Adapting RObust Clustering using linKs for Categorical Variables with Inherent Latent Ordinality

By
Maria Abrar
2010-NUST-MS-PhD-IT-30

Supervisor
Dr. Sharifullah Khan
Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Information Technology (MS IT)

In
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(August, 2014)

Approval

It is certified that the contents and form of the thesis entitled "Adapting RObust Clustering using linKs for Categorical Variables with Inherent Latent Ordinality" submitted by Maria Abrar have been found satisfactory for the requirement of the degree.

Advisor: Dr. Sharifullah Khan

Signature: _____

Date: _____

Committee Member 1: Dr. Khalid Latif

Signature: _____
Date: _____

Committee Member 2: Dr. Ali Mustafa Qamar

Signature: _____
Date: _____

Committee Member 3: Dr. Sohail Iqbal

Signature: _____
Date: _____

Dedication

I dedicate this thesis to my father who was my main motivation in doing Masters and my mother whose prayers are always with me where ever I go. They have given me the opportunity of acquiring education from best institutions of the country and have been always there for me in every step of my life.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's de-sign and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: Maria Abrar

Signature: _____

Acknowledgments

This work would not have been possible without the support of Dr. Sharifullah Khan played a very vital role in this thesis and went out of the way to help me.

Dr Hammad Qureshi, who have helped me throughout MS thesis.

To my committee members: Dr. Khalid Latif, Dr. Ali Mustafa Qamar, Dr. Sohail Iqbal for their feedback, encouragement and support.

To my beloved family, for their continuous support and encouragement. They were my source of inspiration and without their prayers this would not have been possible.

The last but certainly not the least, Allah (SWT), who is the One and Only and controls everything.

Declaration

Maria Abrar, H. A. Qureshi, “Associating Perinatal Mortality With Diet By Adapting Robust Clustering Using Links For Categorical Variables,” VFAST Transactions on Software Engineering, Vol. 3, No. 1, January 2014

Table of Contents

List of Figures

Abstract

Robust Clustering using Links (ROCK) is a hierarchical clustering technique which is used for clustering categorical data. It considers the neighbors of pair of points, and clusters points having a larger number of similar entities i.e. cluster on the basis of links. Two points are linked if they have similar items. However, if there is inherent latent ordinality in data then ROCK tends to fail and does not produce good results. Ordinality may be found in categorical data which is often hidden i.e. not known beforehand. In this study, we present a novel ROCK-based algorithm which can handle such latent ordinality and we show that the technique leads to improved results over traditional ROCK. We apply the technique to a problem from the domain of infant mortality and investigate the impact of the food taken by an expectant mother on infant mortality.

Perinatal Mortality, also known as perinatal death, is death of a neonate within 6 days (early neonatal mortality) or from 7 – 27 days of birth (late neonatal mortality). Food consumed by an expectant mother is said to have an impact on the pregnancy outcome apart from other factors. Using ROCK we cluster expectant mothers as per the food intake. As expected some food items have a greater impact on pregnancy outcome than others. When clustering categorical data such as the one used in this study it is difficult to estimate the inherent latent ordinality in the data i.e. what foods have more impact as compared to other items found in the data which as shown in this work affects the accuracy of the technique applied. To resolve the issue in the context of our problem we propose our novel technique ROCK for Latent Inherent Ordinality (ROCKLIO). The technique involves two stages clustering where the first stage is used to ascertain what food groups have a greater impact on clustering. In the second stage, a new link measure is derived that improves the clustering accuracy by applying weights on the basis of the significance of each food item. Better clustering results are obtained with the novel Robust Clustering using Links with latent Inherent Ordinality as compared to the simple ROCK.

Chapter 1

Introduction

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering [24]. A cluster contains objects which are similar to each other. Clustering is a very useful technique to find out hidden patterns in a large data set.

1.1 Categorical Attributes

Each kind of data comes with its own challenges. In our study we will be dealing with clustering of categorical data. Categorical variables, which are also known as nominal variables, have two or more than two fixed number or variables like gender, colors etc. There is no basic ordering within these values. We cannot order female and male or the colors green, red and white. Categorical data's structure is different from continuous data hence the typical distance functions applied on continuous data fail on categorical data. However categorical data can be clustered by many other available clustering solutions which deal with such data types. These techniques are mentioned in the next section in detail.

In categorical data bases, although there is no inherent order in the data but some of the attribute values would have a higher significance related to others. If we take items sold in a super store, we can identify the items sold together using standard clustering algorithms for categorical attributes, but we won't be able to identify which of the store items are generating more revenue within a cluster, or which one is more in demand using the clustering algorithm. In this case, grouping similar items in clusters is useful

but also identifying the more valuable store items through a clustering algorithm would give a better result. Available clustering technique successfully group the common objects into a cluster (for instance items sold together or items which are sold more) but they fail to identify the significant attributes from the cluster (which are the top three more valuable items sold within a cluster based on some metric, or assign order according to the relative importance in the cluster).

In our study we focus on highlighting these important attribute values from the data by improving an existing data mining algorithm. We will apply the modified algorithm to data acquired from the field of medicine. We studied the causes of perinatal mortality by taking the food intake by expectant mothers, and underlining the important food items in the process of pregnancy. Perinatal mortality is explained in detail below.

1.2 Perinatal Mortality

Perinatal mortality or perinatal death refers to the death of a newborn baby within 7 days of its birth or after 28 or more weeks of gestation. Perinatal mortality rates are sensitive indicators of the social and economic development of any country. High infant mortality rates indicate presence of health care and socioeconomic problems.

Perinatal mortality is an important health issue in Pakistan which has a great social impact. In a report by WHO (2005), it was stated that 99% of the world's perinatal deaths occur in low-income and middle-income countries [3]. According to another WHO report (1998), two-thirds of the world's perinatal deaths occur in only 10 countries, amongst which Pakistan is ranked 3rd [2]. South East Asian countries have the highest number of fatal births known as still births. In developing countries, average still birth rate is 26 per 1000 live births which are five times higher than the rate in the developed countries. In a WHO Bulletin published in 2005 [2], out of 1280 births studied in urban areas, 82 children died within 7 days and 96 died within 28 days of birth (Table 1). According to a 2010 UNICEF report, Pakistan has a perinatal mortality rate of 60-80 per 1000 births [7]. Millennium development goals (MDG) are set each year by UNDP to increase the living standards in developing countries. According to MDGs, Pakistan needs to reduce perinatal mortality rate (PNMR) by two thirds by 2015 (Target 4-A) [8].

Different pre-conceptions are held in different societies pertaining what is a healthy diet during pregnancy. Daily food intake in most societies and cultures is inadequate to meet the requirement for vitamins and minerals during pregnancy due to which extra supplements are prescribed by doctors. Women who are pregnant need certain nutrients (such as good fats, proteins and right amount of carbohydrates) in greater quantity hence

the intakes must be supervised by a registered doctor. Most pregnant mothers in developing countries do not visit clinics regularly for these supplements and hence are not getting appropriate diet which is the cause of high perinatal mortality rate in the country.

