Netherlands, 1991.

[14] Li-Xuan Qin. An Integrative Analysis of microRNA and mRNA Expression—A Case Study. *Cancer informatics*, 6:369–379, 2008.

[15] Ronglai Shen, Qianxing Mo, Nikolaus Schult, Venkatraman E. Seshan, Adam B. Olshen, Jason Huse, Marc Ladanyi, and Chris Sander. Integrative Subtype Discovery in Glioblastoma Using iCluster. *Plos One*, 2012(4):1–9, 2012.

[16] Ronglai Shen, Adam B. Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model

with application to breast and lung cancer subtype analysis. *Oxford Bioinformatics*, 25(22):2906–2912, 2009.

[17] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition*. Academic Press Inc, 4 edition, 2009.

[18] Michael E. Tipping and Chris M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society*, 61:611–622, 1999.

[19] Roel G.W. Verhaak, Katherine A. Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D. Wilkerson, C. Ryan Miller, Li Ding, Todd Golub, Jill P. Mesirov, Gabriele Alexe, Michael Lawrence, Michael O'Kelly, Pablo Tamayo, Barbara A. Weir, Stacey Gabriel, Wendy Winckler, Supriya Gupta, Lakshmi Jakkula, Heidi S. Feiler, J. Graeme Hodgson, C. David James, Jann N. Sarkaria, Cameron Brennan, Ari Kahn, Paul T. Spellman, Richard K. Wilson, Terence P. Speed, Joe W. Gray, Matthew Meyerson, Gad Getz, Charles M. Perou, and D. Neil Hayes. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1):98–110, 2010.

[20] Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu. Spectral Relaxation for K-means Clustering. In *Neural Information Processing Systems (NIPS 2001)*, volume 14, pages 1057–1064, Vancouver, Canada, 2001. MIT Press.

[21] Feng Zhang, Wenli Gu, Matthew E. Hurles, and James R. Lupski. Copy Number Variation in Human Health, Disease,and Evolution. *Annual Review of Genomics and Human Genetics*, 10:451–481, 2009.

[22] Shihua Zhang, Chun-Chi Liu, Wenyuan Li1, Hui Shen, Peter W. Laird, and Xianghong Jasmine Zhou. Discovery of multi-dimensional modules

by integrative analysis of cancer genomic data. *Nuclic acids research*, 40(19):9379—-9391, 2012.