OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

INF FAKULTÄT FÜR
INFORMATIK

Tinnitus Subtypes Identification with Clustering and Result Interpretation using Radial Bar Chart and Surrogate Models

Team Project KMD

Submission Date : 6th *September*2019

Supervisor: Uli Niemann

Authors:

Abhilash Mandal (221196)

Kritika Bhowmik (221276)

Priyanka Mohan (221217)

Shivani Jadhav (223856)

Abstract

Tinnitus is a highly common health condition which severely affects people. As a tinnitus patient, one needs to fill many lengthy questionnaires in order to identify the causes and track the status of the patient throughout the treatment. This can be very tiring and can cause a lack in accuracy of answers. Finding possible subtypes of tinnitus condition using phenotyping would help to categorize these questionnaires thus reducing the load on patients. This would also help medical practitioners to provide type-specific treatment. In this work we employ global and subspace clustering algorithms to find subgroups on tinnitus patient data consisting of 1030 records with 79 features. We used 3 global clustering algorithms, 2 subspace clustering algorithms and 1 global clustering algorithm with linear dimension reduction for this analysis. We built a radial chart visualization to view the clustering results. We used global surrogate models to interpret the black box clustering results to describe the clusters with most discriminating features. For this decision tree was used with the cluster labels to give us decision points to reach each subtype. Mean decrease accuracy for each feature obtained from random forest was used to validate if the decision tree was overfitting or underfitting. Based on the clustering results, we inferred that there exist 2 subgroups. We evaluated our results based on the qualitative and quantitative evaluation. Quantitatively hierarchical clustering gives the best cluster results based on dunn index (0.25) and connectivity (286.73). Qualitatively the results were evaluated based on total scores of questionnaires namely, Tinnitus Questionnaire (TQ) and Long Form General Depression Scale (ADSL). These total scores namely, "tq_tf" and "adsl_adsl_sum" appearing in decision points of decision tree helps in distinguishing between the subtypes as compensated/decompensated tinnitus and clinical/subclinical depression. Thus, clustering helps in finding subtypes of tinnitus and surrogate interpretable models helps in reducing the questionnaires based on these subtypes and aiding doctors in the tinnitus treatment.

Contents

Chapter 1

Introduction

1.1 Motivation

Tinnitus is a health condition in which patients experience the presence of a non-existing sound. The severity of this condition varies to a great extent along with the underlying conditions. Patients usually have to answer a lot of questions when they first visit the doctors for assessment of the severity of tinnitus. These lengthy questionnaires can be tiring and exhausting for patients to answer, which can thus cause a loss in accuracy for answers. This process can be simplified if we decrease the number of questionnaires asked to a patient. Target specific subgroups among tinnitus patients can be created and the relevance of a question to a particular patient can be decided based on his subgroup. A wide variety of factors have been suggested to affect the progression of tinnitus. By subgrouping the patients, the selection of subsets of patients more likely to benefit from different treatments can be optimized. This effort will not only help the drug development process but can also lead us to the preventive and curative treatment of tinnitus.

There has not been much work on finding subgroups of tinnitus patients using statistical approaches. Thus, we employ different clustering algorithms on tinnitus patient data without any prior assumptions about which features are important and infer these clusters to divide the patients to different tinnitus subtypes. We also create cluster feature vectors consisting of the features and their corresponding values to represent a cluster. However, a textual representation of data is not enough for understanding the results. Therefore, we created a visualization in the form of a radial bar chart to view the cluster feature vectors.

Further, it is a black box for a user to understand how input data in clustering algorithm is being processed to get a result. This particularly is important in the medical domain. Therefore, we make use of global surrogate interpretable model. Models such as decision trees are used as surrogate models to make the underlying model more interpretable by reduced approximation of the results. We support our decision tree model by using random forest to get the most important features.

1.2 Research Question

Are there any relevant subgroups of Tinnitus condition that can be identified from patient data?

1.3 Goals

Describing Clusters: Clustering algorithms just provide us records with cluster labels. However, our main goal is to be able to represent or describe the identified clusters so that we can distinguish them from each other. By analyzing the values of the features which best distinguish a cluster can thus be used for describing them.

Reducing Questions: Based on the cluster description, we can identify the subtypes in which the patient belongs to. But our goal is to reduce the number of questionnaires asked to patients to fill and thus decreasing the load on patients and medical practitioners. We can do this by making them answer only the questionnaires consisting of features which are important to distinguish between the subtypes.

1.4 Related work

The interaction of the genotype of an individual with the environment results in a set of characteristics which is nothing but the phenotype of that individual. Lopez et al. [1] in their paper claimed that based on the precise definition of phenotypes, the tinnitus subtyping strategies would help in selecting homogeneous groups of tinnitus patients with similarities that will serve as a strong basis for genetic studies. Tinnitus is considered as a symptom rather than a disease. Genetic studies would help in determining diagnostic markers and markers of resistance to treatment for different subgroups of tinnitus patients which would improve the selection of subjects and optimize the treatment outcome. Thus, targeted treatments for different disease subtypes can be developed. In their study, Lopez et al. investigated the genetic underpinnings of tinnitus by showing the evidence of heritability and by improving patient selection using phenotyping.

Tyler et al. [2] identified subgroups of tinnitus patients which are more likely to benefit from different treatments. A review of different strategies for subgrouping based on aetiology, subjective reports, the audiogram, psychoacoustics, imaging, and cluster analysis was provided. The paper also gives the preliminary outcomes of cluster analysis performed in an attempt to obtain subgroups.

In the paper by Niemann et al. [3], the depression severity was predicted using 11 classification methods. The data was obtained by extracting features from self-report questionnaires and socio-demographic data. subclinical and clinical depression were used as depression status and they were measured by the general depression scale questionnaire (ADSL). The paper concludes that depression in tinnitus patients can be better understood using predictive machine learning models. This understanding will improve the structure of questionnaires for tinnitus patients as well as contribute to the selection of suitable therapy for a particular patient.

Visualizing clusters in high dimensional data is a challenging task. It is even more difficult to analyze the results of subspace clustering because of different selected dimensions for different clusters. Tatu et al. [4] introduced ClustNails system to visualize the selected dimensions interactively. ClustNails consisted of two different views, namely SpikeNails and HeatNails. SpikeNails was used for cluster-centric analysis, where each dimension and their weights were used to represent the visualization. Whereas, HeatNails was used for record-centric analysis, where each row of HeatNail represented dimensions and columns represented data value of the record in that dimension. Since, understanding a visualization which involves all records can be difficult, SpikeNails is better suited for our subtypes. SpikeNails involves dimensions as labels and their values as spikes radially. Each radial SpikeNails chart represented each cluster with their selected dimensions which helped better understand subspace results. We extend this concept to include global clustering algorithms for visualization of the obtained subtypes of tinnitus patients.

1.5 Outline

This report is organized as follows. Chapter 2 gives an overview of the fundamentals of clustering, surrogate interpretable models and radial bar chart visualization. Chapter 3 elaborates about the concept of the workflow implementation and evaluation. Chapter 4 describes the data used in the work and inferred results and findings. Chapter 5 is about the discussion, summary of major findings, limitation and future work. Finally, Chapter 6 summarizes the whole work.

Chapter 2

Foundations

2.1 Fundamentals of Clustering

Clustering is an unsupervised learning approach which helps in identifying groups and patterns. We grouped the tinnitus data using different approaches of clustering. Since this is one of the first attempts to do clustering on Tinnitus patient data, we employ the most common clustering approaches based on *global clustering* and *subspace clustering* methods. The following subsections explain each approach.

2.1.1 Global Clustering

By global clustering, we analyze the data as a whole. In other words, we consider all the dimensions in the data set as is without any form of dimension reductions. The considered global clustering approaches are explained below.

k-means clustering: k-means is a square error based partitioning clustering approach. It is an iterative algorithm which tries to partition the data into k number of clusters. This is done by initially assigning k centers and then assigning the data to the nearest center. Then the centers are recalculated to reduce the squared error of each cluster. The data is again assigned to the newly calculated nearest center. The algorithm runs until there is no change in the cluster assignments. This algorithm accepts mainly one hyper-parameter, namely k number of clusters. k-means prefers spherical clusters and tends to cluster outliers. However, it is an effective and simple clustering algorithm used in many practical problems. Hence, we choose k-means as one of the global clustering approaches.