Births and mortality rates	Births/ deaths	Males	Females	Total[a]
All births,[b] n (%)	1280	619 (51.5)	583 (49.5)	1280 (100)
Stillbirths, [c] rate per 1000 births (95% CI)	43	35.5 (20.7–50.4)	25.7 (15.4–38.7)	33.6 (23.6–43.6)
Early neonatal mortality, rate per 1000 live births (95% CI)	39	35.0 (19.5–50.4)	29.3 (14.7–43.9)	34.8 (24.1–45.5)
Late neonatal mortality, rate per 1000 live births (95% CI)	14	5.5 (0.0–11.8)	19.5 (7.5–31.5)	12.5 (6.0–19.0)
Neonatal mortality, rate per 1000 live births (95% CI)	53	40.5 (23.9–57.1)	48.8 (30.2–67.5)	47.3 (34.9–59.7)
Perinatal mortality, rate-1[d] (95% CI)	82	72.6 (51.2–94.0)	56.9 (37.1–76.7)	70.4 (55.7–85.1)
Perinatal mortality, rate-2[e] (95% CI)	96	77.9 (55.8–100.0)	75.9 (53.3–98.5)	82.5 (66.7–98.3)

Table 1 Stillbirth and neonatal and perinatal mortality rates, by gender and in total, in a prospective study in an urban Pakistani population

2003–2005 [21] CI, confidence interval.

[a] Gender not recorded for 78 births.

[b] Includes the 78 infants without gender data.

[c] The rate of stillbirths was calculated from the number of women with known birth outcomes (n = 1280), while neonatal and perinatal mortality rates were derived from the number whose outcomes were known at 28 days (n = 1121) plus stillbirths, where appropriate.

[d] Rate-1 is stillbirths plus mortality within 7 days per 1000 births.

[e] Rate-2 is stillbirths plus mortality within 28 days per 1000 births.

1.3 Thesis Contribution - Perinatal mortality and Clustering the Data

In this research, we investigate food groups and their effects on pregnancy. This would identify the food items which are more important for pregnant mothers. The food groups under study are:

- Meat
- Vegetables
- Fruits
- Dairy Products
- Cereal
- Fats/Oil

We will study these food items, as variables affecting perinatal mortality, with the help of data mining by applying clustering to a data set of pregnant mothers. The data provided was in textual and categorical form, which made it ideal for our problem statement. Using the algorithm the good food items will be differentiated from the bad food items by relating them to the outcome of the pregnancy using quantitative data analysis (on the food groups) and assign weights according to the impact on the outcome of the pregnancy. Our novel technique provides good clustering accuracy and may lead to the development of novel computer assisted diagnosis techniques.

1.4 Thesis Problem Statement

We have adapted RObust Clustering using linKs for Categorical Variables with Inherent Latent Ordinality and determine the outcome of pregnancy using ROCKLIO. This is an algorithm which will highlight the good food items from the bad food items by relating them with the outcome of a pregnancy after evaluating quantitative data (food groups) and highlighting the order of the variables according to their importance in the outcome of pregnancy.

1.5 Thesis Organization

The thesis report is divided into four major sections. The first section gives a thorough introduction about clustering of categorical attributes and explaining perinatal mortality.

The section of Literature review coming next deals with existing algorithms for categorical attributes and other data mining related work done in the field of perinatal mortality.

We then move on to methodology which explains the whole process of ROCKLIO and ROCK. This section will give an extensive explanation for the methods from preparing the data for the algorithm, application of the algorithm used to calculate hidden ordinality in categorical attributes and how the results of ROCKLIO have been accessed.

After discussing the methodology, we move on to results of our work. This section will cover the end result of the applied algorithm and prove that hidden ordinality can be identified by clustering algorithm.

Chapter 2

Literature Review

Our problem statement's literature review is divided into two sections. The first section deals with the existing data mining studies which revolve around the causes of perinatal mortality, and the other section deals with the data mining techniques for categorical attributes.

2.1 Data-Mining Studies for Perinatal Mortality

Many studies have been carried out to investigate the causes of perinatal mortality with the help of data-mining techniques in the past.

In Denmark, a study was conducted by Kristensen et al., investigating the impact of various factors like smoking, caffeine, alcohol, and age etc. on pregnancy by discretization of variables such as BMI (Body Mass Index), age, parity, smoking, alcohol intake etc. [9]. It was found that obesity doubles the risk of stillbirth and neonatal deaths. Similarly, same conclusion was derived by Doherty et al, in their research of pregnancies 16-18 weeks old in which body mass index (BMI) was correlated with the outcome of a pregnancy by creating associations (calculating ratios) between body mass index (BMI), weight gain, and preterm delivery [10]. K-Nearest Neighbor (K-NN) was used by Qureshi et al. to predict the outcome of pregnancy after examining the relationship between body mass index (BMI), pre-pregnancy weight and weight gain during pregnancy

	Researched By	Main Objective	Technique Used	Conclusion
1	Kristensen et al.	Studied factors affecting pregnancy (smoking, caffeine, alcohol, age, BMI etc.)	Classification and using confidence levels	obesity doubles the risk of still birth
2	Doherty et al.	research of pregnancies 16-18 weeks old, body mass index (BMI) was correlated with the outcome of a pregnancy	Associations and rules inference	Obesity is the major cause of still birth
3	Qureshi et al	predict the outcome of pregnancy using body mass index (BMI), pre-pregnancy weight and weight gain during pregnancy	K-Nearest Neighbor (K-NN)	Weight gain is the most critical factor in pregnancy
4	Siega-Riz et al.	Understand the causes of still birth	Chi square tests	underweight and inadequate weight gain in the third trimester was the main cause of preterm births
5	Jerzy et al.	Comparing less specific and more specific rules	Prediction using more Vs. prediction using less specific rules	Less specific rules are easier to understand and efficient
6	Jerzy *et al*.	Manual methods only 17-38% accurate	LERS (Learning from Examples based on rough sets)	new classification method increased the prediction rate up to 68-90%

Table 2 Literature Review - Data mining solutions for Perinatal Mortality

[11]. The study concluded that weight gain is an important factor in predicting outcome of a pregnancy. Jerzy et al. studied a machine learning technique which predicted pre-term birth on the basis of less and more specific rule induction. According to Jerzy et al., less specific rules are easier to remember and more efficient with respect to more specific

rules in predicting pre-term births [11]. Another study by Jerzy et al. concluded that existing manual methods of predicting pre-term births are only 17-38% accurate, hence a new prediction method was used with the help of LERS (Learning from Examples based on rough sets) [12]. This new classification method increased the prediction rate up to (68-90%). All of the studies discussed in the above analysis were conducted using statistical analysis techniques such as discretization, chi-square tests, frequency and probability estimation and data mining techniques such as regression analysis for prediction or classification using K-nearest neighbors. Our data consists of categorical attributes for which a clustering technique for categorical variables may be used. Our proposed technique would cluster mothers as per the food items they consume and hence lead to insights in to the causes of high infant mortality.

2.2 Data Mining Techniques for handling Categorical Attributes

There are many different solutions present for clustering all kinds of data. We have focused our research on categorical and ordinal data. Clustering categorical data is a topic of extensive research. We selected five techniques which were relevant to the study and the data. In search of the perfect algorithm we looked for two features:

- Deals with categorical data
- Hierarchal clustering technique

The five techniques studied are discussed below.

COOLCAT by Barbará et. al (2002) identifies clusters in a large data set of categorical data and then calculates the entropy of the clusters created [14]. It works in two steps: initialization and the incremental step. In the first step clusters are created from a sample data set by finding the most dissimilar records in the sample set. This is done by maximizing the minimum pairwise entropy of the chosen items. In the second step the remaining records are processed by marking them in a suitable cluster by computing the expected entropy and placing them in the cluster with the closest entropy. This algorithm works well with larger data sets and with continuous data (such as data streams) but does not highlight the ordinality in the data.

Huang et al. (1998) use extension of fuzzy k-means clustering, k-modes clustering for clustering categorical data in which confidence is assigned to the objects of different clusters [16]. These confidence values provide information about the core and boundary objects of the clusters of categorical data. Although this algorithms emphasizes on

boundary and core objects of clusters and ignores the ordinality of the objects within the cluster.

