Finding Patterns in Diabetes Patients Using Functional Clustering

Dr. Satish Kumar Soni[1], Dr. Manish Kumar Soni[2]

[1]MIET Meerut
Email: *satishkumarsoni15[at]gmail.com*

[2]Govt. Indira Gandhi Engineering College, Sagar
Email: *profmanishksoni[at]gmail.com*

Abstract: ***In general diabetes is recognized as a lifestyle caused disease but timely prediction of primary signs of disease can prevent millions to gripe in. In this exploration we are trying to study the practically obtainable clustering procedures that may aid in prediction of pre-diabetic conditions and affecting features, additionally we are trying to present a comparative model using functional clustering and numerical methods with enhanced outcomes and recommendations.***

Keywords: Functional Clustering, Diabetes Dataset, Numerical Methods, Functional Fitting

1. Introduction

In today's world, data is everything. Data in itself is purely facts and figures. Data analysis is now a universal language and more important than ever before. Its importance is to understand problems facing different areas, and to explore data in meaningful ways, it organizes, interprets, structures and presents the data into useful information that provides perspective for the data. Medical science is producing waste amount of data for about all types of diseases, diabetes is one of the lifestyle driven disease creating panic for nearly all countries, From child to old, numerus age groups are affecting and this is a challenge for the researchers to predict the early signs of diabetes. In this study we are trying present a new clustering approach to find the patterns of pre-diabetic signs using Functional clustering methods and numerical techniques in diabetes patient's data. The idea of the clustering is to create similar clusters (grouping of structures) of observations demonstrating insights of some random variable X. clustering is a primary process to find the underlying structure with in objects to understand the similarity or dissimilarity between objects. When data can be represented as functions the functional clustering and numerical methods can be play a vital role [1].

2. Previous Work

In [2] authors mentioned, a functional arbitrary variable Y is a random variable with values in an unbounded dimensional space. Then, the formulated data can be represented as a set of observations {Y1………. Yn} of Y. the problem arises dealing with functional data is that it is supposed to come in the form of infinite dimensional space but practically the data can be analyzed in the form of finite functions and curves. Numerous methods have been proposed by different researchers. The authors in [3] and [4] encompasses to functional data, the canonical learning of two functional variables, principal component analysis, multiple correspondence analysis and linear regression on functional data. An important involvement to functional data is presented in [5]. The basic approach is to first fit the curves or the functions using approximation techniques when analyzing the functional data for this in [6] the authors used B- spline method combined with k-means, in [7] used robust trimmed k-means method, [8] describes the k-means with principle component analysis, the Byes method with non-parametric clustering and wavelet methods proposed in [9], in [10] authors proposed clustering after smoothing and transformation. The clustering with smoothing and transformation can be used to cluster large number of functional data. These all are some good contributions in functional clustering but not limited to. In this study we will go through different smoothing and transformation methods and then apply the numerous clustering techniques to present a model that can compare the outcomes for this the diabetes dataset is used.

3. Materials and Methods

3.1 Dataset

The dataset explored in this study is taken from UCI Irvine Machine Learning Repository [11]. The dataset contains blood glucose levels, HBA1C, insulin and other attributes of 70 patients. Our aim is to fit the data for trend line then cluster that data to find high and low blood glucose for the patients to raise an alarm to alert the doctors. The data file consists four attributes per record i.e. File Names and format: Date in MM-DD-YYYY format, Time in XX: YY format, Code, Value. A sample of raw data is given in the Table 1, all patients data is given in the separate text files where id is the patient's id ranging from 01-70 for each patient, The Code field contains categorical variables like hypoglycaemic symptoms, etc. The detailed description about the dataset can be seen in the link [11] given in the references section.

Table 1: Sample of Diabetes Data in Raw Format for Patient Id 01

Id	Date	Time	Code	Value
1	04-21-1991	09:09	58	100
1	04-21-1991	17:08	62	119
1	04-21-1991	22:51	48	123
1	04-22-1991	07:35	58	216
1	04-22-1991	16:56	62	211
1	04-23-1991	07:25	58	257
1	04-23-1991	17:25	62	129
1	04-24-1991	07:52	58	239
1	04-24-1991	17:10	62	129
1	04-24-1991	22:09	48	340
1	04-25-1991	07:29	58	67
1	04-25-1991	17:24	62	206
1	04-25-1991	21:54	48	288

3.2 Methods

Transformation: Dataset is in text format and in different files so we need to transform the data appropriate to weka software, which is used for clustering, for this we are using python functions to combine the 70 files. Code field is categorical value which is converted to numeric. Some missing values also exist in some attributes are removed, some mixed valued attributes present also removed or in some cases replaced.