Hierarchical Clustering: Hierarchical clustering builds a hierarchical structure of data to find clusters. They can be visualized by the help of a dendrogram and a binary tree. Each leaf node depicts a data point and the root node depicts the whole dataset. Hierarchical has two main types: bottom-up (agglomerative) and top-down (divisive). In agglomerative clusters are merged as we move up and in divisive clusters are divided as we move down. The algorithm merges or breaks clusters depending on their similarity based on the distance metrics and linkage criteria (single, complete and average among others). The dendrogram is cut to get the final set of clusters. Since tinnitus is a medical condition and the entire data can be considered as a cluster in itself (patient medical data will be distinguishable from generally healthy people), we can recursively find subgroups in this data and check if there is a significant difference in these subgroups. Therefore, the hierarchical agglomerative approach seems to be more intuitive for this problem. Hierarchical clustering has a high time complexity and performs better for smaller clusters. One of the biggest advantages is that it is a simple algorithm to implement just like k-means.

Hierarchical k-means: k-means is sensitive to the selection of initial centroids and hierarchical clustering is good at identifying smaller clusters. This approach combines k-means and hierarchical clustering approach and solves the problem of the initial choice of centroids. Initially, hierarchical clustering is performed to build a tree. With the given K value, the tree is cut into k clusters. Then centroids are calculated for each cluster. These centroids are then used for initialization of k-means algorithm as cluster centers. This method combines the positives of both the methods and we use it to observe and compare the results with the previous two approaches.

2.1.2 Subspace Clustering

Subspace clustering tries to solve the problem of high dimensional data being nearly equidistant from each other by finding the clusters in different subspaces. Subspace algorithms can be classified as: top-down search (iterative search) and bottom-up search (grid-based methods). Since we are not using density-based global clustering algorithms, we restrict ourselves to top-down subspace algorithms. Top-down approaches create partitioned clusters i.e. each instance is assigned to only one cluster. We have selected Proclus and Orclus for our exploration using subspace algorithms as they have readily available packages in R.

Proclus: Proclus [5] uses three-phase approach: *Initialization, Iteration and Cluster Refinement*. In the initialization step, it selects a random set of medoids which are farthest from each other. It uses sampling to get data set and set of k medoids. In the next phase, these medoids are selected iteratively to improve the clustering. This is done by choosing the average distance between points and the nearest medoid as the *cluster quality* measure. In the refinement phase, based on the cluster formed new dimensions for each medoid is computed. Proclus can produce cluster with a different number of selected dimensions as it tries to maintain the average dimensions per cluster given by the user. This algorithm accepts mainly two user-specified hyper-parameters, namely K number of clusters and average dimensionality per cluster l. Proclus tends to prefer clusters of spherical shape and can miss certain clusters entirely if the initial set of chosen medoids is smaller. On the other hand, Proclus tends to be faster compared to bottom-up approach *clique* due to usage of sampling [6].

Orclus: Many datasets contain attribute correlations. Orclus [7] is used to tackle this by extending Proclus to include the search for non-axis parallel subspaces. This is done by building new dimensions which for a particular cluster using a covariance matrix. Orclus is based on k-means and has three phases: *Assign clusters, Find Vectors and Merge.* Initially the data points are assigned to the nearest cluster centers with minimal Euclidean distance in corresponding subspaces. In the next phase, for each cluster, subspace dimensionality l_c are determined which is closer to the user input l. Subspace dimensionality is calculated by calculating the covariance matrix and selecting the orthonormal eigenvectors having the smallest value i.e. least spread. From iteration to iteration l_c values decreases gradually to match l. During the *Merge* phase closest current pairs of clusters are merged to get final *k clusters* from initial *k0 clusters*. This algorithm accepts mainly three hyper-parameters, namely the final number of clusters k, the final number of dimensions per cluster l and initial number of clusters $k0$. Orclus is faster and scalable than Proclus because it does sampling with the help of $k0$ to decrease the number of merges. However, because of this, it tends to miss smaller clusters [8].

2.1.3 Clustering with reduced dimensions using PCA

Principal Component Analysis (PCA) reduces the dimension of the dataset while retaining as much variance in the data as possible. It creates new axes called Principal Components (PCs) which correspond to linearly uncorrelated features from the original set of correlated variables. PCA thus reduces the number of dimensions from the original set of dimensions. The PCs are arranged in the order of highest to lowest explained variances. This can be visualized using *scree plots* for PCA. First, n-PCs can be selected based on the desired variance to be explained by PCs. This is can be easily identified by *cumulative scree plot*, which sums up the individual variances of PCs and plots it against the PC count. After selecting the first n-PCs, other clustering algorithms can be employed.

2.2 Understanding with global surrogate interpretable models

Our major task is to reduce the number of questions in the questionnaires and also to group them based on clusters. But clustering methods are a black box for users, which is rather difficult to interpret. Interpretability is, how well a user can understand how the output is derived from a given input data. Therefore, in order to have more interpretability of this black box model we make use of global surrogate interpretable models [9]. These surrogate models are a reduced approximation of the output of the underlying model and with enhanced interpretability. To achieve this we employ the following surrogate models on the obtained clusters.

2.2.1 Decision Trees

Decision trees are used to make a decision path to reach a particular goal. They are represented in the form of a tree in which each node depicts a decision attribute and each leaf depicts one of the class labels. Trees are built while trying to maintain that each leaf is as pure as possible. Decision trees are split based on the hyper-parameters such as information gain or entropy and so on. These splitting criteria or cost functions make the decision tree a greedy algorithm. Since decision trees are easy to interpret, we make use of it as a surrogate interpretable model. We use labels obtained from the clustering algorithms as training data for the decision trees. The obtained decision tree now gives us a sequence of questions to helps us to conclude a subtype of tinnitus.

2.2.2 Random Forest

Random forests are an ensemble of decision trees. Decision trees tend to overfit or underfit a model, due to which the reliability of class label prediction is often questioned. Random forests construct multiple decision trees from a random subset of features during the training phase. Random forest chooses mean of all predictions for regression trees and mode of all labels for classification. These random decision trees thus correct the overfitting and underfitting problem of a single decision tree. The random forest also provides a list of features in the decreasing order of importance. Each feature is assigned a value for *mean decrease in accuracy.* In other words, if this feature is removed during clustering the accuracy of the result decreases by the given value. It helps in validating the decision tree obtained and supports the inference. Decision trees will help us to describe the obtained clusters by showing the features which are providing the most information for the splits.

2.3 Radial bar chart Visualization

As our data is high-dimensional, using basic plots such as scatter plots to visualize the cluster results are not sufficient. Therefore, we selected to visualize the dimensions in the form similar to SpikeNails as introduced in [4]. We modified this concept to show the clusters in the form of radial bar charts.

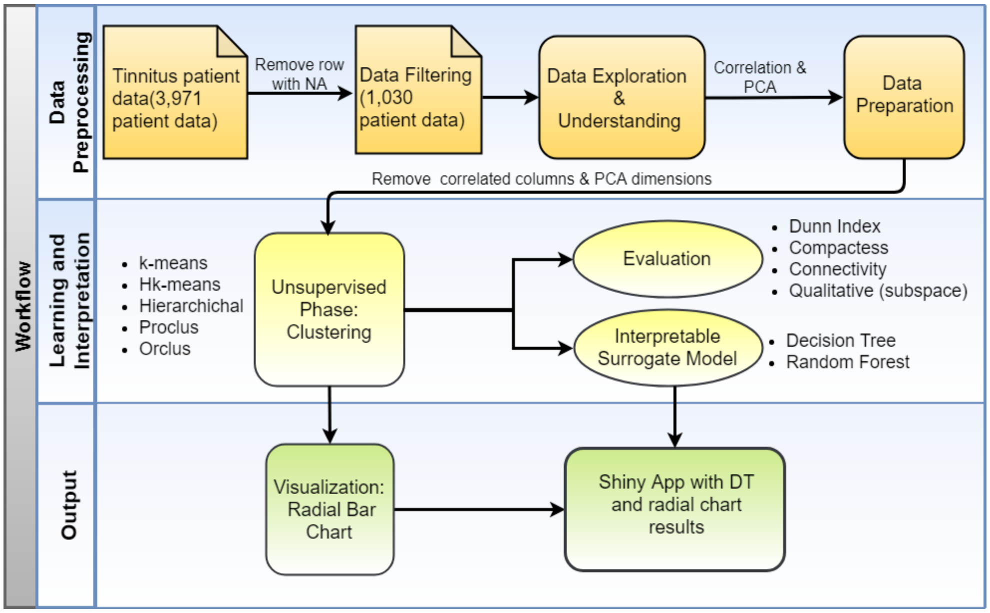
To better understand the clusters it is important that the user should know how different a cluster is from the overall population. The radial bar chart needs as input the cluster feature values calculated from scaled data. Thus the mean of each feature would be zero and the scaled cluster feature value would tell us how much it deviates from the mean. In the radial bar chart, each bar represents a feature of the cluster. The bar height corresponds to the scaled cluster feature value. Thus, the bars going towards the center of the chart are lesser than the mean, while the ones going away are greater than the mean. The user can thus describe the cluster based on the features which highly deviate from the mean.