Another clustering algorithm CURE by Guha et al. [15] mainly focuses on proximity and ignores the interconnectivity detail and ordinality. It identifies non spherical clusters of various sizes. A random sample is taken and partitioned. This is then clustered in iterations to obtain the best clusters. CURE is a hierarchal clustering algorithm in which a constant number of well scattered items are chosen.

	Research done by	**Technique/ Algorithm Applied**	**Algorithm Details**
1	Barbará et. al (2002)	COOLCAT	Used for categorical attributes, incremental heuristic algorithm ,Cluster streams of data and works with entropy
2	Huang et al. (1998)	Fuzzy K means clustering	Used for categorical attributes, uses dissimilarity measure and assigns confidence to the object. Ignores ordinality in the data.
3	Guha et al (2001)	CURE	Hierarchal clustering algorithm, can identify non spherical clusters, not suitable for categorical data due to centroid based clusters
4	Guha et al. (2005)	ROCK	Hierarchal clustering algorithm, used for categorical attributes, Will not highlight any ordinality in data
5	Zhang et al. (1996)	BIRCH	Hierarchal clustering algorithm, used for categorical attributes, only performs in considers spherical clusters

Table 3 Literature Review: Clustering Algorithms

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is a hierarchal clustering algorithm which was developed by Zhang et. al [22]. BIRCH clusters large amount of data in two steps: micro-clustering stage and macro-clustering stage. It creates a Cluster Feature tree which stores the features for hierarchal clustering. The non-leaf nodes, which have children, of the CF tree stores the information of its children by storing the branching factor and the threshold. The branching factor tells the number of children of the non-leaf node and the threshold tells the maximum diameter of the sub clusters.

ROCK, developed by Guha et al. (2005), focuses on interconnectivity amongst features whereas it ignores cluster proximity [3]. It is closest to our problem since it works best with categorical data uses the concept of links instead of distances. A similarity measure and threshold value is used, which identifies neighbors. If the similarity value of pair of points exceed the threshold they are considered as neighbors. The total common neighbors for the items are the links between the neighbors. All the points in the same cluster will have a larger number of neighbors' hence higher value for the links. So the points with the most number of links are clustered together.

All the techniques described above work well with large categorical databases but they do not highlight the important items of the cluster or in other words highlight the hidden ordinality. In this study, ROCK is selected since it focuses on interconnectivity which will help us in our medical data. This interconnectivity will be modified in a way to highlight ordinality in the data along with clustering similar food groups in a cluster.

Chapter 3

Methodology

The whole process of ROCK Latent Inherent Ordinality (ROCKLIO) was done in six phases.

- Understand the data,
- Preprocessing the data provided,
- Modifying the data by reduction and transformation,
- Data Sampling,
- Applying ROCKLIO to the prepared data,
- Accessing the results of ROCKLIO to calculate error margin and success rate of the outcomes.

The below sections explain these phases in detail.

3.1 Understand the data

Data mining is a knowledge discovery process for which domain knowledge is the most important step. Hence we need to get familiar with the data before applying any data mining technique. So the first step in our process was to understand the attributes and the values present in the given data set. We reviewed the data for any anomalies and descriptive statistics were employed in order to gain better understanding of the data. This helped in identifying any discrepancy in the data or any important attribute which may be important for subsequent research.

As a first step, important and unnecessary parameters were marked in the data. This was done by studying the attributes in accordance. The data set consisted of different

parameters related to pregnancy such as weight gained during pregnancy, literacy level of the family, medical history and food intake during pregnancy.

Meat	Beef, Chicken, Fish, Mutton
Vegetables	Beans, Brinjal, B-root, Potato, Mix Vegetables, Carrot, Green Onion, Methi, Palak, Saag, Sarsoon
Fruits	Apple, Banana, Grape, Guava, Malta, Orange
Dairy	Kheer, Milk, Youghart
Cereal	Chana, Chana Daal , Masoor, Legumes, Moong, Arhar
Fats/Oil	Butter, Ghee, Oil, Paneer

Table 4 Food Groups and their Food Items

.

3.2 Preprocessing the data provided

Raw data can be incomplete, noisy or inconsistent. Before applying any kind of data mining technique it needs to be cleaned so that any irregularities in the data are removed. The final cleaned data set is used as the input for the algorithm.

There were no missing attributes in the data however there were invalid entries present. This was done by creating a program which carefully checks each record and its respective values. For instance since we were dealing with food intake and kind of meat should not be present in the vegetable food group. Legumes which are part of cereal food group were present in Meat food group. These were carefully removed by mapping all parameters to their respective food group i.e. in cereals. The data provided for this research was hence processed and all kind of anomalies were eliminated by carefully going through all parameters row by row. This process was repeated multiple times to get a thoroughly cleaned data.

3.3 Modifying the data by reduction and transformation

After cleaning the data from all noise and inconsistencies we obtained cleaned data which can be further looked into for identify important parameters. All the unnecessary values need to be removed from the working data set to avoid any extra processing.

All parameters were examined carefully keeping in view the problem statement. This is an important step since any unnecessary attribute should not become part of the data and any vital attribute should not be removed because it might cause faulty results. Since we focus on diet only hence only the food related parameters were considered in the filtering and remaining were removed after being marked as redundant data.

The data transformation is merely a normalizing step in which the data is prepared to fit in the algorithm. The technique focused in this research works on categorical attributes. Hence we need to convert the textual data present to a categorical data type such as binary data. So these attributes are transformed to the binary data type required in the process.

The transformation was done in two stages. In the first stage a matrix was created against all the possible food items present in all the food groups. For instance the food group parameter: meat had mutton, beef, fish and chicken (shown in Table 4). So this one parameter was transformed to four sub parameters for each food item in the food group. In the second phase the matrix was populated with the values of 1 and 0 for each food item present against the patient number. So if the patient has taken in chicken only the matrix contained a value of 1 against chicken and 0 for rest of the three food items. This step was repeated for all the food groups and food items till we had the complete binary data converted. This process transformed our textual data into binary data.

3.4 Sample the data

The data is separated into training and test groups. The training (90% of the total data, Alive and Perinatal Mortality separately) and test group (10% of the total data, alive and Perinatal Mortality separately) is randomly selected with for each iteration.

The whole algorithm was applied to the training set and test set was untouched. In the last step of our process we compare our results with the test data outcomes.

3.5 ROCK

ROCK for Inherent Latent Ordinality can be applied on categorical data and the ordinality of the data can be discovered. The data available for this study consists of medical records of patients with singleton pregnancies. All food related information was extracted from the medical data set which is shown in Table 4. There were six food groups with thirty seven categorical food items with 1039 patients.

The algorithm takes input I i.e. a random sample data chosen from the total data set also known as the training data. The Alive and Perinatal Mortality outcome data is separated in to 2 subsets where C_A represents positive outcomes i.e. Alive and C_B represents negative outcomes i.e. Perinatal Mortality. ROCK is applied separately to both the groups to identify the clusters in the data.

$$sim(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

Figure 1 Jaccard Similarity

procedure cluster(S, k)
begin
1. $link :=$ compute_links(S)
2. **for each** $s \in S$ **do**
3. $\quad q[s] :=$ build_local_heap$(link, s)$
4. $Q :=$ build_global_heap(S, q)
5. **while** size$(Q) > k$ **do** {
6. $\quad u :=$ extract_max(Q)
7. $\quad v :=$ max$(q[u])$
8. $\quad$ delete(Q, v)
9. $\quad w :=$ merge(u, v)
10. $\quad$ **for each** $x \in q[u] \cup q[v]$ **do** {
11. $\quad\quad link[x, w] := link[x, u] + link[x, v]$
12. $\quad\quad$ delete$(q[x], u)$; delete$(q[x], v)$
13. $\quad\quad$ insert$(q[x], w, g(x, w))$; insert$(q[w], x, g(x, w))$
14. $\quad\quad$ update$(Q, x, q[x])$
15. $\quad$ }
16. $\quad$ insert$(Q, w, q[w])$
17. $\quad$ deallocate$(q[u])$; deallocate$(q[v])$
18. }
end

Figure 2 ROCK Algorithm [1]

ROCK computes the links between all the pair of points in each group C_A and C_B and creates clusters based on the interconnectivity of points. Any pair of point is defined as a neighbor if they satisfy the similarity threshold .Links are created between neighbors who are identified using a similarity function. In our research, we have used Jaccard coefficient as the similarity function. This generates uniform sized clusters for categorical data which become the first level clusters.

