Sampled Clustering on Phenotypic Data of Plant

MS(Research) Thesis

By

Mithun Singh

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

January 2024

Sampled Clustering on Phenotypic Data of Plant

A THESIS

submitted to the

INDIAN INSTITUTE OF TECHNOLOGY INDORE

in partial fulfillment of the requirements for

the award of the degree

of

MS(Research)

By

Mithun Singh

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

January 2024

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Sampled Clustering on Phenotypic Data of Plant** in the partial fulfillment of the requirements for the award of the degree of **MS(Research)** and submitted in the **Department of Computer Science and Engineering, Indian Institute of Technology Indore,** is an authentic record of my own work carried out during the time period from August 2021 to January 2024.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

08/07/2024

Signature of the Student with Date

(Mithun Singh)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

08/07/2024

Signature of Thesis Supervisor with Date

(Prof. Kapil Ahuja)

Mithun Singh has successfully given his MS(Research) Oral Examination held on _____

Signature of Chairperson, OEB Date:	Signature of External Examiner Date:	Signature of Thesis Supervisor Date:
Signature of PSPC Member #1 Date:	Signature of PSPC Member #2 Date:	Signature of Convener, DPGC Date:
Signature of Head of Discipline Date:		

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my heartfelt gratitude to a number of persons who in one or the other way contributed by making this time as learnable, enjoyable, and bearable. First, I would like to thank my supervisor **Prof. Kapil Ahuja**, a constant source of inspiration during my work. Without his constant guidance and research directions, this research work could not be completed. His continuous support and encouragement have motivated me to remain streamlined in my research work.

I am thankful to **Prof. Aruna Tiwari** and **Prof. Ram Bilas Pachori**, my PSPC members for taking out some valuable time to evaluate my progress throughout the course. Their good comments and suggestions helped me to improve my work at various stages. I am also grateful to **DPGC** and **HOD** of Computer Science and Engineering for their help and support.

My sincere acknowledgment and respect to **Prof. Suhas Joshi** , Director, Indian Institute of Technology Indore for providing me the opportunity to explore my research capabilities at Indian Institute of Technology Indore.

I would like to express my heartfelt respect to my parents for the love, care, and support they have provided to me throughout my life.

Finally, I am thankful to all who directly or indirectly contributed, helped, and supported me.

Mithun Singh

Abstract

Clustering species of the same plant into different groups is an important step in developing new species of the concerned plant. Phenotypic (or Physical) characteristics of plant species are commonly used to perform clustering. Hierarchical Clustering (HC) is popularly used for this task, and this algorithm suffers from low accuracy and high computational complexity. In one of the recent works [1], the authors used the standard Spectral Clustering (SC) algorithm to improve the clustering accuracy. Further, they used Pivotal Sampling, a probabilistic algorithm, to reduce the cost of clustering. They demonstrated the efficacy of their algorithm on soybean species.

Use of the standard SC algorithm is not ideal. Based upon spectral graph theory and the involved Cheeger's inequality, we propose a novel base "a" SC instead. We further improve our algorithm by integrating the technique of local scaling (instead of global one) as proposed in [2]. We perform experiments on 1865 rice species. Using eigenvalue analysis, we argue that both our algorithms should lead to a better clustering. We further demonstrate that the standard SC is 49.86% better than HC while our base "a" SC and base "a" locally scaled SC are 64.93% and 66.33% better than HC, respectively. We also demonstrate that our latter clustering algorithm is better than another popular clustering technique of Gaussian Mixture Model (22.05% better).

In the Pivotal Sampling algorithm, the most crucial aspect is computing the inclusion probability of every unit. In the earlier application of this sampling to plant phenotypic data, this probability was taken to be inversely proportional to the deviation (sum of difference between specie's characteristic values from their respective base values). Earlier, a maximum function was used to find the base values that did not capture the problem behaviour well. We now propose use of a median function that is more intuitive. Again, we perform experiments on 1865 rice species. Using a statistical analysis, we show that this choice leads to better sampling. We also demonstrate that a median function works better in not just Pivotal Sampling but another popular probabilistic sampling (Poisson Sampling), and that Pivotal is slightly better than Poisson. Finally, we demonstrate that the standard SC with maximum based

Pivotal Sampling is about 50.10% better than HC while our base "a" locally scaled SC with median based Pivotal Sampling is 61.24% better than HC.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

SC Spectral Clustering

HC Hierarchical Clustering

GMM Gaussian Mixture Model

KNN K-Nearest Neighbour

UPGMA Unweighted Pair Group Method with Arithmetic Mean

PT Population Total

HT Horvitz-Thompson

Chapter 1

Introduction

Phenotypic characteristics (or physical characteristics) of plant species are often used in clustering them into separate categories [3, 4]. This is done so that plant species from different categories (or diverse plant species) could be selectively chosen for developing new species having better characteristics [5] (or called breeding).

Hierarchical Clustering (HC) is one of the most commonly used clustering algorithms in this domain [6]. This algorithm suffers from low accuracy issues and is also computationally expensive. In one of the recent works [1], authors used the standard Spectral Clustering (SC), which is considered to be one of the most accurate clustering algorithms. Since SC, like HC, is computationally costly, authors sampled the plant species before clustering. Pivotal sampling, which is probability based, was used due to its high accuracy and low cost (requires a single pass through the data). Pivotal sampling has the added advantage of maintaining independence from the data's order or location. The authors demonstrated the usefulness of their algorithm on soybean species.

The standard SC can be improved further. In this algorithm, one of the crucial step is building the similarity matrix. If $d_{p_1p_2}$ denotes the distance between two units p_1 and p_2, then the element $W_{p_1p_2}$ of the similarity matrix W should be inversely proportional to $d_{p_1p_2}$ (because larger the distance between two units, smaller should be the similarity between them). In the earlier work [1], authors had taken $W_{p_1p_2}$ to be the most commonly used choice of $e^{-\frac{d_{p_1p_2}}{2\sigma^2}}$, where σ defines the decay of the distance.

Based upon spectral graph theory, specifically the Cheeger's inequality, in this work we propose a base "a" exponential function instead of a natural exponential.

Having a global decay factor (σ) as above is not ideal. In our implementation, we use the concept of local scaling ([2]), which according to the theory works better. That is, $W_{p_1p_2}$ is taken as $a^{-\frac{d_{p_1p_2}}{\sigma_{p_1}\sigma_{p_2}}}$, where σ_{p_1} and σ_{p_2} are computed via a K-nearest neighbour (KNN) search.