Chapter 3

Concept and Implementation

3.1 Project Concept

Our work consists of mainly three phases namely *Data Preprocessing, Learning & Interpretation and Visualization.* as depicted in the workflow diagram in Fig. 3.1. In *Data Preprocessing* phase we clean the data by removing rows which have any *NA values* from present 3971 rows to 1030 rows. We further remove uncorrelated columns and scale the data based on z-score.



Workflow
Data Preprocessing
Tinnitus patient data(3,971 patient data)
Remove row with NA
Data Filtering (1,030 patient data)
Data Exploration & Understanding
Correlation & PCA
Data Preparation
Remove correlated columns & PCA dimensions
Learning and Interpretation
• k-means
• Hk-means
• Hierarchichal
• Proclus
• Orclus
Unsupervised Phase: Clustering
Evaluation
• Dunn Index
• Compactess
• Connectivity
• Qualitative (subspace)
Interpretable Surrogate Model
• Decision Tree
• Random Forest
Output
Visualization: Radial Bar Chart
Shiny App with DT and radial chart results

Figure 3.1: Concept workflow describing three phases of the implementation.

In *Learning & Interpretation* phase we run above-mentioned clustering algorithms to get cluster labels. With the help of these cluster labels, we then perform internal evaluation (Dunn Index and Connectivity). These criteria are used to assess the cluster quality for each global clustering method. Subspace algorithms are evaluated based on qualitative criteria by seeing their clustering results. As explained in section 2.2, we then make use of the surrogate interpretable model to make our clustering results more understandable for the user. We use decision trees built on the original data set with class labels as their cluster assignment. Random forest is then used to get the feature importance for each cluster in descending order. This helps us to see if the decision tree chooses the most important features at its decision points. Thus, the decision tree gives us a reduced decision path to arrive at a particular subtype of tinnitus, thus decreasing the number of questionnaires.

In the last phase, we are creating a radial bar chart visualization to show the found clusters. We also build a Shiny R application which accepts various user-specified hyper-parameters. This app lets a user select hyper-parameters like the number of clusters, algorithm, algorithm-specific hyper-parameters to show radial bar chart and decision tree output.

3.2 Implementation

3.2.1 Elbow-method using WSS

We used the elbow method to find the optimal value of k for our dataset for clustering. In this method we plot the*Within Sum of Square (WSS)* value of k-means result for increasing values of k. Initially, there is a steep drop in WSS with each addition of cluster. Then there is just a marginal drop in WSS which looks like an "elbow" in the graph. The best possible value of k is where this gradual drop in WSS starts. We observed that k=4 was a good candidate for the number of clusters.

Bootstrap sampling: We created samples of data (with replacement) from the original data set. The total WSS of each sample for each k was calculated and the elbow method was plotted with a box plot. Large variance or bigger boxes in the box plot would mean that the clustering is not stable. However, we did not find abrupt changes in the trend of the plot with bootstrap samples thus confirming stable clustering results.

3.2.2 Average Silhouette

The silhouette coefficient measures the cohesion in the clustering results. It will calculate how close each data point is to the points in its own cluster compared to those in other clusters. This value lies between [-1,1] and values closer to 1 are better.
On plotting the average silhouette score against the number of clusters for k-means we found that the highest value was obtained for k=2. Thus, we also selected k=2 as one possible candidate.

3.2.3 k-Means clustering

For k-means we used the "kmeans" function provided by the `cluster` library in R. This function takes as input the data, number of clusters *k*, number of random sets of centroids (*nstart*) and maximum number of allowed iterations*iter.max*. The *nstart* hyper-parameter chooses the number of times k means should run with a different set of initial centroids and *iter.max* decides the maximum number of iterations that are allowed for the algorithm to converge. We chose the *nstart* as 50 and *iter.max* as 15 for this use case.

3.2.4 Hierarchical agglomerative clustering

For hierarchical clustering, we choose the agglomerative algorithm. The library `cluster` in R provides a function “agnes” which generates a hierarchical tree. Agnes takes as input a distance matrix and the method to be used for calculating inter-cluster distance. The function provides a “agglomerative coefficient” which evaluates the clustering structure of the resulting tree. The closer the value to 1 the better is the result. We compared the “ac” value by using different combination of distance measures- “euclidean”, “manhattan”, “maximum”, “canberra”, “binary” and inter-cluster distance measures- “average”, “single”, “complete”, “ward”.

We observed that “ward” method gave the best *ac value* for all the distance measures. So we chose the combination of “euclidean” and “ward” method as the hyper-parameters. On plotting the dendrogram we observed that the inter-cluster distance is high for k=2, k=3, k=4 and then rapidly decrease until each point is a cluster. We finally selected k=2 and k=4 since for k-means we identified 2 and 4 as the optimal number of clusters.

3.2.5 Hierarchical k-means clustering

For hierarchical k-means, we used the “hkmeans” function provided by the R library. Similar to hierarchical and k-means clustering method, the combination of “euclidean” and “ward” with iter.max set to 15 was used as hyper-parameters. We obtained the results for both k=2 and k=4 which were then compared with the results of other approaches.

3.2.6 Proclus subspace clustering

We used the `subspace` package in R library to use the “proclus” function. This function takes as input the data, number of clusters to be found and the average number of dimensions per cluster. We selected the number of clusters to be 2 and 4 with an average number of dimensions to be 20.

3.2.7 Orclus subspace clustering

We used the `orclus` package available in R library. The “orclus” function takes as input the data, final number of clusters (k), the final number of dimensions per cluster(l) and the initial number of clusters (k0).

We tuned the hyper-parameters l and k0 by “sparsity.coefficient” and “within.projenss” measures obtained along with the clustering result where lower values for sparsity coefficient and wss are expected. We obtained optimum results for:
k=2 : l=25, k0=20
k=4 : l=20, k0= 31

3.2.8 PCA with k-Means clustering

We used “prcomp” function in R library to compute set of new PCs for a given data. This function takes as input the data as a mandatory hyper-parameter. After getting the new principal components, we plotted the scree plots and cumulative scree plots to see the variance explained. We chose explained variance threshold as 90% of the actual data and got the corresponding number of principal components as 40. We saved these first 40 principal components as a new dataset and then performed k-means as described in section 3.2.3.

3.2.9 Decision tree

We used `caret` package available in R library to use “trainControl” and“train” functions. “trainControl” takes as input resampling method (*method*) and number of folds for resamples (*number*). We set *method*=boot and *number*=10. “train ” function trains and fits the model with the input data and takes as input the class labels, data, the classification model (*method*), type of split (*split*), resamples to be used (*trControl*) and count of set hyper-parameters values that need to be evaluated (*tuneLength*). We set *method*=rpart, *split*=information, *trControl*= output of “trainControl” and *tuneLength*=10.