ROCK clusters the data according to the connectivity and ignores the overall importance of the item in the data set i.e. ordinality. As shown in Table 5, some of the food items occur more frequently in the clusters compared to others. For instance, Beef and Mutton represent a match more frequently in C_B but not in C_A. There are other food items (for e.g. oil) also present in C_A and C_B but they do not play a significant role i.e. are not responsible for creation of any links. Chicken is present in C_A with a probability of 0.49 but missing in C_B. Our approach assigns higher weightage to items which impact the clustering results more. We will see in the results section that assigning more weightage to more critical items improves the results greatly. These weights help in resolving the hidden ordinality in the data which ROCK fails to deal with. This gives us the top food items occurring abundantly in the clusters obtained for both groups C_A and C_B.

3.7 ROCKLIO

ROCKLIO is then applied to the data set using the strong links of the previous stage for determining a better link computation measure. The links from ROCK show the connectivity of food items within the Alive and Perinatal Mortality group. Food items which are found more abundantly in the group are more responsible for creating links. The weight values for various items are computed for both i.e. Perinatal and Alive group separately.

The two set of selected food items from ROCK are analyzed to estimate the significance factor of Alive and Perinatal outcomes. Significance factor is calculated by comparing the food item and its links in both the clustered groups i.e. Alive and Perinatal Mortality outcome. Higher significance factor shows higher effect of the food item in a specific cluster and low significance factor shows lower impact on the outcome. The significance factor is derived from the probability of a food item to help form a link between tuples or patient records.

Procedure ROCK-LIO-Comp-Weights (I; t)
Begin
C^A = TrainingData (I, 'Alive');
C^B = TrainingData (I, 'PerinatalMortality');
LinksA = ROCK (C^A, t);
LinksB = ROCK (C^B, t);
For $Item_i$=$Item_1$ **to** $Item_N$ **in I**
S_i^A = ItemsProbability ($Links^A$, $Item_i$)
S_i^B = ItemsProbability ($Links^B$, $Item_i$)
End For

W_i^A=Weights_compute(S_i^A);
W_i^B=Weights_compute(S_i^B);
End

Figure 3 ROCKLIO Algorithm

Procedure ROCK-LIO-Comp-Weights (I; t)
Begin
C^A = TrainingData (I, 'Alive');
C^B = TrainingData (I, 'PerinatalMortality');
LinksA = ROCK (C^A, t);
LinksB = ROCK (C^B, t);
For $Item_i$=$Item_1$ **to** $Item_N$ **in I**
S_i^A = ItemsProbability ($Links^A$, $Item_i$)
S_i^B = ItemsProbability ($Links^B$, $Item_i$)
End For

W_i^A=Weights_compute(S_i^A);
W_i^B=Weights_compute(S_i^B);
End

Figure 4 ROCKLIO Weights Computation

Weights are applied on the basis of the significance factor calculated which is in turn dependent upon the probability values shown in Table 5. The higher the probability value the higher is the significance of the food item. Based upon the significance factor, weight of a certain food item is computed. For each food group the weights are computed separately for the positive and negative perinatal outcomes.

Three different sets of weights were tried on the links. The range selection varied in each configuration, weights which produce the best outcome were chosen by applying gradient descent algorithm. The gradient descent algorithm estimated the weights in every iteration, and minimizes the error in every step of the way.

Food	**Alive**	**Perinatal Mortality**
Beef	0.00	0.26
Chicken	0.49	0.33
Fish	0.47	0.07
Mutton	0.03	0.35
Butter	0.33	0.37
Ghee	0.38	0.53
Oil	0.06	0.05
Paneer	0.24	0.03

Table 5 Probability Value

‘p’ represents the probability of an item to factor in the formation of a link. The weights configurations shown are divided into three different set of ranges which cover the significance factor ranges. Using the different weight configurations, the algorithm was applied on the test group separated initially and the results were compared.

The food items which are have a higher probability value in the alive data set have a lesser value in Perinatal Mortality, since these food items are more visible in either one of the two clusters. This relation is shown in equation (1) where ω represents the weights and α and β show the Alive and Perinatal Mortality food items. Hence using the equation we can easily categorize the food item as vital for pregnancy or a risk based on its weight in the equation (1). Using the significance factor calculated, and the probability value, the weights are assigned to the result according the formulae (2) and (3) shown below.

	Probability Ranges	Significance Factor (Weights)
Weight Configuration 1	0 > p >.05	1
	.05 > p > .10	2
	.10 > p > .15	3
	p > .15	4
Weight Configuration 2	0 > p > .10	1
	.10 > p > .20	2
	.20 > p > .30	3
	p > .30	4
Weight Configuration 3	0 > p > .15	1
	.15 > p > .30	2
	.30 > p > .40	3
	p > .40	4

Table 6 Weights Configuration

$$(\omega_1\alpha_1 + \omega_2\alpha_{2+\ldots+}\omega_N\alpha_N) = 1 - (\omega_1\beta_1 + \omega_2\beta_{2+\ldots..+}\omega_N\beta_N)..(1)$$

$$\chi = \frac{(\omega_1\alpha_1 + \omega_2\alpha_{2+\ldots..+}\omega_N\alpha_N)}{\Sigma^{\iota=0}\,\omega_\iota} .....(2)$$

$$\psi = \frac{(\omega_1\beta_1 + \omega_2\beta_{2+\ldots..+}\omega_N\beta_N)}{\Sigma^{\iota=0}\,\omega_\iota} .....(3)$$

Where α=food items of C_A, β=food items of C_B, ω = weights on the food items and this is divided by total number of items in a food group. The functions are used to calculate the new link measure based upon the weights and the occurrence or the use of a food item. This is then normalized by dividing the sum of the weights of the different items in the food group. The calculation is performed for each food group. They are applied separately for the Alive group's clusters and Perinatal Mortality group's clusters.

3.8 Accessing the results of ROCKLIO

The results obtained were then evaluated by tests to calculate their usefulness and reliability. The output was compared with the actual outcomes in the test data using W. M. Rand's (1971) Adjusted Rand Index [24]. Adjusted Rand Index measures the similarity between any two data clusters and gives a value between 0 and 1. It takes the two clusters and compares the items which are common in both clusters, producing 0 with no similarity and 1 for exactly the same clusters.

Testing the result, adjusted rand index [24] is applied to calculate the similarities between the two clusters of iterations. The iterations were repeated ten times with different test data used at each instance.

Chapter 4

Results and Discussion

We applied our technique on a real life data belonging to patients from a small town Gadap (population: 448,490) in Karachi, Pakistan. The study consisted of 3 councils from a total of 8. This population belonged to low socioeconomic group who had monthly income less than 3000/= per month (US $ 0.01 per rupee). The data consists of a variety of patients from the district which belonged to different sects of the society. Patients with singleton pregnancies and without any history of diabetes mellitus or hypertension were made eligible for this analysis. The main emphasis of the data lies on the food intake taken in by all the 1039 patients. There are six food groups and 34 food items included in the groups (shown in Table 4). The data is separated into target and control groups. The target (90% of the total data, alive and dead separately) and control group (10% of the total data, alive and dead separately) is randomly selected with replacement for each step of the process.