Smoothing: Non-linear smoothing is applied before clustering to find nonlinear fitted function. Nonlinear Least Squares with Levenberg-Marquardt Function for Non-Linear Functional Fitting is applied. The Levenberg-Marquardt (L-M) algorithm is an iterative technique which associates the Gauss-Newton method and the steepest descent method. The algorithm works well for most cases and turn out to be the classic of nonlinear least square method.

1) Calculate the $a^2(x)$ value from the given initial values: x.
2) Pick a uncertain value for φ, say φ = 0.001
3) Solve the Levenberg-Marquardt function for δx and evaluate $a^2\ (\alpha + \delta x)$
4) If $a^2\ (\alpha + \delta x) \geq a^2\ (x)$, increase φ by a factor of 10 and go back to step 3
5) if $a^2\ (\alpha + \delta x) \leq a^2\ (x)$, decrease φ by a factor of 10, update the parameter values to be δx and go back to step 3
6) Stop until the χ^2 values computed in two successive iterations are small enough (compared with the acceptable values).

This will give the nonlinear fitted function.

The **K-means clustering** algorithm [12] is a distance based partitioning algorithm which minimizes distance between similar objects and maximizes distance between dissimilar objects. The objective functions of k-means is given by the following formula:

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\|x_i^{(j)} - c_j\|.^2$$

Where $\|x_i^{(j)} - c_j\|$

is a distance function between data points $x_i^{(j)}$ and centroid c_j

The distance function can warily consider type of datasets. Typically, Euclidian Distance is used.

The steps of algorithm is as follows:

1) Select number of k and randomly assign k points these are the initial cluster centres.
2) Assign each point to the closest k based on the distance calculated.
3) If assigned all points to some k, recalculate the k cluster centers.
4) Repeat steps 2 and step 3 until cluster centers don't move further.
5) Outcomes k group of similar point.

DBSCAN clustering algorithm [13] is Density-Based Spatial Clustering of Applications with Noise most commonly used density-based algorithm. It uses the concept of density reachability and density connectivity. **Density Reachability**: when an object "x" is reachable to object "y" having ε distance from "x" to "y" and sufficient neighbour in ε distance. **Density Connectivity:** this is a chaining process i.e. if "x" is neighbour of "r", "r" is neighbour of "s", "s" is neighbour of "y" then "x" is neighbour of "y".

The steps of algorithm is as follows:

1) Start with random data points not visited yet.
2) Calculate ε and find neighbourhood of data points based on ε.
3) If find adequate neighbourhood nearby this data point then clustering process starts and point is marked as visited else this point is labelled as noise.
4) If found point is participating in the cluster then its ε neighbourhood is also the part of the cluster and the above practice from step 2 is repeated for all ε neighbourhood points. The process is iterated until all points in the cluster is calculated.
5) Take a fresh point not taken and processed, for the detection of a further cluster or noise.
6) This process continues until all points marked as visited.

4. Experimental Results

Proposed Method:

In this study we have considered the clustering techniques and numerical methods to present a new approach for analysing the diabetes dataset. Experimental setup is created using weka for clustering and python for transformation, cleaning, functional fitting. We have used nonlinear functional fitting methods to model the diabetes dataset and after that used clustering methods to group the similar functions to find the high and low blood sugar of the patients. We have used Nonlinear Least Squares with Levenberg-Marquardt Function for Non-Linear Functional Fitting to model each patient's high and low blood sugar value separately of the diabetes dataset. The graph plot of Nonlinear Least Squares with Levenberg-Marquardt Function for non-linear fitting sample is given in the figure 1 for

Predicted High Blood Sugar Values and figure 2 for Predicted Low Blood Sugar Values