3.2.10 Random forest

We used the `random forest` package available in the R library to implement random forest. We tuned the hyper-parameters for the random forest for each algorithm and selected the values giving the optimal accuracy. These hyper-parameters were namely the number of features/data points sampled randomly as candidates at each split (*mtry*) and the number of trees to grow (*ntree*). After applying the random forest function, the error rate and the confusion matrix is provided as output. Also, the important features are sorted with respect to the drop in the “mean decrease accuracy” provided by the random forest. This list of “mean decrease accuracy” is used to validate our decision tree results. If the features in decision tree match with the top features appearing in this list then our decision tree is not overfitted or underfitted.

3.2.11 Visualization

We have created the radial bar chart visualization with the help of the d3 library and linked it to R using the `r2d3` library. The d3 library provides various inbuilt functions to create and modify Scalable Vector Graphics (SVG) objects [10]. These objects are nothing but a format to represent graphics based on XML. The SVG objects act as building blocks to create the entire visualization. The first step is to decide the format in which the chart receives the data to represent. D3 usually accepts data in the JSON format. However, the user can send the R objects directly to the D3 chart with the help of “r2d3()” function. We then created a data structure to store all the information related to a cluster. We used the labels obtained from the clustering algorithm and assigned it to the respective records in both the scaled and unscaled data. Next, we grouped the records based on labels and calculated the mean of each feature in each group. The result obtained gives us a summary of each cluster which we can call a cluster feature vector. We obtained the cluster feature vectors from the scaled data and unscaled data and sent it along with some additional information to the radial bar chart.

The second step is to process this data record by record and create the chart. In the radial bar chart created each bar height corresponds to the scaled cluster feature value. On hovering over the bar the actual unscaled cluster feature value can be seen along with the general mean value for the feature. The radial bar chart has a scale in the form of concentric circles and axis values going above and below zero. The circle at 0 thus represents the population mean and the bars deviate away from this circle. The bars are not ordered or grouped in any form.

3.2.12 Shiny App

A simple application was created using Shiny App which shows the decision tree and radial bar charts for the selected method of clustering. The method of clustering (k-means, Hk-means, hierarchical, Proclus, Orclus) and the number of clusters (2 to 10) are accepted as user input. The user can select the desired approach and cluster-specific hyper-parameters. The user can then toggle between the tabs to view the decision tree and the radial charts. Shiny app UI and Decision trees are dynamic and their shapes change as we change the window size. The UI of the Shiny app for decision tree is shown in Fig. 3.2 and radial bar chart in Fig. 3.3.

3.3 Evaluation

The results of global and subspace clustering are not comparable. This is because the clusters exist in different dimensions and thus not comparable. We performed the evaluation using two approaches namely, quantitative and qualitative. While quantitative evaluation was performed on global clustering approaches, qualitative evaluation was performed on all methods including subspace clustering. These approaches are detailed below.

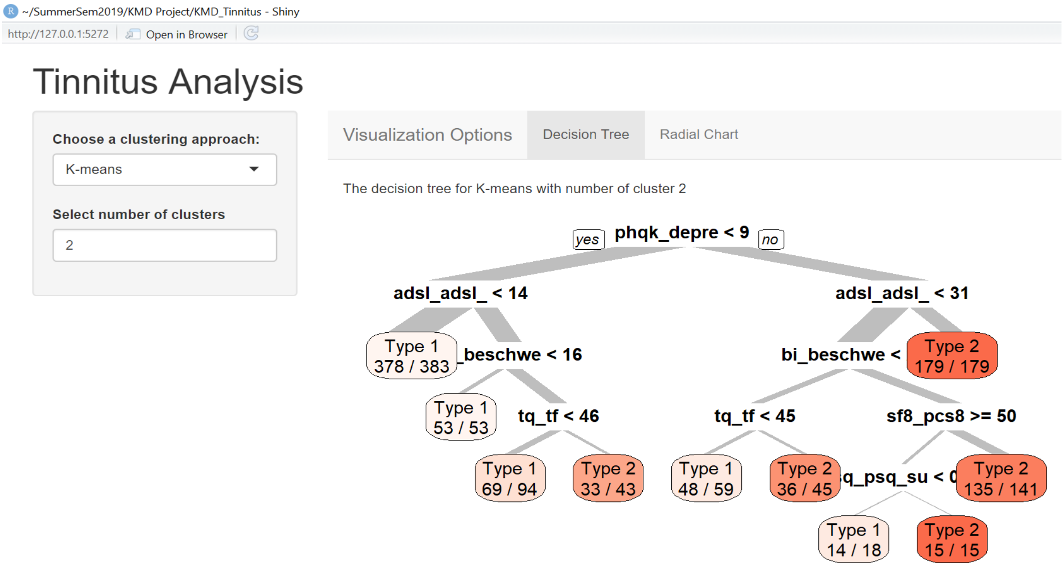
3.3.1 Quantitative:

To compare the clustering results obtained from k-means, agglomerative hierarchical clustering and Hk-means clustering, we chose the following quantitative evaluation measures provided by `clvalid` library in R:
Dunn Index: Dunn Index calculates the ratio of the smallest distance between the points in different clusters to the largest distance between the points within the cluster. Higher values for Dunn Index is desirable. Dunn Index ranges from zero to one.
Connectivity: The connectivity index measures the extent to which data points are placed in one cluster compared to its nearest neighbours. It ranges from zero to infinity. The lower the connectivity better are the results.
Compactness: Measures how close are the objects within the same cluster. It is based on measures such as cluster-wise within average distances between the observations. The value ranges from zero to infinity. Smaller values indicate a better clustering structure.

3.3.2 Qualitative:

For qualitative analysis, we analyzed the decision trees and the cutoff values of "tq_tf" and "adsl_adsl_sum" based on Niemann et al. [3] as follows: (i) 'tq_tf': 0-46 (compensated tinnitus); 47-84 (decompensated tinnitus) (ii)'adsl_adsl_sum': 0-15 (subclinical depression); 16-60 (clinical depression).

Since evaluation for subspace clustering is rather difficult, we make use of these values along with the decision tree to evaluate the results.