We show the benefit of the above two algorithms with experiments on 1865 rice species. We initially justify both the above choices with an eigenvalue analysis. Next we show that the standard SC is about 49.86% better than HC while our base "a" SC and base "a" locally scaled SC are 64.93% and 66.33% better than HC, respectively. We also demonstrate that our latter clustering algorithm gives better results than another popular clustering technique of Gaussian Mixture Model (22.05% better).

Computing inclusion probabilities of units in the Pivotal Sampling algorithm is an important step. To compute these probabilities for plant data, the notion of deviation was introduced in [1]. If $(\kappa_j)_i$ represents the value of the j^{th} characteristic for the i^{th} specie, the total deviation for the i^{th} specie is given by

$$dev_i = \sum_{j=1}^{m} (base((\kappa_j)_{i=1,\ldots,n}) - (\kappa_j)_i), \tag{1.1}$$

where m denotes the total number of the characteristics and the *base* function captures the representative value of the j^{th} characteristic over all the species (n). Since a specie with a lower deviation is more likely to represent the entire population, the inclusion probability of a specie was taken to be inversely proportional to this deviation. In the earlier work [1], authors had taken *base* to be the maximum function. In this work, we propose the use of a median function instead, which is more intuitive.

We again perform experiments on 1865 rice species. Initially, we support our above choice via a statistical analysis. Next, we show that use of the median function is better not just in Pivotal Sampling but also in another competitive probabilistic algorithm (Poisson Sampling). Further, we demonstrate that overall Pivotal works better than Poisson. Finally, we show that the standard SC with maximum based

Pivotal sampling is about 50.10% better than HC while our base "a" locally scaled SC with median based Pivotal Sampling is 61.24% better than HC.

Next, we discuss relevant past work done in clustering of rice species using phenotypic data. We do not plan to be exhaustive.

1.1 Literature Survey

[7] employed the Unweighted Pair Group Method with Arithmetic Mean (UPGMA), which is a type of Hierarchical Clustering. The paper investigated clustering patterns of 10 species of Thai indigenous upland rice using phenotypic traits. Cluster analysis resulted in three distinct groups. This highlighted the potential for utilizing traits such as flag leaf length, harvest index, total dry weight, total grain weight, and filled grains in obtaining higher grain yield.

[8] used UPGMA-based dendrogram analysis, also a type of Hierarchical Clustering. The paper investigated the clustering of 114 rice species of North East India using phenotypic traits. The species were clustered into seven groups. Grouping of rice species was done on 15 traits which include root length, shoot length, fresh root length, dry root length, root volume, root angle, peripheri root, etc. This was used for the assessment of genetic variation in rice breeding programmes.

[9] employed Neighbor-joining tree analysis, which is a type of Hierarchical Clustering. In this paper, the authors clustered 15 red rice varieties in Bangladesh using phenotypic traits. The species were clustered into five groups (I, II, III, IV, V). A total of 15 traits, e.g., culm diameter, plant height, panicle length, five panicle weight, filled grains per panicle, etc. were used for the detection of economically desired traits for future molecular breeding programmes.

[10] used UPGMA for Clustering. The paper investigated the clustering of 41 rice species cultivated by Indian farmers using phenotypic traits. The species were clustered into four groups. The phenotypic traits, such as the number of grains per panicle, thousand-grain weight, and the measurements of length and breadth for both whole and milled grains, were used in breeding programs.

All the above works have two major drawbacks. First, they used Hierarchical Clustering, which is not very accurate and computationally expensive. Second, they experimented on small datasets (a few tens to a few hundreds of species).

We overcome both these drawbacks here. As mentioned in the introduction, we use Spectral Clustering (old from [1] and new), which is known to be highly accurate. To offset the high computational cost we use sampling (old from [1] and new). Also, we experiment with about $2,000$ species of rice.

The rest of the manuscript has three more chapter. Chapter 1 presents both our old and new Spectral Clustering. In Chapter 2, we discuss Pivotal Sampling and its new application. Finally, Chapter 4, provides conclusion and future work.

Chapter 2

Clustering on Phenotypic Data

In this chapter, we explore clustering methods for phenotypic data of rice species, with focus on Spectral Clustering. In Section 2.1, we discuss the standard Spectral Clustering algorithm along with the distance measures commonly used in it. Section 2.2 introduces a novel modification to the standard Spectral Clustering, which involves using a base "a" exponential function, instead of the natural exponential function, to build the similarity matrix. We theoretically justify this choice as well. In Section 2.3, we combine our above novelty with another improvement of local scaling in the Spectral Clustering algorithm. Section 2.4 describes the data, discusses its normalization, and provides eigenvalue analysis to support our novel algorithm. In Section 2.5, we first compute the ideal number of clusters. Second, we give results for variants of Spectral Clustering and the current standard Hierarchical Clustering. Here we show that our proposed base "a" locally scaled Spectral Clustering works better. Finally, and third, we give results for another popular clustering algorithm (Gaussian Mixture Model) and show that our proposed algorithm is the best. Section 2.6 provides the conclusion of the chapter.

2.1 Standard Spectral Clustering

Spectral Clustering is one of the most popular modern clustering algorithms. It is simple to implement and can be solved efficiently by standard linear algebra software.

Given a set of points $S = \{p_1, p_2, ..., p_n\}$ in R^m that we want to cluster into k subsets, the algorithm consists of below steps [11].

- Form a similarity matrix A such that

$$A_{ij} = e^{\left(-\frac{d_{p_i p_j}}{2\sigma^2}\right)}, \tag{2.1}$$

  with i, $j \in \{1, ..., n\}$ and $A_{ii} = 0$. Here, $d_{p_i p_j}$ denotes the distance between two points p_i and p_j and σ defines the decay of the distance.

- Construct the Laplacian matrix

$$L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \tag{2.2}$$

  where D is a diagonal matrix whose (i, i) element is the sum of the elements of A's i^{th} row.

- Let e_1, e_2 .., e_k be the first k eigenvectors of L. Then, form the matrix $X = [e_1, e_2...., e_k]$ by stacking the eigenvectors as columns of this matrix.

- Form Y by normalizing X's rows to unit length.

- Cluster Y using the $k-$Means clustering.

The earlier work that we extend here [1] used this standard Spectral Clustering.