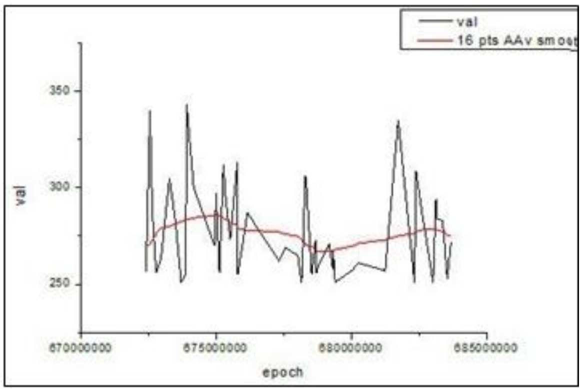

Figure 1: Predicted High Blood Sugar Values

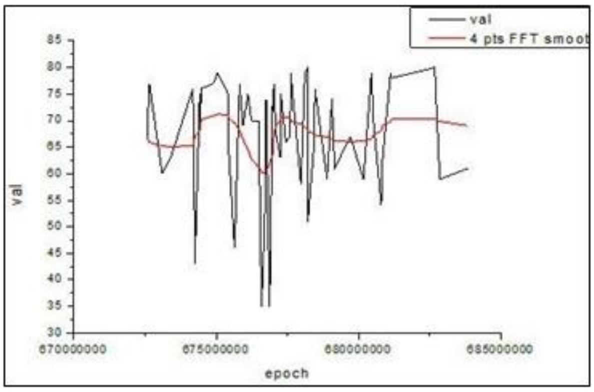

Figure 2: Predicted Low Blood Sugar Values

The sample of predicted values of high blood sugar is shown in the table 2 and predicted values of low blood sugar is shown in table 3.

Bellow the table depicts the results of raw dataset clustering after applying transformation for different clustering techniques

Table 4: Results of Raw Dataset Clustering After Applying Transformation for Different Clustering Techniques

Evaluation	Clustered Instances	Within cluster sum of squared errors	Time taken (Sec.)	Seeds	Number of iterations
Clustering Methods					
K-Means	0 1980 (48%) 1 2111 (52%)	239.88	0.23	10	4
DBSCAN	0 1980 (48%) 1 2111 (52%)	154.7	0.60	60	4

Bellow the table depicts the results of non-linear predicted dataset clustering after applying transformation and functional fitting on diabetes dataset for different clustering techniques

Table 5: Results of Non-Linear Predicted Dataset Clustering

Evaluation	Clustered Instances	Within cluster sum of squared errors	Time taken (Sec.)	Seeds	Number of iterations
Clustering Methods					
K-Means	0 2004 (49%) 1 2087 (51%)	180.052	0.09	10	3
DBSCAN	0 2115 (52%) 1 1976 (48%)	110.45	0.07	11	2

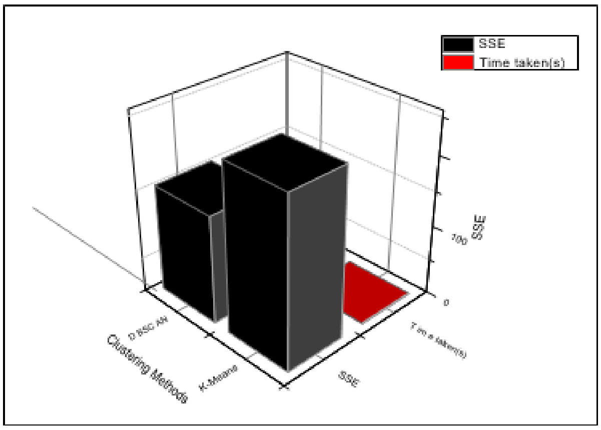

Figure 3: The performance graph for table 4

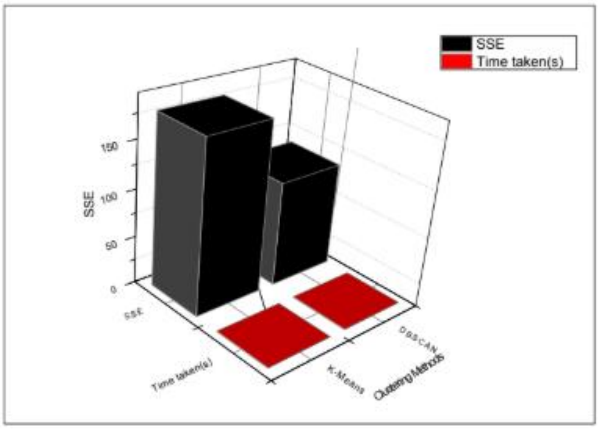

Figure 4: The performance graph for table 5

The following figures presenting the graphical aspects of the different clustering techniques before applying functional fitting clustering and after applying functional fitting clustering

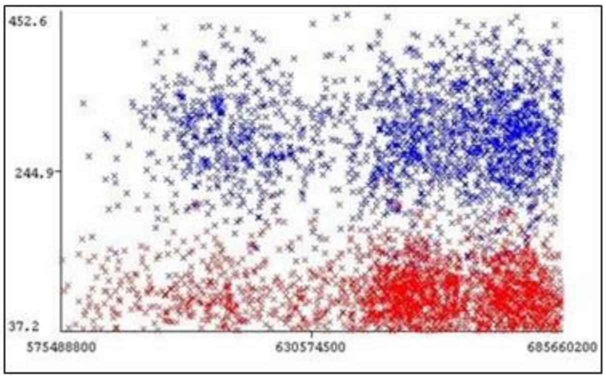

Figure 5: k-means clustering before applying Functional fitting(x->epoch, y ->value)