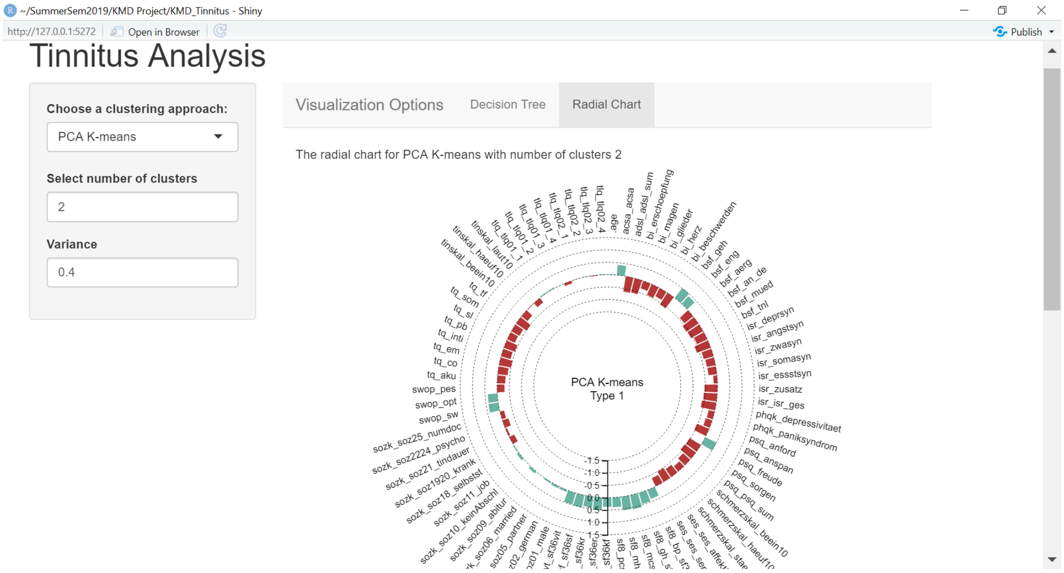


Choose a clustering approach: K-means

Select number of clusters: 2

The decision tree for K-means with number of cluster 2

Figure 3.2: Shiny App UI for decision tree for k-means with k value 2



Choose a clustering approach: PCA K-means

Select number of clusters: 2

Variance: 0.4

The radial chart for PCA K-means with number of clusters 2

[4]23

Figure 3.3: Shiny App UI for radial bar chart for PCA k-means with k value 2

Chapter 4

Application Study

4.1 Tinnitus Data

The data used for this project is extracted from questionnaires filled by patients with Tinnitus condition from January 2011 to October 2015 at the tinnitus center of the university medicine in Berlin. The cohort data is obtained from in total of 3971 patients. Patients involved were 18 years and above. These patients suffered from tinnitus condition for more than 3 months. The data comes from a total of 7 questionnaires filled by patients over the period when they started and ended the tinnitus specific treatment at the center. These 7 questionnaires are: (i)Long Form General Depression Scale (ADSL) [11], (ii) Perceived Stress Questionnaire (PSQ) [12], (iii) Short Form 8 Health Survey (SF8) [13], (iv) Tinnitus Questionnaire according to Goebel and Hiller (TQ) [14], (v) Tinnitus Localisation and Quality (TLQ) [15], (vi) Visual analogue scales measuring tinnitus loudness, frequency and distress (TINSKAL), and (vii) a sociodemographics questionnaire (SOZK) [16]. The Charité University Medicine Ethics Committee granted the ethical approval (reference number EA1/115/15) and written consents were given by the patients. All data used in the analysis has been pre-anonymised. From a total of 3971 records, records of the patients who missed any questionnaires were dropped. The remaining 1490 records were then pre-processed and any record with any missing values was deleted. The data then consisted of 78 features and 1030 records. We further identified 4 features with a correlation higher than 0.9 and dropped them for global clustering and PCA. We further dropped the patient identifier. Final data used for the analysis had 73 features and 1030 records.

4.2 Results

The clustering results of above-mentioned algorithms were visualized using the radial bar chart. Radial bar chart of hierarchical clustering is shown in Fig. 4.1 and 4.2. We choose to show the results of global clustering based on the quantitative evaluation values as mentioned in table 4.1. Hierarchical clustering showed best values for dunn index (0.249) and connectivity (286.73).

Table 4.1: Evaluation Table: Evaluation values for Global clustering methods

Clustering	Cluster	Dunn Index	Connectivity	Compactness
k-means	2	0.227	304.55	10.88
Hierarchical	**2**	**0.249**	**286.73**	11.09
Hk-means	2	0.227	304.55	10.88
PCk-means	2	0.227	304.55	10.88
k-means	4	0.216	752.64	10.37
Hierarchical	**4**	**0.214**	**673.78**	10.62
Hk-means	4	0.216	752.64	10.37
PCk-means	4	0.216	752.64	10.37

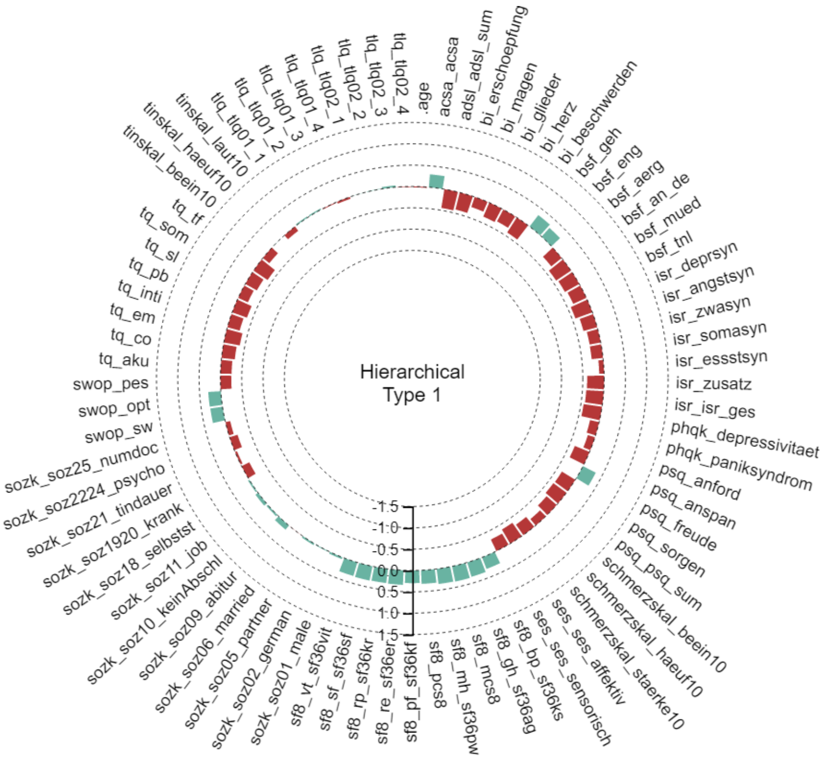


Figure 4.1: Radial bar chart for hierarchical clustering with k value 2 of Type 1

Type 1 cluster (Refer Fig. 4.1) can be described as a subtype with less severe body pain complaints (bi_beschwerden, bi_magen among others) and low depressive disorder sum score (adsl_adsl_sum). Therefore, patients have higher scores of optimism (swop_opt) and assessment of the quality of life (acsa_acsa). Since these patients have compensated tinnitus, their various parameters from the short form health survey (SF8) scores are also higher.

We found that *Type 2* cluster (Refer Fig. 4.2) has opposite characteristics compared to *Type 1*. *Type 2* cluster can be described as decompensated tinnitus with high body pain complains (bi_beschwerden, bi_magen among others). Depressive disorder sum score (adsl_adsl_sum) is higher and hence there is lower optimism (swop_opt) and lower assessment

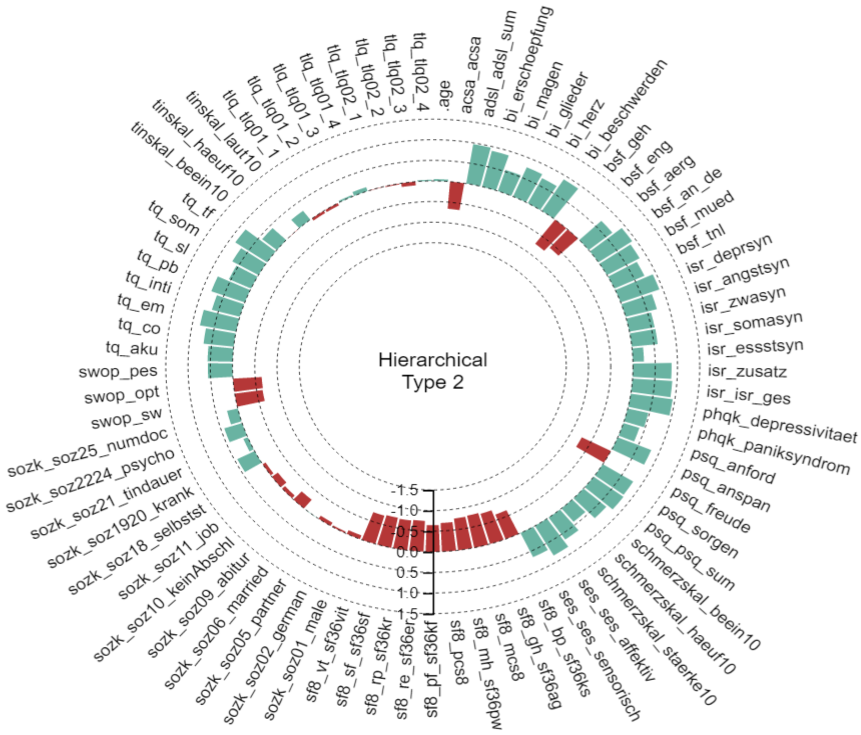


Figure 4.2: Radial bar chart for hierarchical clustering with k value 2 of Type 2

of the quality of life (acsa_acsa) among *Type 2* patients. Thus, their various parameters from short form health survey (SF8) scores are also lower. We found that location of tinnitus (tlq_tlq01_1 among others), type of tinnitus sound (tlq_tlq02_1 among others) and sociodemographics questionnaires (SOZK) does not deviate much from the mean of the features.