There are many ways to the distance between points p_i and p_j in (2.1), i.e., $d_{p_i p_j}$. Some common ones are Euclidean, Squared-Euclidean, and Correlation, and are given below.

- **Euclidean**: It represents the straight-line distance between two points in Euclidean space, and is calculated as follows:

$$d_{ij} = \sqrt{\sum_{l=1}^{m} (p_i^l - p_j^l)^2}, \tag{2.3}$$

  where p_i^l is $(\chi_l)_i$ and p_j^l is $(\chi_l)_j$.

- **Squared-Euclidean**: It is the square of the Euclidean distance, and is given as follows:

$$d_{ij} = \sum_{l=1}^{m} (p_i^l - p_j^l)^2, \tag{2.4}$$

with p_i^l and p_j^l are defined as above.

- **Correlation**: It captures the correlation between two non-zero vectors, and is expressed as follows:

$$d_{ij} = 1 - \frac{(p_i - \bar{p_i})^t (p_j - \bar{p_j})}{\sqrt{(p_i - \bar{p_i})^t (p_i - \bar{p_i})}\sqrt{(p_j - \bar{p_j})^t (p_j - \bar{p_j})}}, \tag{2.5}$$

where $\bar{p_i}$ and $\bar{p_j}$ represent the means of vectors p_i and p_j, respectively, multiplied by a vector of ones, and the t indicates the transpose operation.

2.2 Base "a" Spectral Clustering

Spectral Clustering is based upon spectral graph theory. To derive our new algorithm, we first revisit few concepts from this domain. We form a graph from the given data as follows [12]: (a) use data points as vertices and, (b) connect each point with the remaining points with an edge having weight equal to the corresponding element of similarity matrix A.

Definition 2.2.1 (Conductance [13]). *Given a graph $G = (V, E)$ with V partitioned into S and $\overline{S}$, the conductance of S is defined as*

$$\phi(S) = \frac{|E(S, \overline{S})|}{Vol(S)}, \tag{2.6}$$

where numerator is the fraction of edges in cut$(S, \overline{S})$ and denominator is the sum of vertices in S. The conductance of G is defined as

$$\phi(G) = \min_{vol(S) \leq \frac{vol(V)}{2}} (\phi(S)), \tag{2.7}$$

or the smallest conductance among all sets with at most half of the total volume.

Theorem 2.2.1 (Cheeger's Inequality [13]). *For any graph G,*

$$\frac{\lambda_2}{2} \leq \phi(G) \leq \sqrt{2\lambda_2}, \tag{2.8}$$

where λ_2 is the 2^{nd} smallest eigenvalue of L given by (2.2).

From the above theorem, we infer that $\phi(G)$ is close to zero (or G can be grouped into 2 clusters) if and only if λ_2 is close to zero. Note that λ_1 is always zero. This characterization carries over to higher multiplicities as well. G can be grouped into k clusters if and only if there are k eigenvalues close to zero [14].

We propose using a base "a" exponential function instead of the natural exponential function in (2.1) of the standard spectral clustering algorithm. That is,

$$A_{ij} = a^{\left(-\frac{d_{p_i p_j}}{2\sigma^2}\right)}, \tag{2.9}$$

where "a" $>$ "e". This results in A_{ij} of (2.9) being smaller than A_{ij} of (2.1).

Theorem 2.2.2. *The elements of non-normalized Laplacian matrix $L = D - A$ get smaller in absolute sense when we use (2.9) instead of (2.1), with "a" > "e", to build A. Here, D is the diagonal matrix whose (i, i) element is the sum of i^{th} row of A. Further, this leads to reduction in upper bound of eigenvalues of L.*

Proof. The first part of the Theorem is obvious. Since elements of A get smaller with the proposed change of base, the elements of D also get smaller (D is formed via elements of A). Thus, elements of $D - A$ or L get smaller in the absolute sense. For the second part of the proof, we use the fact that the spectral radius of the matrix is bounded above by its norm or $\rho(L) \leq ||L||$. □

Conjecture 2.2.3. *The above theorem holds true when we change the non-normalized Laplacian matrix $L = D - A$ with the normalized Laplacian matrix $L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$.*

We are unable to prove this theoretically. However, this holds true experimentally. We demonstrate in the Analysis section later in this paper that the change of the base as discussed in the above conjecture leads to a reduction in the eigenvalues of L. Thus,

from the Cheegers's Inequality (2.8), we infer that we should get a better clustering when we use base "a" exponential function instead of the natural exponential function in building the similarity matrix. This is supported by experiments in the Results section.

2.3 Base "a" Locally Scaled Spectral Clustering

Next, to further improve our clustering, we depart from the conventional practice of utilizing a global scaling factor (σ) in (2.9). Instead, we adopt the concept of a local scaling factor specific to each data point, as proposed by [2]. Now, the similarity between the two points is defined as

$$A_{ij} = a^{\left(-\frac{d_{p_i p_j}}{\sigma_i \sigma_j}\right)}. \tag{2.10}$$

The determination of the local scale σ_i involves analyzing the local statistics within the neighborhood of a given point. We employ a simple yet effective approach for scale selection. That is,

$$\sigma_i = d_{p_i p_K}, \tag{2.11}$$

where p_K is the K^{th} neighbor of p_i. The selection of K is independent of the scale and based upon the data dimensionality.

In the Analysis section, we show that this choice of similarity function leads to further reduction in eigenvalues of L (more than just use of base "a" exponential function). Thus, again by Cheegers's Inequality (2.8), this choice of the similarity function should lead to better clustering than both the natural exponential function and base "a" exponential function clustering. This is again supported by experiments in the Results section.

2.4 Analysis

Few settings of our algorithms from previous sections are as follows: (a) The best value of "a" (the base of the exponential function used to build the similarity matrix) for us turns to be "30". (b) The most fitting value of K (neighbor of a point in local scaling) comes to 180.

Below, in Section 2.4.1 we describe the rice data. Section 2.4.2 discusses the normalization of data. Finally, in Section 2.4.3 we do eigenvalue analysis to justify the use of base "30" as well as local scaling in Spectral Clustering.

2.4.1 Data Description

Our algorithm can be applied to any plant data, however, in this work, we do experiments on the rice data. This data is taken from The International Rice Information System (IRIS) (`www.iris.irri.org`)- a platform for meta-analysis of rice crop plant data [15]. It consists of 12 phenotypic (or physical) characteristics of 1865 rice species. A snapshot of this data is given in the Table 2.1.