As discussed in methodology, we extracted the relevant data from the whole data set and converted it to binary data so that we can easily apply ROCK and then ROCKLIO.

4.1 ROCK Results

As a result of the ROCK, with the similarity threshold 0.50, we easily identified the major food items which affect the perinatal outcome. Two of the major clusters found in iteration 1 and their variations in Cluster A and Cluster B are shown in Figure 1. We can clearly see from the table that Beef (26%) and Mutton (31%) are present with a greater percentage in negative perinatal outcomes. These clusters show the links which satisfy the similarity threshold of 0.5.

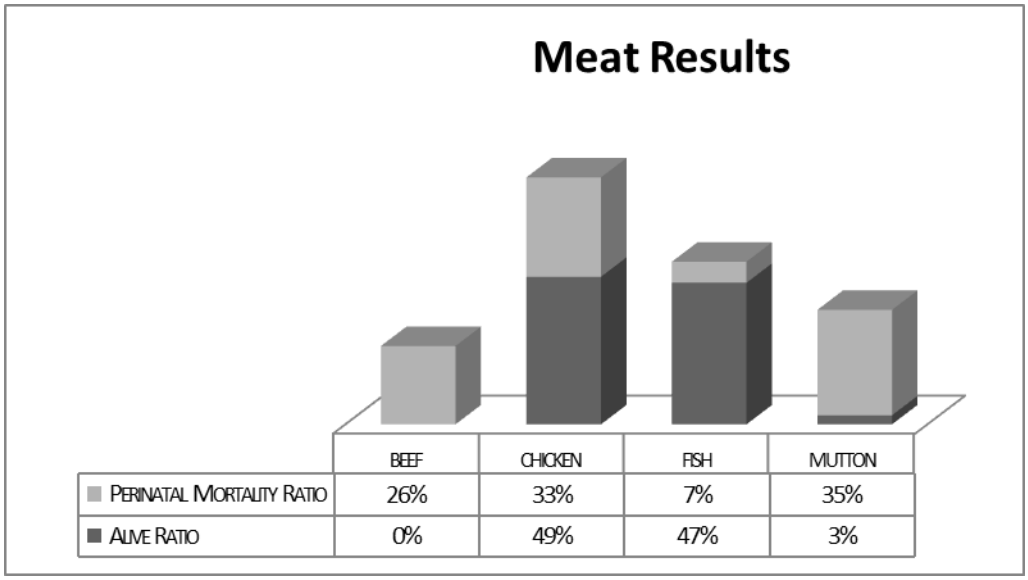

Figure 5 ROCK Meat Clusters

Another clear cluster is seen in Fats and Oil food group (Figure 2) which shows the presence of Paneer in successful pregnancies. According to these details, food items were marked good for pregnancies and some of them were marked as a risk.

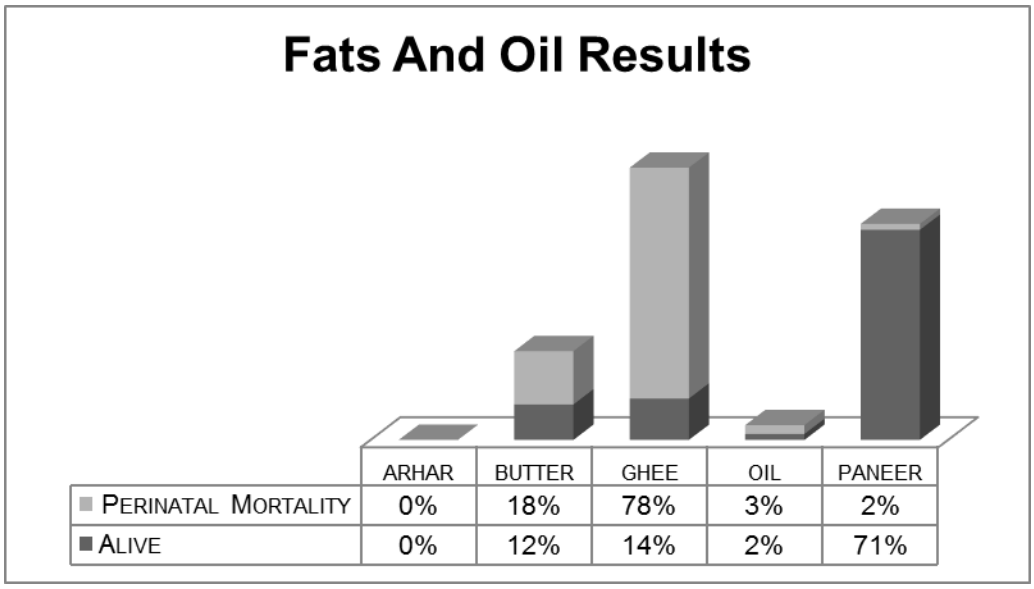

Figure 6 ROCK Fats and Oils Clusters

Figure 3 shows the result of clusters from the vegetable food group. There is no relation found in vegetables except Beet root food item which clearly tells expectant mothers having beet root are most likely to be at risk.

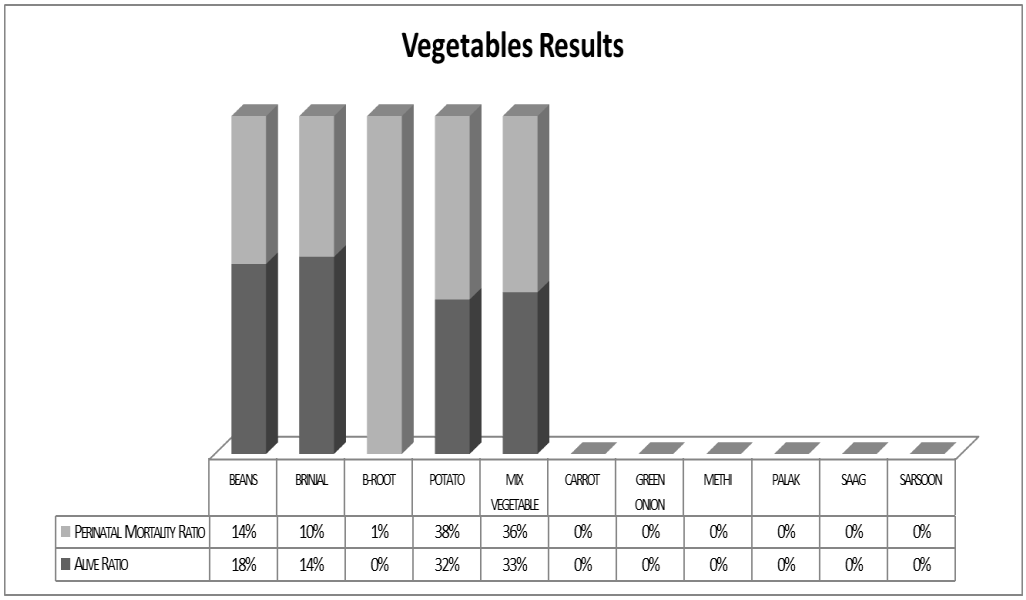

Figure 7 ROCK Vegetable Results

In fruits all food items were equally found in both the alive and perinatal groups shown in Figure 4.

