Data Smoothing Numerical Methods and Their Applications in Unsupervised Learning for Prediction of Diabetes in Patients

Satish Kumar Soni, R S Thakur, A K Gupta

Abstract: ***Machine Learning in its fullest can provide much more accurate and enhanced analysis for medical diagnosis. In this paper we are trying to portray how the data related to diabetes can be used to predict if a person has diabetes or not. In more specific way this paper will explore the utilization of numerical methods smoothing with unsupervised learning to predict the early signs of disease like diabetes and rest.***

Index Terms: Unsupervised Machine Learning, Numerical Methods, Functional Smoothing, Diabetes Prediction, Clustering.

I. INTRODUCTION

Unsupervised learning techniques or clustering is an initial yet important approach of machine learning which describes the underlying structure or the existing patterns within the data [1]. There are a number of clustering techniques that can be applied to the numerous datasets of wider fields such as crop disease prediction in agriculture, transport management, medical diagnostics, IoT, cyber & network security, image segmentation, object detection in videos etc. Clustering is basically grouping of similar datasets in one group and dissimilar datasets in others by which we can identify dense and sparse regions, therefore discover interesting patterns and correlations among the datasets [2]. The overall performance of clustering techniques depends upon the quality of the clusters produced, so in this paper we will use the different clustering approaches to find the appropriate quality of clusters for diabetes dataset and compare among them [3]. The functional data analysis using curves and error estimation was first introduced by Ramsay (1982) in his paper "When the Data are Functions" [4], prior to that in recent decades, many researchers have employed datasets in the form of pieces of curves and functions basically in the biomedical studies. Now a day it is common to identify the patterns applying this type of approaches. With the development of high performance computing devices, the denser curves and functions gone ease which makes functional smoothing clustering more prevalent [5]. Important initial step for analysing functional data is data smoothing, researchers may also directly use raw curves for inferences instead of smoothing before further analysis.

Revised Manuscript Received on July 05, 2019.

Satish Kumar Soni, PhD (pursuing) in Computer Sciences, Barkatullah University, Bhopal. MCA from RGPV University, Bhopal

Dr. Ramjeevan Singh Thakur, Associate Professor(MCA), MANIT, Bhopal,

Dr. Anil Kumar Gupta, Dept., of Computer Science & Applications Barkatullah University, Bhopal,

However, in Hitchcock, Casella and Booth (2006) and Hitchcock, Booth and Casella (2007) the drawn study clarifies that most of the times the smoothing dramatically improves the accuracy and performance of the further pattern analysis such as clustering and classification [6][7]. Recent data smoothing methods popular among researchers include curve fitting, regression splines, roughness penalty-related smoothers, kernel-based smoothers, and smoothing splines. Past study revealed that there is no unified system that can be applied repeatedly to a huge collection of datasets for perfect data fitting (Ramsay and Silverman 2005) [8]. In this paper we are applying unsupervised learning techniques in the diabetes dataset to accurately predict the no of patients as clusters who may be infected by the diabetes disease. For these different clustering techniques such as k-means, EM, DBSCAN, X-means, Farthest First are used integrating smoothing techniques for better accuracies.

II. DATA PREPARATION

Loading Dataset: In this paper we are using an existing and open access dataset of "Pima Indians Diabetes Database" provided by the UCI Machine Learning, a huge open access database repository available online for machine learning. The dataset contains records of female patients at least 21 years old of Pima Indian heritage, 768 records with 8 attributes and no missing values [9].