Sr. No.	**Cudicle Repro -duction**	**Cultural Repro -duction**	**Cuneiform Repro -duction**	**Grain Length**	**Grain Width**	**Grain Weight per 100 Seed**
1	5	147	16	8.7	3.1	2.9
2	6	150	27	7.1	3.3	2.1
–	–	–	–	–	–	–
1875	3	56	16	7.7	3.4	2.8

Sr. No.	**HDG 80HEAD**	**Lightness of Color**	**Leaf Length**	**Leaf Width**	**Plant Post Harvest Traits**	**Stem Height**
1	102	25	72	1.1	29	54
2	123	20	73	1.5	27	45
–	–	–	–	–	–	–
1875	69	10	31	1	16	23

Table 2.1: Phenotypic data of rice plant.

2.4.2 Normalization

Let us consider a dataset consisting of n species with m distinct characteristics. We begin by normalizing the characteristics as follows [1]:

$$(\chi_j)_i = \frac{(x_j)_i - \min(x_j)}{\max(x_j) - \min(x_j)}. \tag{2.12}$$

Here,$(\chi_j)_i$ and $(x_j)_i$ are the normalized and the actual value of the j^{th} characteristic for the i^{th} specie, respectively. Next, we represent each specie as

$$p_i = \begin{bmatrix} (\chi_1)_i \\ (\chi_2)_i \\ \vdots \\ (\chi_m)_i \end{bmatrix},$$

for $i = 1, 2, .., n$.

2.4.3 Eigenvalue Analysis

Table 2.2 illustrates the first 30 smallest eigenvalues of the Laplacian matrix obtained from similarity matrix built using natural exponential function, base "30" exponential function, and base "30" & locally scaled exponential function. Additionally, Figure 2.1 plots these eigenvalues.

This table and figure validates our Conjecture 2.2.3. That is, the eigenvalues associated with base "30" exponential function are closer to zero as compared to the eigenvalues associated with the natural exponential function. Thus, as mentioned earlier, using Cheegers's Inequality (2.8) base "30" exponential function should result in better clustering than the natural exponential function based clustering. This turns to be true experimentally, which we demonstrate in the results section.

Second, we further observe that, as claimed in Section 2.3, eigenvalues corresponding to base "30" & locally scaled exponential function are more closer zero than the

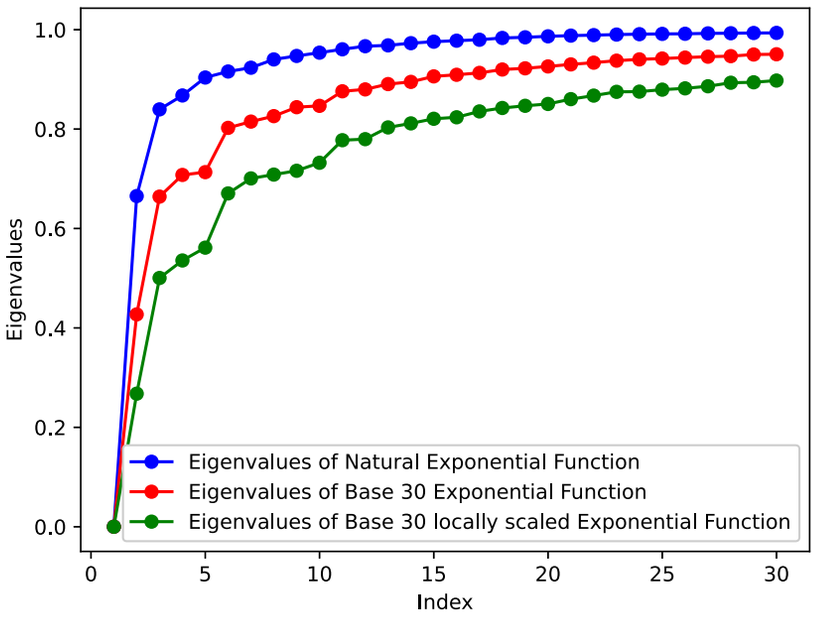

Figure 2.1: First 30 eigenvalues obtained using natural exponential function, base "30" exponential function, and base "30" & locally scaled exponential function for building the similarity matrix.

prior two function choices. Thus, again by using Cheegers's Inequality (2.8), this function should give the best clustering. This turns to be true experimentally as well, which we again demonstrate in the results section.

2.5 Results

In this results section, we perform three kinds of experiments. In Section 2.5.1, we determine the ideal number of clusters. Section 2.5.2 gives the results of variants of Spectral Clustering and current standard Hierarchical clustering, and also demonstrates that our proposed algorithm (base "30" locally scaled Spectral Clustering) works better. Finally, in Section 2.5.3, we experiment with another popular clustering algorithm (Gaussian Mixture Model), and show that our proposed algorithm is the best.

Sr. No.	1	2	3	4	5	6	7	8	9
Base (e) Fun	0.0000	0.6653	0.8397	0.8677	0.9038	0.9159	0.9235	0.9402	0.9472
Base (30) Fun	0.0000	0.4272	0.6642	0.7078	0.7134	0.8025	0.8151	0.8261	0.8441
Base (30) locally scaled Fun	0.0000	0.2677	0.5005	0.5357	0.5615	0.6706	0.7008	0.7085	0.7162

Sr. No.	10	11	...	25	26	27	28	29	30
Base (e) Fun	0.9538	0.9606	...	0.9916	0.9918	0.9926	0.9929	0.9934	0.9936
Base (30) Fun	0.8468	0.8761	...	0.9420	0.9439	0.9457	0.9468	0.9501	0.9506
Base (30) locally scaled Fun	0.7323	0.7777	...	0.8794	0.8820	0.8862	0.8933	0.8942	0.8977

Table 2.2: Small eigenvalues of Laplacian matrices obtained using different similarity matrices.

2.5.1 Ideal Number of Clusters

Determining the ideal number of clusters is always a dataset specific task, and we have many methods to approximate this number. A commonly used method is the Eigen-Gap heuristic [11] tailored for Spectral Clustering. This heuristic says that given λ_i's as the eigenvalues of the Laplacian matrix, k is the ideal number of clusters when the eigenvalues λ_1, λ_2, .., λ_k have small magnitude while λ_{k+1} has comparatively much higher magnitude.