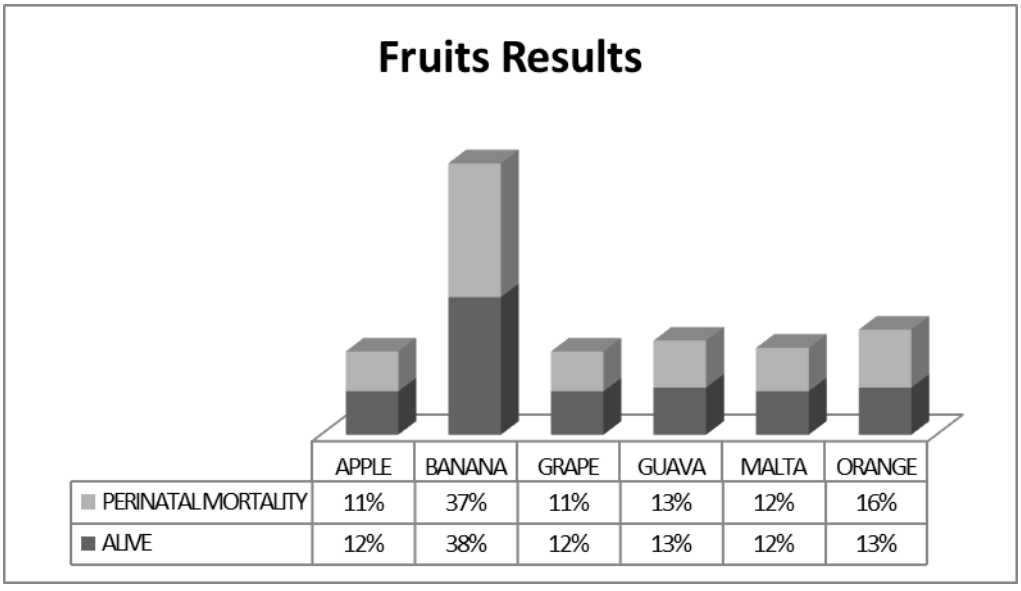

Figure 8 ROCK Fruit Results

Milk, Kheer and Yogurt were all present in almost equal amounts in both groups of perinatal mortality and alive (Figure 5).

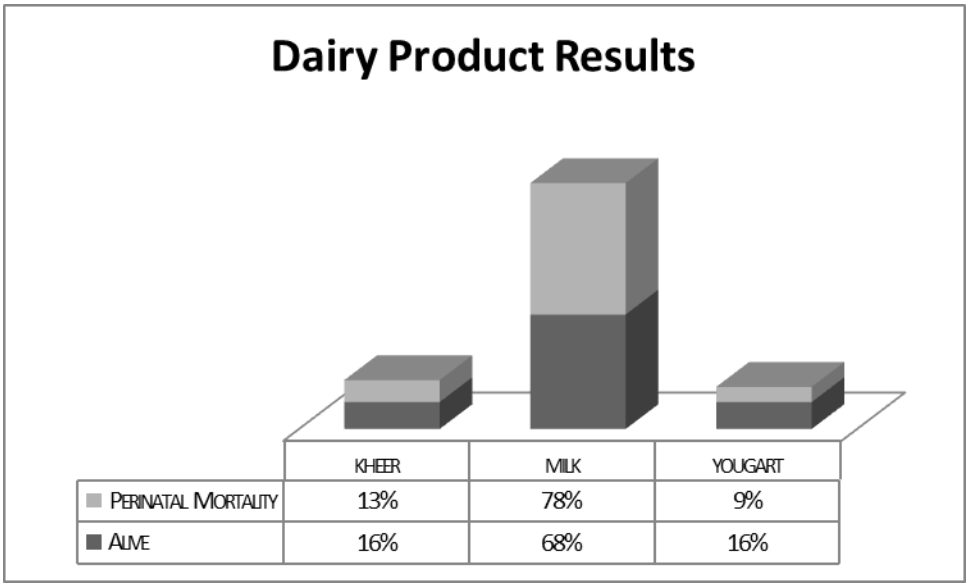

Figure 9 ROCK Dairy Products Results

Lastly in Cereals food group, we see one distinct food item Chana daal which is present in perinatal mortality cluster and not in alive group. Moong was found abundantly in alive cluster hence termed as a good food item for healthy pregnancy.

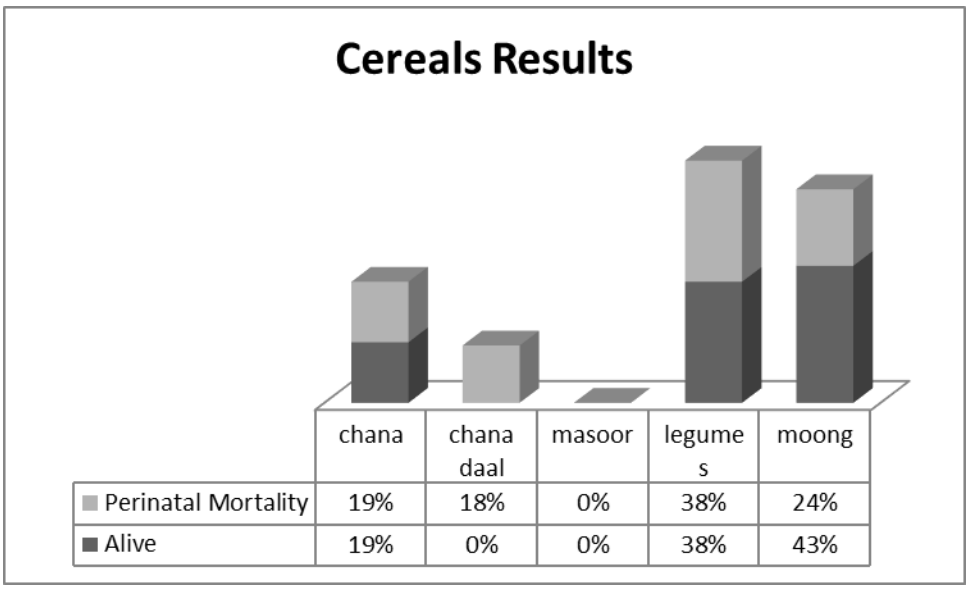

Figure 10 ROCK Cereals Results

4.2 ROCKLIO Results

Using the information obtained from ROCK, weights were applied separately according to Table 6. According to the result of the first iterations Chicken, Fish, Moong and Paneer were given the highest ranks for alive cluster, and Beef, Mutton, Chana Daal and Ghee was given the highest ranks for perinatal mortality. Using the

weights we calculated the “f value” for each patient and respective food group. The result was then applied to the same ROCK algorithm in the first iteration.

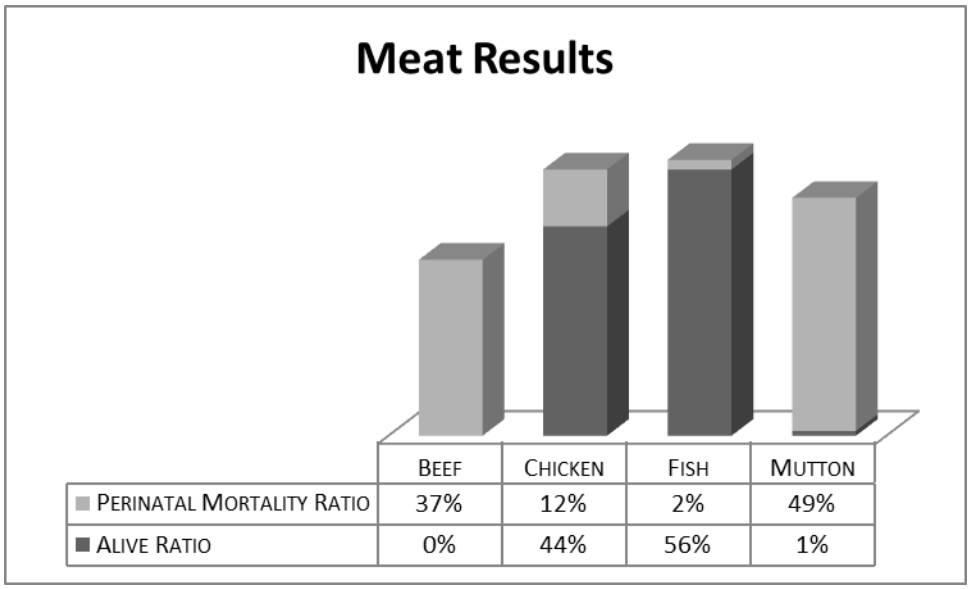

Figure 11 ROCKLIO Meat Results

Comparing Figure 7 with Figure 1, we can see a marked difference in the clusters formed. The weights given in ROCKLIO gave ordinality to the food items. The result shows that beef and mutton are harmful whereas chicken and fish are beneficial for pregnant mothers.