After performing qualitative analysis as mentioned in Section 3.3.2, we found that k-means gives the best decision tree according to the cutoff values of "tq_tf" and "adsl_adsl_sum". As shown in Fig. 4.3, all records with depressive disorder sum score (adsl_adsl_sum) less than 14 was classified as *Type 1* and greater than 31 was classified as *Type 2*. For any records falling in between these values, we check the pain complain (bi_beschwerden) value of the patient. Patients having values lower than 16 are *Type 1*, whereas for values higher than 25 results in *Type 2*. For values between 16 and 25, tinnitus severity score (tq_tf) are checked. Any record having values lesser than 46 were majorly classified as *Type 1* and greater than 45 as *Type 2*. Similarly, other decision points are used to describe subtypes. We could validate the decision points of the decision tree with the most important features obtained using the random forest as shown in Table 4.2. We can see that phqk_depress, adsl_adsl_sum, bi_beschwerden, tq_tf and others are in the top 10 most important features for the clustering. Thus we could conclude that there can exist one subtype (*Type 1*) of tinnitus with subclinical depression and compensated tinnitus and another subtype (*Type 2*) with clinical depression and decompensated tinnitus.

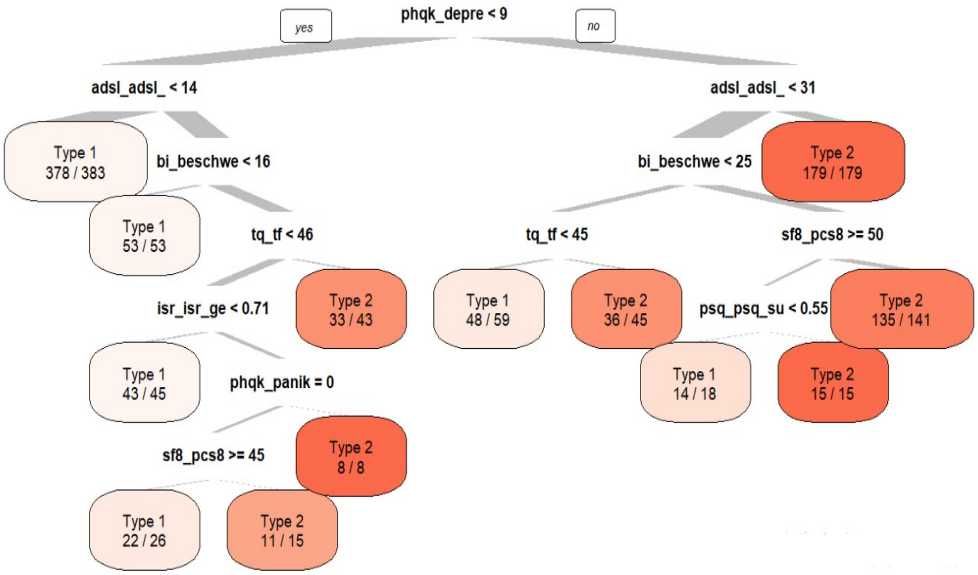


Figure 4.3: Interpretable decision tree for k-means clustering for k value 2. Values inside the nodes show the ratio of number correct classifications to the total number records in the node.

Table 4.2: Mean decrease accuracy for top 10 features obtained after performing random forest on k value 2 for k-means clustering

Feature	Description	Mean Decrease Accuracy
adsl_adsl_sum	Depressive disorder sum score	24.24
bi_beschwerden	Total complain pressure score	23.82
bi_erschoepfung	Exhaustion score	23.56
phqk_depresivitaet	phqk depression score	23.55
sf8_mcs8	Short form health survey mental score	22.74
isr_isr_ges	isr total score	22.38
bsf_an_de	Anxious depressivity score	21.39
isr_deprsyn	isr depression score	20.28
tq_tf	TQ total score	19.39
psq_psq_sum	PSQ stress sum score	18.88

For subspace clustering, we choose to show and interpret the results obtained from Proclus clustering. This is because Proclus uses the original dimensions to form subspaces whereas Orclus transforms the original dimensions into new projected subspaces. These transformed dimensions cannot be interpreted medically. For k=2, Proclus identified two clusters as shown in Fig. 4.4 and Fig. 4.5. Each chart shows only the selected dimensions for that cluster. For *Type 1* (Refer Fig. 4.4), the selected features related to pain complaints (bi_magen, bi_herz, bi_beschwerden) and panic syndrome(phqk_paniksyndrom) show values lower than the general mean. This type also shows higher than mean values for optimistic thinking

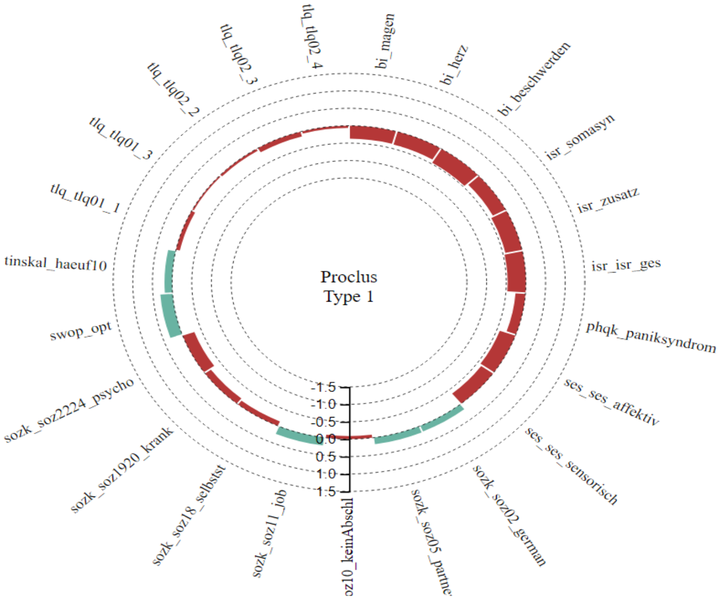


Figure 4.4: Radial bar chart for Proclus clustering with k value 2 of Type 1

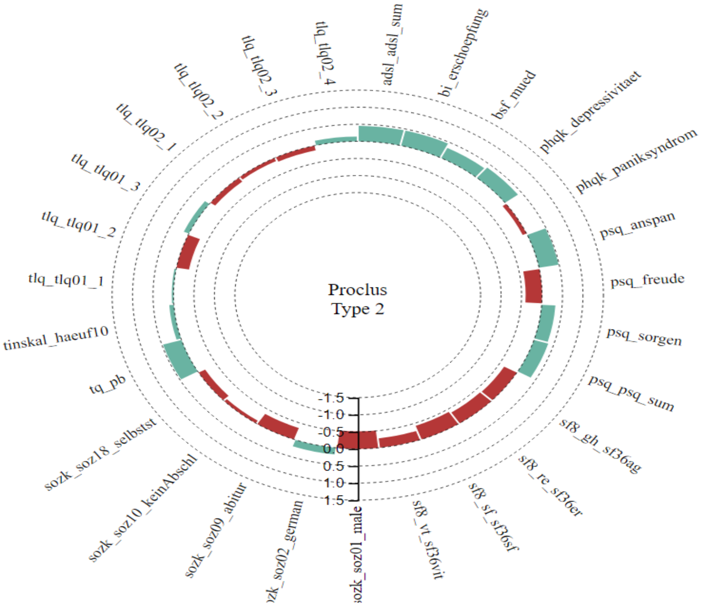


Figure 4.5: Radial bar chart for Proclus clustering with k value 2 of Type 2

(swop_opt) and have a job and partner (sozk_soz11_job and sozk_soz05_partner).

For *Type 2* (Refer Fig. 4.5, the selected features related to depression (adsl_adsl_sum, phqk_depressivitaet, phqk_paniksyndrom), perceived stress (psq_psq_sum, psq_sorgen), pain complaints (bi_erschoepfung) and tinnitus psychological distress score (tq_pb) are higher than the general mean. These patients also showed a lower score for short form health survey scores (SF8).

The decision tree for Proclus as shown in Fig. 4.6 also classifies the records with depression score (adsl_adsl_sum) less than 15 as *Type 1*. For depression score more than 15 and pain complaints (bi_beschwerden) greater than 36 it was classified as *Type 2*. Further, it was classified based on anxiety score. We can thus describe *Type 1* as a cluster with low pain complaints, low stress and people with job and family whereas *Type 2* as a cluster with higher depression, stress, pain complaints, and tinnitus causing psychological distress. This is similar to the descriptions obtained from global clustering results.