As mentioned earlier, Figure 2.1 displays the 30 smallest eigenvalues derived from the Laplacian matrix obtained using similarity matrix built using three functions from the previous section. While the eigenvalues of natural exponential function show a jump at eigenvalues numbered 1 and 2, the eigenvalues of both base "30" exponential function and base "30" locally scaled exponential function show a jump at eigenvalues numbered $1, 2, 5$ and 10. Plant biologists are typically interested in more number of clusters, and hence, we choose 5, 10, 15, and 20 as the ideal number of clusters for our experimentation.

Clusters	Distance	Old SC	Base "30" SC	New SC	HC	Old SC Vs HC %	Base "30" SC Vs HC %	New SC Vs HC %
5	Euclidean	0.1249	0.1235	0.1258	0.07	28.88	33.8	34.15
	SqEuclidean	0.215	0.2214	0.2242	0.1349			
	Correlation	**0.2200**	**0.2284**	**0.2290**	**0.1707**			
10	Euclidean	0.0952	0.1103	0.1113	0.0662	35.64	51.55	55.43
	SqEuclidean	0.1592	**0.1952**	**0.2002**	0.1079			
	Correlation	**0.1747**	0.1913	0.1987	**0.1288**			
15	Euclidean	0.0841	0.0981	0.0946	0.026	51.55	71.23	72.33
	SqEuclidean	0.1342	**0.1714**	**0.1725**	0.0677			
	Correlation	**0.1517**	0.1693	0.1719	**0.1001**			
20	Euclidean	0.0784	0.0881	0.0874	0.0128	83.35	103.15	103.40
	SqEuclidean	0.1109	0.1569	0.1587	0.0432			
	Correlation	**0.1454**	**0.1611**	**0.1613**	**0.0793**			
Average percentage gain						49.86	64.93	66.33

Table 2.3: Silhouette values of different clustering algorithms.

2.5.2 Comparison of variants of Spectral Clustering with Hierarchical Clustering

Here, we compare four clusterings. First is the standard Spectral Clustering as described in Section 2.1 (also natural exponential function based Spectral Clustering), and used in [1]. We refer to this as Old SC. Second is our proposed base "30" exponential function based Spectral Clustering as elaborated in Section 2.2. We call this the Base "30" SC. Third is, again our proposed, base "30" locally scaled exponential function based Spectral Clustering as descibed in Section 2.3. We call this New SC. Finally, the fourth is Hierarchical Clustering, which is mentioned in the literature. We refer to this as HC.

The results of this comparison are given in Table 2.3. Here, the first column denotes the number of the clusters that are chosen based upon the previous analysis. The second column contains the distance metrics used to build the similarity matrix in the clustering algorithms. Columns three through six list the silhouette values of the respective algorithms. Best values in a cell are highlighted in bold. Finally, columns seven through nine give the percentage gain of Old SC, Base "30" SC, and New SC

over HC, respectively. The best values are used to compute this gain.

We conclude that most improvement in clustering quality is obtained when we move from Old SC to Base "30" SC (49.86% to 64.93% gain over HC). Further, New SC leads to a little bit more improvement (64.93% to 66.33% gain over HC). To sum up, New SC works the best.

2.5.3 Results using another Popular Clustering

In this section, we present results of another popular clustering. Table 2.4 presents the silhouette values for Gaussian Mixture Model (GMM) Clustering. Here, we cluster the data by considering four cases, namely when the covariance matrix is full and has shared covariance (Type 1), the covariance matrix is diagonal and has shared covariance (Type 2), the covariance matrix is full and has unshared covariance (Type 3), and the covariance matrix is diagonal and has unshared covariance (Type 4). Comparing the results of this clustering with our New SC from Table 2.3, we find that our proposed algorithm works the best (22.05% better).

Clusters No. / Clustering		5	10	15	20
GMM Clustering	Type 1	0.1855	0.1445	0.1098	0.1014
	Type 2	**0.2096**	**0.1573**	0.1220	**0.1351**
	Type 3	0.1906	0.1023	0.0867	0.0669
	Type 4	0.1673	0.1386	**0.1304**	0.0607

Table 2.4: Silhouette values for GMM clustering.

2.6 Conclusion

In this chapter, we focused on clustering rice species using their phenotypic characteristics. Initially, we discussed the standard Spectral Clustering algorithm. Then, we proposed our novel base "30" Spectral Clustering. Next, we combined our novelty of base "30" with an existing technique of local scaling in the Spectral Clustering algorithm. Further, we discussed the rice data, normalized it, and performed an eigenvalue analysis to justify our novel algorithm.

Finally, we performed experiments. Here, we first computed the ideal number of clusters. Second, we compared the proposed variants of the Spectral Clustering algorithm with the current standard Hierarchical Clustering. Here, we showed that our base "30" locally scaled Spectral Clustering works better. Finally, and third, we gave results for another popular clustering (GMM), and demonstrated that our proposed algorithm is the best.

Chapter 3

Sampled Clustering on Phenotypic Data

In this chapter, we explore sampling to reduce the complexity of clustering. In Section 3.1, we discuss Pivotal Sampling, which is an unequal probabilistic sampling algorithm. Section 3.2 discusses the application of this sampling to phenotypic data and introduces novel modifications in computing the inclusion probability employed in sampling (deviation from the median provides better results in comparison to the deviation from the maximum). Next, in Section 3.3, we perform analysis to support our novelty. Here, we also experiment with another popular sampling algorithm; Poisson Sampling. We first show median based approach works better in case of Poisson Sampling as well, and second, we show that Pivotal is slightly better than Poisson. Further, Section 3.4 compares variants of Sampled Spectral Clustering with Hierarchical Clustering demonstrating that our novel algorithm (base "30" locally scaled Spectral Clustering with median based Pivotal Sampling) works the best. Further, we end this chapter with a summary in Section 3.5.

3.1 Pivotal Sampling

Pivotal Sampling [16] is an effective non-uniform probabilistic sampling technique without replacement, which employs a fixed sample size. Consider a finite population U with a size of n, where each unit is uniquely labeled $i = 1, 2, ..., n$. A sample S, a subset of U, can have a size determined either randomly ($N(S)$) or fixed (N). An

integral aspect of this unequal probability sampling technique involves acquiring the inclusion probabilities for all units in the population, denoted as π_i with i = 1, 2, ..., n. This technique involves sequentially scanning of the population, wherein at each iteration there is a duel between two units, i and j, with inclusion probabilities of π_i and π_j, respectively, to get selected or rejected from the sample.