Normalizing the Dataset: There is a saying in machine learning "Better data beats fancier algorithms", which suggests better data gives you better resulting models [10]. The first step is to explore data for the available attributes and features. Then second step is data cleaning in this step many factors considering cleaning diabetes dataset: removal of duplicate and irrelevant instances, multiple attributes with similar categories, missing or data points with 0 values, detection of unwanted outliers. In diabetes dataset we are only observing the cleaning process for null or 0 values and potential outliers because this is a standard dataset. By graphical observation we have seen that no missing values in this dataset but some of the outliers are seen in some attributes which can be dramatically harmful for the analysis. Such as in Blood Pressure attribute some values are 0 but it is known that a living person cannot have 0 diastolic Blood Pressure, 35 such entries found. Same way in Plasma Glucose Levels attribute the 5 entries found 0 which is not possible even in fasting Blood Sugar count and in BMI attribute 11 entries found 0 which is again not a valuable

measurement. We abandon the remaining features since these three plays vital role for diagnosis of diabetes. After a through data cleaning we come to the point that this dataset is incomplete and we should remove the instances with 0 values in column "Blood Pressure", "BMI" and "Glucose" [10].

III. ALGORITHM SETUP

Clustering Methods:

K-means: k-means clustering is one of the classical clustering techniques applied in past decades. This is a distance based clustering in which the number of clusters should be specified in advance, these are the centroids (mean vectors) from which the distance of all points calculated and try to minimise the sum of squared distances between data points and respective centroids. The number of k chosen randomly, initially first few points considered as k [2].

Algorithm: - k-means

Let the data points are denoted by the set of X= {x1, x2, x3, x4…….,xn} and the set of clusters by K={k1,k2,k3,k4,……..,kn}

Step 1: Initialize random cluster centers i.e. k (fig.1)

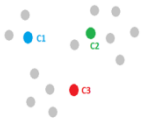

Figure 1

Figure 2

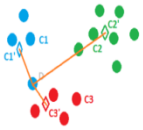

Figure 3

Figure 4

Figure 5

Step 2: Calculate the distance of each point using given formula and Assign them to the nearest cluster center (fig.2):

$$d = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (1)$$

Step 3: Recalculate the cluster centers (fig.3) as mean of assigned observations using this formula:

$$Vi = (1/ki)\sum_{j=1}^{ki} xi \quad (2)$$

Step 4: Repeat step 2 and step 3 until convergence

Step 5: Final Cluster

DBSCAN: Density-Based Spatial Clustering of Applications with Noise introduced in Ester et al. 1996 [11]. Finds core points of high density and draws clusters from them. Good for data which contain clusters of similar density. The main features of this algorithm are [12]:

- Unlike the k-means DBSCAN does not require to specify the number of clusters in advance to be generated.
- It can find any shape of clusters. The cluster doesn't have to be spherical.
- It can identify outliers too.

Algorithm: - DBSCAN

Let $X=\{x_1, x_2, x_{3...}x_n\}$ is the set of data points. DBSCAN requires two parameters: ε (epsilon) and the minimum number of points required to form a cluster (minPts).

Step1 :DBSCAN creates shapes of different dimensions around each point in the dataset and then calculates how many points belong to that shape using ε.

Step2 : It identifies these shapes as clusters.

Step3 : Then it iteratively expands the cluster assignments, by going through each instance point within the cluster, and verifying the number of other data points in close proximity.

Step4 : Repeat this process until the new assignment exists.

Step5 : Final Cluster

EM: EM assigns a probability function to each data points which specify the probability of it belonging to each of the clusters. EM can self-check that how many clusters should be created by cross validation or one can specify a priory how many clusters to generate.

Algorithm: - EM

Step1: Select number of k randomly or may choose any method for that. These are the primary cluster means.

Step2: Calculate variance of each cluster.

Step3: **E Step:** Find the subsequent probabilities of the hidden variable given present parameter values.

Step4:**M Step:** Re-estimate the attribute values given the present subsequent probabilities.

Step5: Using cluster mean & variance create multivariate Gaussian distribution for each cluster

Step6: Calculate the probability of each data point grouped into each cluster using the Gaussian distribution, each observation belongs to each cluster with a certain probability.

Step7: Repeat 2-6 and monitor the likelihood function until convergence is noticed.