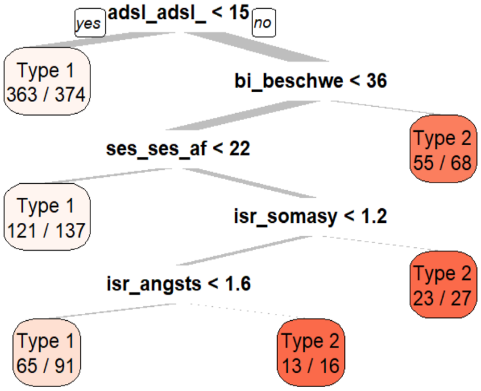


Figure 4.6: Interpretable decision tree for Proclus clustering for k value 2.

4.3 Findings

After our clustering analysis, we made a few observations in the radial bar chart. (i) All global clustering algorithms and PCA showed similar clustering results. Some features deviated more from their general mean value, particularly adsl_adsl_sum, bi_beschwerden, bi_eschoepfung, bsf_geh, psq_freude among few others. While certain features such as TLQ questionnaires, sozk_soz06_married, sozk_soz18_selbstst and few others with values closer to the general mean population do not give us any new information and thus can be ignored. In other words, one of the clusters for k value 2 had higher than general mean value for certain features like adsl_adsl_sum ,bi_beschwerden among others, whereas low values for tq_tf. (ii) We inferred from our evaluations that there may exist 2 natural clusters. For k value 2 and for k value 4, we found that in 4 clusters, each natural cluster is being divided into two. This can be seen from the feature and their respective values for the clusters as shown in Fig. 4.7 and 4.8. (iii) One of the subtypes found in subspace for a particular k value is always found to be similar to that of global clustering. We can see from Fig. 4.7 and 4.9 that the clusters are similar, just the standard deviation values for each features are different.

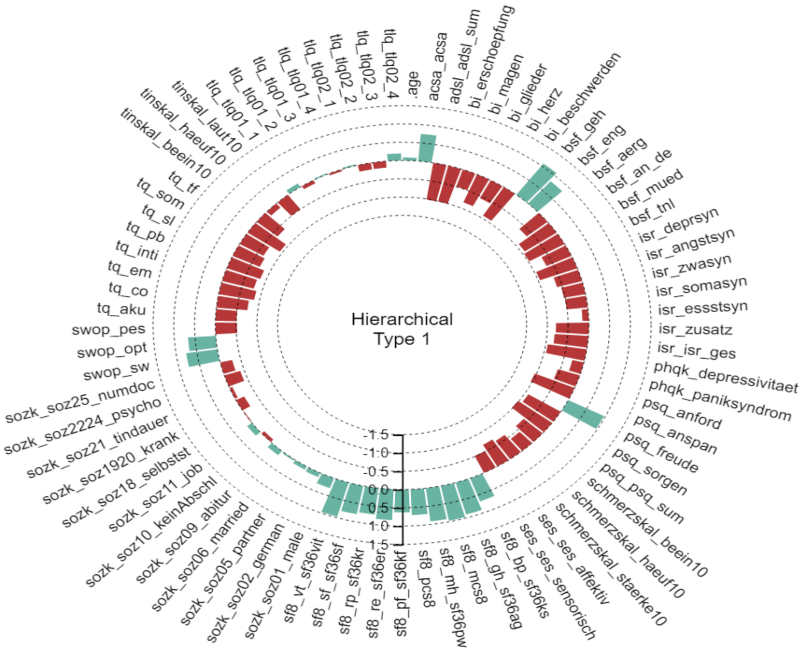


Figure 4.7: Radial bar chart for hierarchical clustering with k value 4 of Type 1 showing higher bar heights.

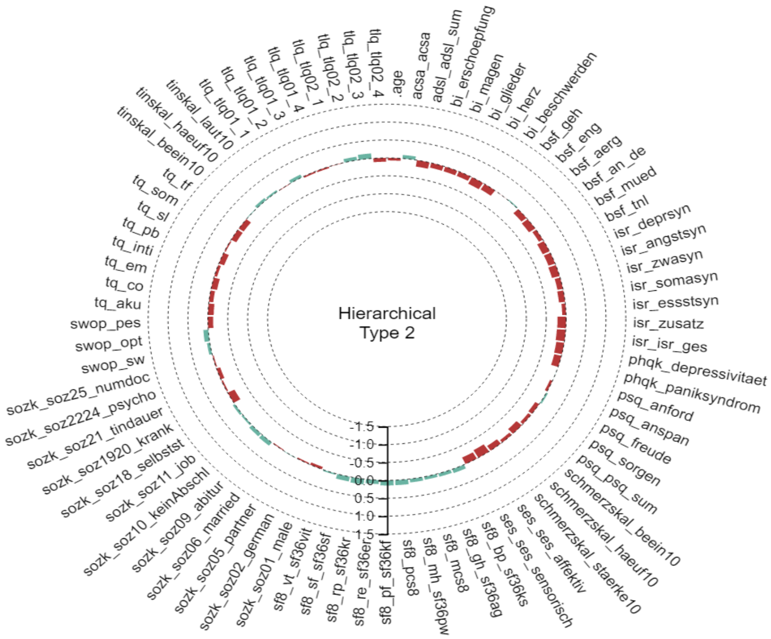


Figure 4.8: Radial bar chart for hierarchical clustering with k value 4 of Type 2 showing lower bar heights.

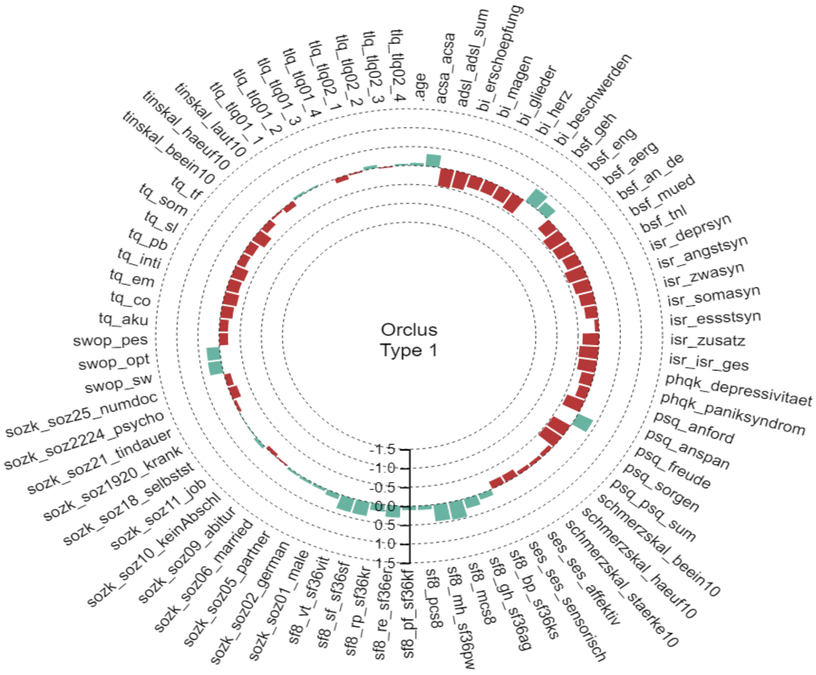


Figure 4.9: Radial bar chart for Orclus with k value 4

Chapter 5

Discussions

Clustering algorithms can find patterns in data that are not easily identifiable by a human. Finding patterns in tinnitus patient data using clustering gave us interesting subtypes which are distinguishable from each other. In our limited knowledge of the domain, we found that k=2 gives the best result for reduction of questionnaires according to subtypes. However, a medical expert's opinion can differ and more suitable subtypes can be found.

We also found that all global clustering approaches were giving similar clustering output in the radial bar chart. Apart from a few cluster assignments that changed the bar values, the trends of the subtypes were similar. This further confirms that the subtypes exist and follows a similar pattern when clustering is done.