- **Selection step**: If $(\pi_i + \pi_j \geq 1)$, then one of the units will be selected (i.e., sampled).
- **Rejection step**: If $(\pi_i + \pi_j < 1)$, then one of the units will be rejected (i.e., not sampled).

In the Selection step, we set one of the probabilities to 1 while in the Rejection, we set one of the probabilities to 0. By repeating this process, we get our sample of the desired size. Since there is for a single pass or scan of the data, the method's computational complexity is O(n).

3.2 Application

In our earlier work, the probabilities in the pivotal sampling algorithm were taken as a function of variations in the characteristics of the plant species. Plant height, number of pods per plant, days to pod initiation, seed yield per plant, etc., are some characteristics examples. Specifically, this probability for the i^{th} specie was taken to be inversely proportional to

$$dev_i = \sum_{j=1}^{m}((max(\chi_j)_{i=1,..,n}) - (\chi_j)_i), \tag{3.1}$$

where dev_i denotes the deviation for the i^{th} specie, $max(\chi_j)_{i=1,..,n}$ denotes the maximum value of the j^{th} characteristic over all the species, and $(\chi_j)_i$ denotes the value of j^{th} characteristic for the i^{th} specie.

This method of computing deviation implies that the higher the characteristic values of a specie, the higher its probability. Thus, the higher chance that it would

be selected in the sample (and would represent all the species). However, species with higher characteristic values representing all species is not ideal. Take 'days to pod initiation' characteristic for example. Here, it could be that most species take less number of days to pod initiation. Hence, having a specie that takes more number of days for pod initiation, to represent all would be inaccurate.

Hence, we propose the use of the median function instead of the maximum function in (3.1). That is,

$$dev_i = \sum_{j=1}^{m}((median(\chi_j)_{i=1..,n}) - (\chi_j)_i). \tag{3.2}$$

This would ensure that the samples consist of species with characteristic values closer to the median values. Thus, such sampled species would represent all the species better. We support our this choice by performing statistical analysis as well as experiments.

The final inclusion probability of i^{th} specie is obtained by normalizing as follows [1]:

$$\pi_i = N\frac{\frac{1}{dev_i}}{\sum_{i=1}^{n}\frac{1}{dev_i}}. \tag{3.3}$$

We then iteratively apply the selection and rejection steps of the pivotal sampling technique in order to obtain a sample of size N.

Next, we perform earlier discussed clustering techniques on the sampled data of size N. Since the total data is of size n (with $n > N$), there is a need to reverse map the remaining $n - N$ data points to the required cluster. For this, we define the notion of average similarity, which between the non-clustered specie $\tilde{p}$ and the cluster C_k is given as

$$AS(C_k, \tilde{p}) = \frac{1}{\#(C_k)}\sum_{y \in C_k} a^{\left(-\frac{d_{\tilde{p}y}}{\sigma_{\tilde{p}}\sigma_y}\right)}. \tag{3.4}$$

Here, $\#(C_k)$ represents the number of species in cluster C_k. We compute the average similarity of $\tilde{p}$ with all the k clusters and assign it to the cluster with which $\tilde{p}$ has the maximum similarity.

3.3 Analysis

In order to assess the quality of the obtained samples, we employ the estimation of the following metrics: the Population Total (PT) estimator and the Horvitz-Thompson (HT) estimator. We define the PT estimator for a specific characteristic as the summation of all the values associated with that characteristic. That is, for a characteristic j, it is computed as

$$(Y_{PT})_j = \sum_{i=1}^{U} (\chi_j)_i, \tag{3.5}$$

where U is the set of all species (or #(U)=n) and $(\chi_j)_i$ is the value of the j^{th} characteristic for the i^{th} specie.

The HT estimator is defined as follows:

$$(Y_{HT})_j = \sum_{i=1}^{S} \frac{(\chi_j)_i}{\pi_i}, \tag{3.6}$$

where S is the set of samples (or #(S)=N), $(\chi_j)_i$ is defined as above, and π_i denotes inclusion probability of i^{th} specie. The sampling is better when the values of the HT estimators are closer to the PT estimators. The size of the sample N taken is 500.

Table 3.1 lists the PT estimators, HT estimators when using the maximum function in (3.1), and HT estimators when using the median function in (3.1) for all the 12 characteristics. The first thing we observe from the table is that the PT estimators are close to both the sets of HT estimators, and hence, both the samplings are good. Second, we also observe that the PT estimators are closer to the HT estimators when using the median function than the HT estimators when using the maximum function.

To better demonstrate the second observation, we plot the difference between the PT and the HT estimators (when using both the maximum function and the median function) for all characteristics in Figure 3.1. We can notice from this figure that the median function leads to a smaller difference in 9 of the 12 characteristics as compared to the maximum function. Thus, median based sampling is better.

Here, we also compare with another popular sampling, i.e. Poisson Sampling.

Characteristics	PT estimators	HT (Max) estimators	HT (Median) estimators
1	9.037800e+03	9.099514e+03	9.032072e+03
2	2.121750e+05	2.113126e+05	2.103986e+05
3	3.057500e+04	3.120878e+04	3.089743e+04
4	1.614470e+04	1.619068e+04	1.612586e+04
5	5.661400e+03	5.625295e+03	5.652058e+03
6	4.630100e+03	4.599130e+03	4.538882e+03
7	1.865000e+05	1.885044e+05	1.852332e+05
8	3.343930e+04	3.508258e+04	3.456751e+04
9	1.024092e+05	1.034117e+05	1.026595e+05
10	2.569800e+03	2.552394e+03	2.562038e+03
11	4.660790e+04	4.639347e+04	4.629499e+04
12	7.214200e+04	7.184603e+04	7.227474e+04

Table 3.1: PT and HT estimators with Pivotal Sampling.