Same results can be seen in all the food groups as shown in the Figures below.

The inherent ordinality or significance of categorical data is successfully catered to by application of weights to these attributes as seen in the results table.

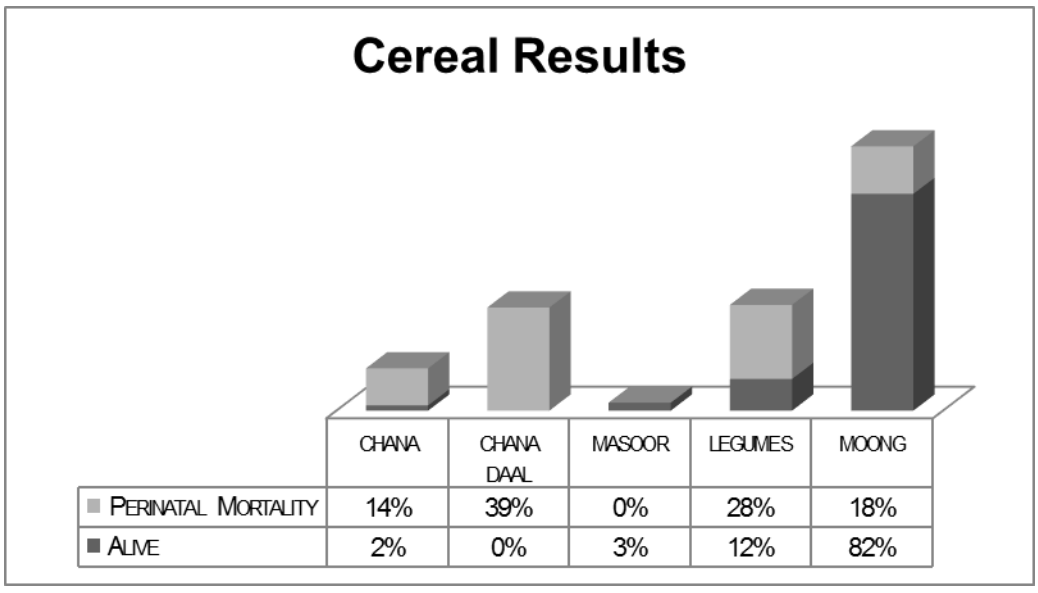

Figure 12 ROCKLIO Cereal Results

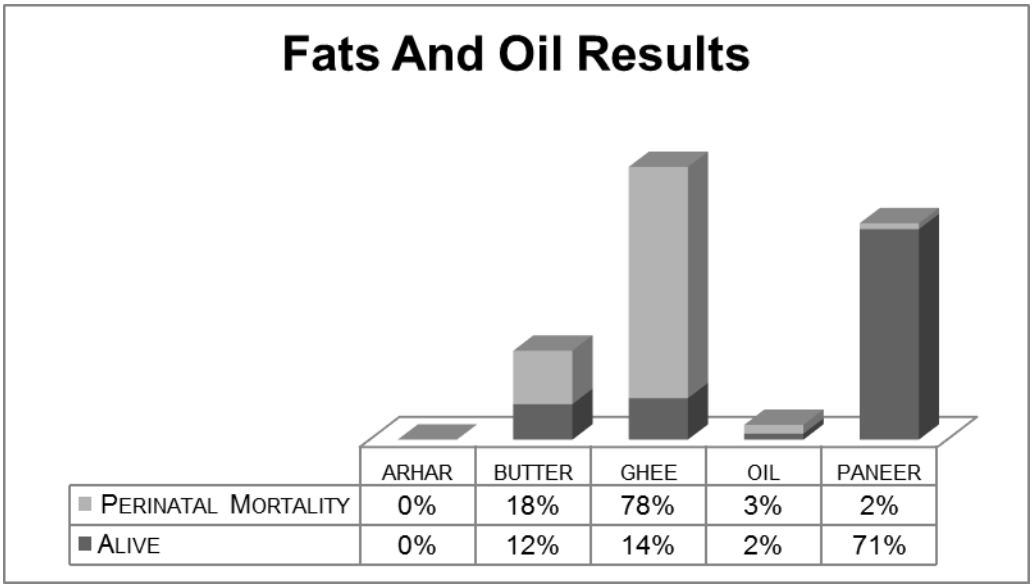

Figure 13 ROCKLIO Fats and Oils Results

These results identified food items as either good for pregnancy or which cause perinatal mortality. Food items like chicken, fish, paneer, banana, moong were highlighted as good for pregnant mothers in all of the iterations performed. Similarly items like mutton, beef were considered bad for pregnancy and likely to cause perinatal mortality.

The new output was tested against the actual outcomes using W. M. Rand's (1971) Adjusted Rand Index [24]. Adjusted Rand Index measures the similarity between any two data cluster and gives a value from 0 to 1. According to the Adjusted Rand Index, the value for the non-weighted analysis came out as 0.5041, and after applying ROCKLIO on the data, the clusters had an average of 0.7381. The iterations were repeated for different training sets ten times and each time the output was compared with the actual outcome of pregnancy, results are shown in Table 10 below. As we moved towards the tenth iterations,

Food	Alive	Dead	Variation
beef	0%	26%	-26%
chicken	49%	33%	17%
fish	47%	7%	41%
mutton	3%	35%	-31%

Table 7 Meat Results

Food	Alive	Dead	Variation
butter	33%	37%	-5%
ghee	38%	53%	-15%
oil	6%	5%	0%
paneer	24%	3%	20%

Table 8 Diary Products Results

Iteration#	Rand Index
Without ROCKLIO	0.5214
Iteration 1	0.659
Iteration 2	0.681
Iteration 3	0.741
Iteration 4	0.703
Iteration 5	0.777
Iteration 6	0.787
Iteration 7	0.798
Iteration 8	0.674
Iteration 9	0.798
Iteration 10	0.763
Average Index	0.738

Table 9 Adjusted Rand Index Results

the weights became more mature and consistent. Hence we successfully assigned weights to our categorical attributes with an average rand index of 0.74 using the best weights range.

Confirming our result, medical communities have also marked protein, fatty acids, iron and multivitamins as an important factor in birth outcomes [25] [26]. Comparing these studies with the result of this study, chicken and fish which is a source of protein, paneer source of fatty acids and cereals which contain Iron have been considered positive for a neonate.

Chapter 5

Conclusion

5.1 Discussion

Our focus in the research was to solve the limitation of inherent ordinality present in algorithms dealing with categorical variables. Along with the clustering algorithm, we also look into the medical issue of perinatal mortality. Multiple researches have been conducted in the field of science to evaluate the causes of perinatal mortality. All these have been discussed along with the various techniques available to cluster categorical data. But none of the algorithm available cater the issue of inherent ordinality in the data.