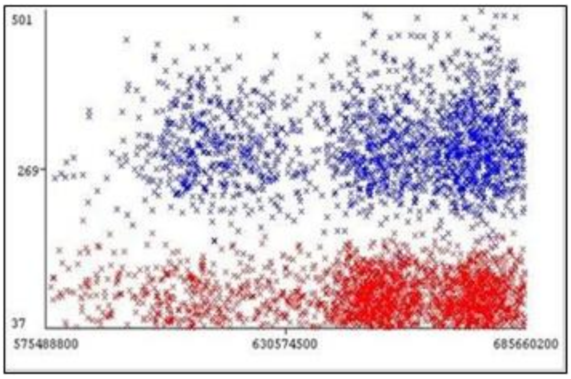

Figure 6: k-means clustering after applying functional fitting(x->epoch, y ->value)

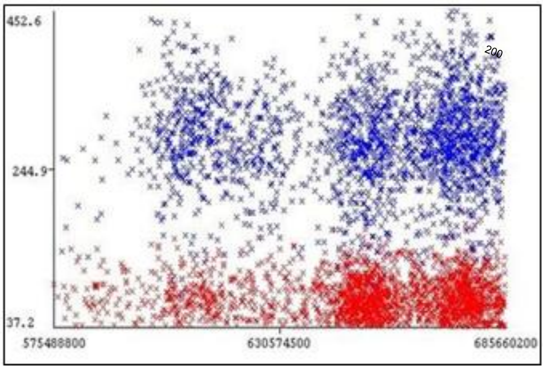

Figure 7: DBSCAN clustering before applying functional fitting(x->epoch, y ->value)

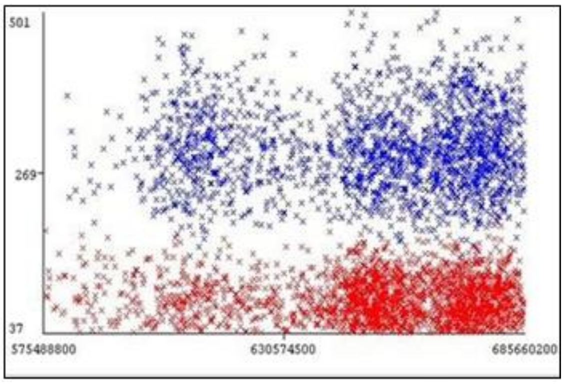

Figure 8: DBSCAN clustering before applying Functional fitting(x->epoch, y ->value)

5. Conclusion

The results given in the table 4 and table 5 shows the comparison between both clustering methods and it is clear that the results have improved after applying functional fitting clustering technique over raw data clustering techniques. The performance of both the clustering methods has comparable. The graphical representation of both the methods have also verifies the results shown in the table 4 and table 5. In coming days, we will try more practice with other numerical methods and functional clustering techniques for exploratory data analysis.

References

[1] C. P. Julien Jacques, "Functional data clustering: a survey," *Advances in Data Analysis and*, vol. 8, no. 3, p. 24, 2014.

[2] F. Ferraty and P. Vieu, "Nonparametric functional data analysis," *Springer Series in Statistics, New York*, 2006.

[3] P.Besse, "Descriptive study of a process," `eme cycle` Paul University, Sabatier, Toulouse, 1979.

[4] G. Saporta, "Exploratory methods of analyzing temporal data," Buro notebooks, 1981, pp. 37-38.

[5] R. Boumaza, "Contribution to the descriptive study of a qualitative random function, PhD thesis," Paul University, Sabatier, Toulouse, France, 1980.

[6] C. Abraham, P. A. Cornillon, E. Matzner-Løber, and N.Molinari, "Unsupervised curve clustering using B-splines," *Scand. J. Statist.*, vol. 30, p. 581–595, 2003.

[7] L.A. García-Escudero, and A. Gordaliza, "A proposal for robust curve clustering," J. Classificn., 2005.

[8] T. Tarpey, and K. K. J. Kinateder, "Clustering functional data," *Jurnal of Classification,* vol. 20, p. 93–114, 2003.

[9] S. Ray and B. Mallick, "Functional clustering by Bayesian wavelet methods," *J. R. Statist. Soc*, vol. 68, no. B, p. 305–332, 2006.

[10] N. Serban and L. Wasserman, "clustering after transformation and smoothing," *J. Am. Statist. Ass.*, vol. 100, p. 990–999, 2005.

[11] M. Kahn, "UCI Irvine Machine Learning Repository," Washington University, St. Louis, MO, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Diabetes. [Accessed may 2020].

[12] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1967.

[13] M. Ester, H.P. Kriegel, J. Sander, X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise". Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231, 1996.