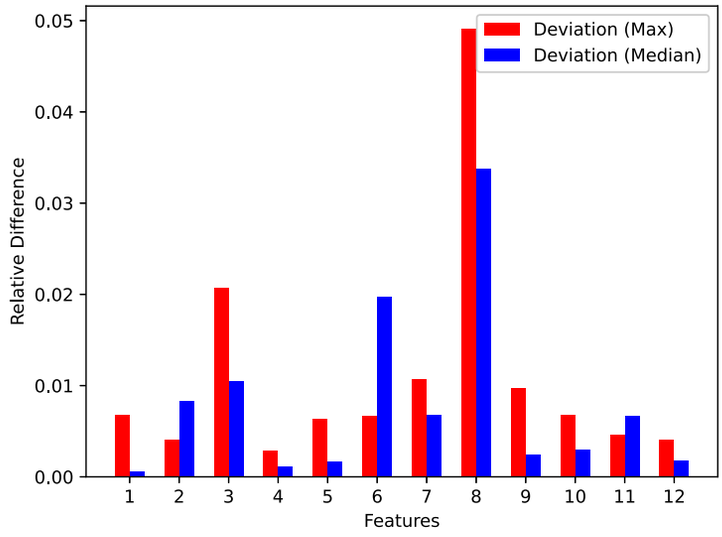

Figure 3.1: Difference between PT and HT estimators when pivotal sampled with maximum function and median function.

Again, Table 3.2 lists the PT estimators, HT estimators when using the maximum function in (3.1), and HT estimators when using the median function in (3.1) for all the 12 characteristics. We plot the difference between the PT and the HT estimators (when using both the maximum function and the median function) for all characteristics in Figure 3.2. From both the table and figure we can see that median function leads to a smaller difference in all the 12 characteristics as compared to the maximum function. Thus, median based sampling is again better.

If we compare the best of Pivotal Sampling (median based) with the best of Poisson Sampling (median based), we find that the Pivotal is slightly better. Hence, we use this sampling for our experiments in the next section.

Characteristics	PT estimators	HT (Max) estimators	HT (Median) estimators
1	9.037800e+03	8.475070e+03	8.775710e+03
2	2.121750e+05	1.963557e+05	2.099071e+05
3	3.057500e+04	2.918635e+04	2.991180e+04
4	1.614470e+04	1.521900e+04	1.613994e+04
5	5.661400e+03	5.307340e+03	5.659630e+03
6	4.630100e+03	4.399400e+03	4.636690e+03
7	1.865000e+05	1.768083e+05	1.852229e+05
8	3.343930e+04	3.242312e+04	3.382171e+04
9	1.024092e+05	9.535065e+04	1.007785e+05
10	2.569800e+03	2.427490e+03	2.552710e+03
11	4.660790e+04	4.329817e+04	4.599378e+04
12	7.214200e+04	6.767168e+04	7.234891e+04

Table 3.2: PT and HT estimators with Poisson Sampling.

3.4 Results

Here, we compare three algorithms. First, is the standard Spectral Clustering (natural exponential Spectral Clustering) along with the maximum function based Pivotal Sampling as used in [1]. We refer to this as Old Sampled SC. Second is the base "30" locally scaled Spectral Clustering along with the median function based Pivotal Sampling (our novel algorithm). We call this New Sampled SC. Finally, the

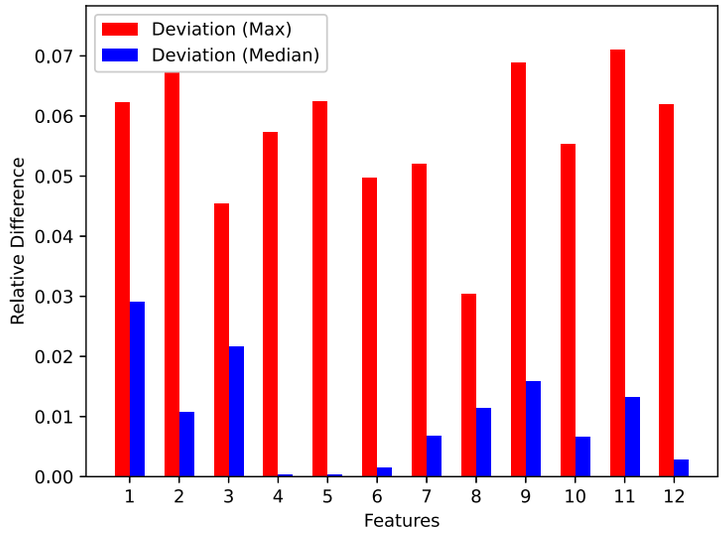

Figure 3.2: Difference between PT and HT estimators when poisson sampled with maximum function and median function.

third is the Hierarchical Clustering. We refer to this as HC. The most fitting value of K (neighbor of a point in local scaling) comes to be 400.

The results for this are given in Table 3.3. Here, the first column denotes the number of the clusters. The second column contains the distance metrics used to build the similarity matrix in the clustering algorithms. Columns three, four, and five list the silhouette values of the three algorithms mentioned in the above paragraph. Columns six and seven give the respective gain. As evident, the Old Sampled SC gives 50.10% gain over HC while our New Sampled SC improves this gain to 61.24%. Thus, our new algorithm is substantially better than the current best and also computational cheaper.

3.5 Summary

This chapter also dealt with the clustering of rice phenotypic data but after sampling of data to reduce the complexity of clustering. Initially, we discussed Pivotal Sampling. Then, we discussed the novel application of Pivotal Sampling to rice data. The novelty involved using a more intuitive function (median instead of a maximum)

Clusters	Distance	Old Sampled SC	New Sampled SC	HC	Old Sampled SC Vs HC %	New Sampled SC Vs HC %
5	Euclidean	0.1218	0.127	0.07	33.33	38.08
	SqEuclidean	0.2111	0.2233	0.1349		
	Correlation	**0.2276**	**0.2357**	**0.1707**		
10	Euclidean	0.0851	0.1043	0.0662	32.76	53.26
	SqEuclidean	0.1602	0.1943	0.1079		
	Correlation	**0.171**	**0.1974**	**0.1288**		
15	Euclidean	0.0815	0.0858	0.026	57.14	62.94
	SqEuclidean	0.1348	0.1586	0.0677		
	Correlation	**0.1573**	**0.1631**	**0.1001**		
20	Euclidean	0.0751	0.0774	0.0128	77.18	90.67
	SqEuclidean	0.1115	0.1367	0.0432		
	Correlation	**0.1405**	**0.1512**	**0.0793**		
Average percentage gain					50.10	61.24

Table 3.3: Silhouette values of variants of Sampled Spectral Clustering and Hierarchical Clustering

in computing the inclusion probabilities. We supported this design choice via statistical measures. Next, we demonstrated that this choice of median is better in not just Pivotal Sampling but in Poisson Sampling as well. We also showed that Pivotal Sampling works slightly better than Poisson Sampling.