Perinatal Mortality, also known as perinatal death, is death of a neonate within 6 days (early neonatal mortality) or from 7 – 27 days of birth (late neonatal mortality). Food consumed by an expectant mother is said to have an impact on the pregnancy outcome apart from other factors. For the past few years, perinatal mortality rate has been increasing in developing and under-developed parts of the world. Two-thirds of the world's perinatal deaths occur in only 10 countries, and Pakistan is ranked third amongst these countries. These deaths have not been studied widely, in fact they have been under-reported and these reports have not even been considered in any attempts made to improve birth outcomes in developing nations. Nutritional, socioeconomic, demographic and health advice seeking behavior factors are responsible for a higher mortality rates in countries such as Pakistan

5.2 Our Contribution

Many studies have been carried out to study ordinal variables, but not much work has been done to find hidden ordinality in categorical data. We use ROCK and statistical analysis to identify any ordinality in categorical attributes. This leads to a new metric for computing links amongst the various data items. In the new metric, weights are assigned to attributes with greater impact. Hence, ROCKLIO produces better results.

When clustering categorical data such as ours it is difficult to estimate the inherent latent ordinality in the data which affects the accuracy of the technique applied. To resolve the issue in the context of our problem we propose the novel ROCK with Latent Inherent Ordinality (ROCKLIO). The technique involves two time clustering where in the first phase is used to ascertain what food groups have more impact on clustering and then a new link measure is derived that improves the clustering accuracy by applying weights on the basis of the significance of each food item. Better clustering results are obtained with the novel ROCKLIO as compared to simple ROCK. We have modified Robust Clustering using Links (ROCK) which is a hierarchical clustering technique used on categorical attributes. We have used ROCK since it considers interconnectivity of the clusters, and groups points having a larger number links/neighbors. ROCK links two points if they have similar items. We have added functionality to the existing ROCK algorithm by using weights and significant factors. Weights are computed on the basis of item’s significance in its respective cluster.

The results for the algorithms marked all the food items in Alive or perinatal mortality cluster. Food items like chicken, fish, paneer, banana, moong were present abundantly in Alive clusters hence are good for pregnancy in all of the iterations performed. Food items like mutton, beef were considered bad for pregnancy and are more likely to cause perinatal mortality.

We present ROCKLIO which clusters categorical attributes and calculate weights on the attributes accordingly by finding the latent inherent ordinality in the data. We have worked on effects of food intake on perinatal mortality and applied ROCKLIO, which identified the major food items effecting pregnancy

5.3 Future Direction

Our algorithm can be successfully used to cluster pregnant mothers as per the major food items in take and their relation to mortality may be ascertained. In a developing

country like Pakistan such research work is necessary to decrease the perinatal mortality and to increase the health awareness in the medical society.

This algorithm has been applied to a fixed number of attributes and items, it can be further tested on a higher number of attributes. The algorithm can also be used for other databases from various fields which contain categorical values.

Reference

[1] Guha, S., Rastogi, R., & Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. Information systems, 25(5), 345-366.

[2] http://www.who.int/bulletin/volumes/87/2/08-050963/en/index.html

[3] Geneva: World Health Organization, 2005, "The world health report 2005: make every mother and child count"

[4] Agarwal, P., Alam, M. A., & Biswas, R. (2011). Issues, Challenges and Tools of Clustering Algorithms. International Journal of Computer Science Issues (IJCSI), 8(3).

[5] http://www.pakservice.com/Health/health_home.html

[6] http://www.unicef.org/infobycountry/pakistan_pakistan_statistics.html#94

[7] http://en.wikipedia.org/wiki/Millennium_Development_Goals#Goal_4:_Reduce_child_mortality_rates

[8] J. Kristensen, M. Vestergaard, K. Wisborg, U. Kesmodel, and N. Secher,"Pre-pregnancy weight and the risk of stillbirth and neonatal death," BJOG An International Journal of Obstetrics and Gynaecology, vol.112, no. 4, pp. 403–408, 2005.

[9] D. Doherty, E. Magann, J. Francis, J. Morrison, and J. Newnham, “Prepregnancy body mass index and pregnancy outcomes” International Journal of Gynecology and Obstetrics, vol. 95, no. 3, pp. 242–247 2006.

[10] Hammad Qureshi, Rehan Hafiz, Mahjabeen Khan, Syed Mustafeel Aser Quadri , “Association of Pre-pregnancy Weight and Weight Gain with Perinatal Mortality”

[11] Jerzy W. Grzymala-Busse, Linda K. Woolery, “Improving Prediction of Preterm Birth, Using a New Classification Scheme and Rule Induction”

[12] Jerzy W. Grzymala-Busse, Linda K. Woolery, “A Comparison of Less Specific Versus More Specific Rules for Preterm Birth Prediction”

[13] Han, J., Kamber, M., & Pei, J. (2006). Data mining: concepts and techniques. Morgan kaufmann

[14] Barbará, D., Li, Y., & Couto, J. (2002, November). COOLCAT: an entropy-based algorithm for categorical clustering. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 582-589). ACM.

[15] Guha, S., Rastogi, R., & Shim, K. (1998, June). CURE: an efficient clustering algorithm for large databases. In ACM SIGMOD Record (Vol. 27, No. 2, pp. 73-84). ACM.

[16] Huang, Z., & Ng, M. K. (1999). A fuzzy k-modes algorithm for clustering categorical data. Fuzzy Systems, IEEE Transactions on, 7(4), 446-452

[17] Li Liu, Hope L Johnson, Simon Cousens, Jamie Perin, Susana Scott, Joy E Lawn, Igor Rudan, Harry Campbell,

Richard Cibulskis, Mengying Li,Colin Mathers, Robert E Black, for the Child Health Epidemiology Reference Group of WHO and UNICEF, “Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000”

[18] Imtiaz Jehan, Hillary Harris, Sohail Salat, Amna Zeb, Naushaba Mobeen, Omrana Pasha, Elizabeth M McClure, Janet Moore, Linda L Wright & Robert L Goldenberg, “Neonatal mortality, risk factors and causes: a prospective population-based cohort study in urban Pakistan”, WHO Bulletin Volume 87, Number 2, February 2009, 130-138

[19] Han, J., Kamber, M., & Pei, J. (2006). Data mining: concepts and techniques. Morgan kaufmann

[20] Barbará, D., Li, Y., & Couto, J. (2002, November). COOLCAT: an entropy-based algorithm for categorical clustering. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 582-589). ACM

[21] Guha, S., Rastogi, R., & Shim, K. (1998, June). CURE: an efficient clustering algorithm for large databases. In ACM SIGMOD Record (Vol. 27, No. 2, pp. 73-84). ACM

[22] Zhang, Tian, Raghu Ramakrishnan, and Miron Livny. "BIRCH: an efficient data clustering method for very large databases." ACM SIGMOD Record. Vol. 25. No. 2. ACM, 1996.

[23] Huang, Z., & Ng, M. K. (1999). A fuzzy k-modes algorithm for clustering categorical data. Fuzzy Systems, IEEE Transactions on, 7(4), 446-452

[24] W. M. Rand (1971). "Objective criteria for the evaluation of clustering methods". Journal of the American Statistical Association (American Statistical Association) 66 (336): 846–850

[25] Abu-Saad, Kathleen, and Drora Fraser. "Maternal nutrition and birth outcomes." Epidemiologic Reviews 32.1 (2010): 5-25

[26] Lechtig, Aaron, et al. "Effect of food supplementation during pregnancy on birthweight." Pediatrics 56.4 (1975): 508-520.