It was surprising for us that clustering done on reduced dimensions (PCk-means) did not improve the results. This was visible in the evaluation results in Table 4.1. We hypothesize that this might be because of two reasons: (i) majority of the features from 73 features were inherently categorical. Because of this, we had a lot of principal components that explained less than a single explanatory variable, thus resulting eigenvalue less than 1 in scree plots (Refer Fig. 5.1). (ii) We already removed highly correlated columns as a part of preprocessing. It will be interesting to see if there is a change in evaluation results if we perform advanced algorithm of global clustering methods. However, that is not currently in the scope of this work.

5.1 Summary of major findings

The subtypes found for all the global clustering algorithms were similar with respect to their feature vectors. And for a higher number of clusters, algorithms divided subtypes depending on their values. Lastly, one of the subtypes from subspace clustering always was to be found to be similar to that of other global clustering algorithms.

5.2 Limitations

The major limitation of our work is that our results are biased towards selected dataset. This is due to the fact that we excluded patients whose data were incomplete in the seven questionnaires. This was because they did not answer them both before and after treatment. It can very well happen that, after the inclusion of more patients, more subtypes of tinnitus

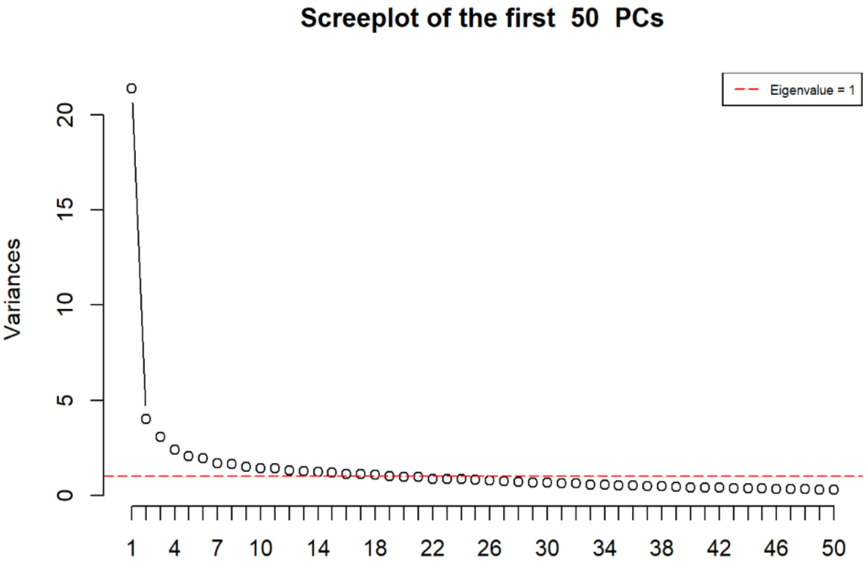


Figure 5.1: Scree Plot for each components explaining individual variances. Red horizontal line describes that each component describes less than a single explanatory variable.

can be found. Another limitation is the absence of ground truth. Since we did not have any ground truth, we cannot say with surety which method is better for a particular number of clusters. Another limitation of this work is limited data. A larger amount of patients data from various backgrounds, including more representation of different genders, nationalities, age groups among others can give us different results. Also, since the data collected was only from one organization, we can have some local demographic problems that could be inherent to the studied patients, such as economic status, work stress, noise pollution, unfriendly neighbourhoods and so on. These factors can hugely influence the results. Data from various organizations throughout Germany can remove this local demographic bias in the subtypes. And one of our last limitations is the usage of simple clustering algorithms. Since this is one of the first attempts to find subgroups with a statistical approach, we confined ourselves to the basic and popular clustering algorithms. However, these algorithms have their own problems and advanced versions of these algorithms can give better results.

5.3 Future work

To solve the limitation of ground truth, we can have user studies to validate the ground truth. Another future work is to cluster on data gathered from other health centers and validate our results and find discrepancies in our results with the help of a domain expert. And finally, density-based clustering algorithms and other advanced clustering algorithms can be implemented.

Chapter 6

Summary

In this project, we performed clustering on tinnitus patients dataset to find any possible subtypes. We observed that global clustering algorithms give best results for k=2. Based on evaluation measures, hierarchical clustering produced the best clustering results. Based on interpretable models using “tq_tf” and “adsl_adsl_sum” scores, we found k-means gave the best result. Radial bar chart visualization for each cluster gave a better idea about how each feature behaves with respect to the global feature mean population. Since our main aim was to identify the subtypes and reduce the questionnaires, we successfully achieved that by building interpretable models. Questions were reduced from 78 to approx 5 to 10. In future work, we aim to improve the results so that more subtypes with fine granularity can be found by using advanced algorithms.

Bibliography

[1] J. A. Lopez-escamez, T. Bibas, R. F. F. Cima, and P. V. D. Heyning, “Genetics of Tinnitus : An Emerging Area for Molecular Diagnosis and Drug Development,” *Frontiers in neuroscience*, vol. 10, no. August, pp. 1–13, 2016.

[2] R. Tyler, C. Coelho, P. Tao, H. Ji, W. Noble, A. Gehringer, and S. Gogel, “Identifying Tinnitus Subgroups With Cluster Analysis,” *American journal of audiology*, vol. 17, no. 2, pp. 1–19, 2008.

[3] U. Niemann, P. Brggemann, B. Bcking, B. Mazurek, and M. Spiliopoulou, “Development and internal validation of a depression severity prediction model for tinnitus patients based on questionnaire responses and socio-demographics,” *Scientific Reports. Under review*, 2019.

[4] A. Tatu, L. Zhang, E. Bertini, T. Schreck, and D. Keim, “ClustNails : Visual Analysis of Subspace Clusters,” vol. 17, no. 4, pp. 419–428, 2012.

[5] C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park, “Fast Algorithms for Projected Clustering,” *SIGMOD Record (ACM Special Interest Group on Management of Data)*, vol. 28, no. 2, pp. 61–72, 1999.

[6] P. Lance, H. Ehtesham, and L. Huan, “Subspace Clustering for High Dimensional Data: A Review,” *SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, vol. 6, no. 1, pp. 90–105, 2004.

[7] C. C. Aggarwal and P. S. Yu, “Finding generalized projected clusters in high dimensional spaces,” *SIGMOD Record (ACM Special Interest Group on Management of Data)*, vol. 29, no. 2, pp. 70–81, 2000.

[8] L. Parsons, E. Haque, and H. Liu, “Evaluating Subspace Clustering Algorithms,” *Proceedings of the fourth SIAM international conference data mining, workshop clustering high dimensional data and its applications.*, no. April, p. 9, 2004.

[9] C. Molnar, *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.* 2019.

[10] M. Bostock, V. Ogievetsky, and J. Heer, “D3 data-driven documents,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[11] L. S. Radloff, “The CES-D scale: A self-report depression scale for research in the general population,” *Applied psychological measurement*, vol. 1, no. 3, pp. 385–401, 1977.

[12] H. Fliege, M. Rose, P. Arck, O. B. Walter, C. Kocalevent, Rueya-Daniela Weber, and B. F. Klapp, "The Perceived Stress Questionnaire (PSQ) reconsidered: validation and reference values from different clinical and healthy adult samples," *Psychosomatic medicine*, vol. 67, no. 1, pp. 78–88, 2005.

[13] M. Bullinger and M. Morfeld, "Der SF-36 Health Survey," *Gesundheitsokonomische Evaluationen*, pp. 387–402, 2008.

[14] G. Goebel and W. Hiller, "Tinnitus-Fragebogen:(TF); ein Instrument zur Erfassung von Belastung und Schweregrad bei Tinnitus; Handanweisung," *Hogrefe, Verlag für Psychologie*, 1998.

[15] G. Goebel and W. Hiller, "Psychische Beschwerden bei chronischem Tinnitus: Erprobung und Evaluation des Tinnitus-Fragebogens (TF)," *Verhaltenstherapie*, vol. 2, no. 1, pp. 13–22, 1992.

[16] P. Brüggemann, A. J. Szczepek, M. Rose, L. McKenna, H. Olze, and B. Mazurek, "Impact of multiple factors on the degree of tinnitus distress," *Frontiers in Human Neuroscience*, vol. 10, no. June, p. 341, 2016.