Further, experimentally we demonstrated that our novel algorithm (base "30" locally scaled Spectral Clustering with median based sampling) is better than old natural exponential Spectral Clustering with maximum based sampling. We also showed that our new algorithm is substantially better than the currently prevalent Hierarchical Clustering and also cheaper than it.

Chapter 4

Conclusion and Future Work

Phenotypic data of plants is commonly used to group species into different categories, which is further used in breeding programs. Hierarchical Clustering (HC) is a common algorithm that is used for implementing such groupings. Since this algorithm is not very accurate and expensive as well, recently authors in [1] proposed the use of the standard Spectral Clustering (SC) to improve accuracy and Pivotal Sampling to reduce the complexity. They demonstrated the usefulness of their algorithm via experiments on the soybean plant.

In this work, we proposed a novel base "30" locally scaled SC that improved the standard SC. We performed experiments on 1865 rice species. Initially, an eigenvalue analysis justified the use of this new algorithm. Next, we demonstrated that our base "30" locally scaled SC led to larger improvement over HC (66.33%) as compared the improvement of standard SC over HC (49.86%). We also showed that our new algorithm worked better than another commonly used clustering of GMM (22.05% better).

Further, we proposed a better application of Pivotal Sampling to plant phenotypic data. Earlier, a maximum function was used to compute the inclusion probabilities in Pivotal Sampling, which was not intuitive. Instead, we proposed the use of a median function that worked better. We performed experiments on 1865 rice species. Initially, we performed a statistical analysis that justified our choice on not just Pivotal Sampling but also on another popular probabilistic sampling (Poisson). Next, we

showed that Pivotal Sampling performed better than Poisson Sampling. Finally, we demonstrated that our base "30" locally scaled SC with median Pivotal sampling gave 61.24% improvement over HC, which is substantially better than the standard SC with maximum Pivotal Sampling's 50.10% improvement over HC.

There are multiple future work directions here. First, in one of the seminal works [17], the authors have listed sufficiency conditions for SC to work well. It would be very useful to translate those conditions to plant data. Second, it would be useful to experiment with other accurate clusterings and samplings (e.g., see [18]). Third, although phenotypic characteristics are useful for clustering, genetic data of plant species carries more information. In our earlier work [19], we had explored the possibility of using genetic data for clustering, however, reduced data was used there. It would be interesting to experiment with the full data exhaustively.

Bibliography

[1] A. A. Shastri, K. Ahuja, M. B. Ratnaparkhe, and Y. Busnel, "Probabilistically sampled and spectrally clustered plant species using phenotypic characteristics," *PeerJ*, vol. 9, p. e11927, 2021.

[2] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering." in *Advances in neural information processing systems*, 2004, pp. 1601–1608.

[3] P. Painkra, R. Shrivatava, S. K. Nag, and N. Markam, "Clustering analysis of soybean germplasm (glycine max l. merrill)," *The Pharma Innovation Journal*, vol. 7, no. 4, pp. 781–786, 2018.

[4] P. Sharma, S. Sareen, M. Saini, A. Verma, B. S. Tyagi, and I. Sharma, "Assessing genetic variation for heat tolerance in synthetic wheat lines using phenotypic data and molecular markers," *Australian Journal of Crop Science*, vol. 8, no. 4, pp. 515–522, 2014.

[5] S. Swarup, E. J. Cargill, K. Crosby, L. Flagel, J. Kniskern, and K. C. Glenn, "Genetic diversity is indispensable for plant breeding to improve crops," *Crop Science*, vol. 61, no. 2, pp. 839–852, 2021.

[6] A. Kahraman, M. Onder, and E. Ceyhan, "Cluster analysis in common bean genotypes (Phaseolus Vulgaris L.)," *Turkish Journal of Agricultural and Natural Sciences*, vol. 1, pp. 1030–1035, 2014.

[7] S. S. Islam, J. Anothai, C. Nualsri, and W. Soonsuwon, "Genetic variability and cluster analysis for phenological traits of thai indigenous upland rice (oryza sativa l.)." *Indian Journal of Agricultural Research*, vol. 54, no. 2, 2020.

[8] H. Verma, J. Borah, and R. Sarma, "Variability assessment for root and drought tolerance traits and genetic diversity analysis of rice germplasm using SSR markers," *Scientific reports*, vol. 9, no. 1, p. 16513, 2019.

[9] M. Islam, M. Khalequzzaman, M. Prince, M. Siddique, E. Rashid, M. Ahmed, B. Pittendrigh, and M. Ali, "Diversity and population structure of red rice germplasm in Bangladesh," *PLoS One*, vol. 13, no. 5, p. e0196096, 2018.

[10] V. Pachauri, N. Taneja, P. Vikram, N. K. Singh, and S. Singh, "Molecular and morphological characterization of indian farmers rice varieties (oryza sativa l.)," *Australian Journal of Crop Science*, vol. 7, no. 7, p. 923, 2013.

[11] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395–416, 2007.

[12] S. Gharan, "Cheeger's inequality continued, spectral clustering. Lecture notes on design and analysis of algorithms I (University of Washington)," 2020.

[13] ——, "Cheeger's inequality and the sparse cut problem. Lecture notes on recent advances in approximation algorithms (University of Washington)," 2015.

[14] J. R. Lee, S. O. Gharan, and L. Trevisan, "Multiway spectral partitioning and higher-order cheeger inequalities," *Journal of the ACM (JACM)*, vol. 61, no. 6, pp. 1–30, 2014.

[15] C. G. McLaren, R. M. Bruskiewich, A. M. Portugal, and A. B. Cosico, "The international rice information system. A platform for meta-analysis of rice crop data," *Plant Physiology*, vol. 139, no. 2, pp. 637–642, 2005.

[16] J.-C. Deville and Y. Tille, "Unequal probability sampling without replacement through a splitting method," *Biometrika*, vol. 85, no. 1, pp. 89–101, 1998.

[17] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*. MIT Press, 2001, pp. 849–856.

[18] S. Jain, A. A. Shastri, K. Ahuja, Y. Busnel, and N. P. Singh, "Cube sampled K-prototype clustering for featured data," in *2021 IEEE 18th India Council International Conference (INDICON)*, 2021, pp. 1–6.

[19] A. A. Shastri, K. Ahuja, M. B. Ratnaparkhe, A. Shah, A. Gagrani, and A. Lal, "Vector quantized spectral clustering applied to whole genome sequences of plants," *Evolutionary Bioinformatics*, vol. 15, pp. 1–7, 2019.