Farthest First: Farthest first algorithm first introduced by Hochbaum and Shmoysat(1985) and has same course of action as k-Means. This applies the concept of centroid and the assignment of objects in cluster having maximum distance. The starting points are values which are far more from mean of values. Here clusters assignment is different and at initial cluster, get link with high Session Count, like at cluster-0 more than in cluster-1 and so on. This arrangement places clusters farthest from points. This point should be within dense area of dataset. Points farthest from center clustered first. Where less iterations needed this method can give better performance.

Algorithm: - Farthest First

Step1 : First choose any dataset randomly as center.

Step2 : Find the dataset that is farthest from first point.

Step3 : Find the third point which is farthest from two existing points.

Step4 : Then find the point that has not yet selected.

It is the farthest point from {1, 2… j-1} and mark it as point j. Use d(x,S) = minyϵS d(x,y) to identify the distance.

Optimisation Methods:

Exponential curve fitting: -

For optimising the results of these clustering techniques we are using numerical methods for functional approximation and curve estimation of given dataset. Considering the dataset and correlations between the parameters and the variations depicted in the measurements we are looking the functional curve of the form $y=Ae^{kx}$ in data. For all our data points (xj,yj), we are computing wj=ln(yj) and the constants a and b such that the line w=a+bx is a line of best fit to the data (xj,wj).Then e^a and b are good estimates for A and k respectively. Meanwhile the selection of "Line of best fit" is a huge issue [13]. Here we will use **least squares** to find the significant numbers for a and b which minimizes

$$\sum_{j=1}^{n}.\left(wj-(a+bxj)\right)^2 \qquad (3)$$

It fits out that the best b and a are given by the following formulas:

$$b=\frac{\sum_{j=1}^{n}xjwj-\frac{1(\sum_{j=1}^{n}xj)(\sum_{j=1}^{n}wj)}{n}}{\sum_{j=1}^{n}xj^2-1/n(\sum_{j=1}^{n}xi).^2} \qquad (4)$$

and

$$a=\frac{1}{n}\sum_{j=1}^{n}wj-\frac{1}{n}b\sum_{j=1}^{n}xj \qquad (5)$$

where b is given by the previous formula [13].

This gives the functions with well fitted smoothed data points which on clustering gives better performance and further performance improvements for prediction of diabetes.

Weighted Moving Average curve fitting: -

The Weighted Moving Average method figures the average of a set of input values over a specified interval on the curve. In this function, the weights are given more to recent data points. This function can be used as a piece wise curve fitting, that may help to reduce noise and make it easier to predict data trends [14].

The formula for the weighted Moving Average Curve Fitting is being considered is as follows:

$$W_t = (n * Y_t / k) + ((n-1) * Y_{t-1} / k) \qquad (6)$$

here Y is the data point, n is the number of intervals, and k is the sum of n-based multipliers [14].

These two numerical methods for curve fitting and estimation of best possible predicted data points are used in this study to give better clusters with improved quality and performance.

IV. APPLYING ALGORITHMS ON THE NORMALIZED DATASET

There are two parts of this study First we apply the clustering methods on the data taken from the UCI web repository and compare the results of these methods, Second we apply numerical techniques based functional approximation and curve fitting on the data, then apply the clustering techniques and finally we compare the results of clustering without optimization and clustering with optimization which will give the approach that can give better results for diabetes prediction from medical datasets. The analysis is done in Weka 3.8 and python anaconda navigator tools some of the data processing is done in Microsoft Excel 2016.

Applying Clustering Without Optimisation: -

After applying the techniques corresponding results are given in Table.1 with different parameters which shows the respective performance of the clustering techniques. Notably the K-means gave comparative performance in diabetes datasets without optimization in methods.

Table.1Performance of clustering methods without optimization

Method	Iterations	Time (s)	SSDE	Clusters Identified
K-means	9	0.30	33.06	2
DBSCAN	10	0.14	45.02	1
EM	58	2.46	35.23	2
Farthest First	5	0.01	38.01	2

Applying Clustering with Optimisation: -

The proposed optimization done here and after applying the optimisation the analysis is carried out with same dataset and the tools setup, results are shown in Table.2.

Table.2 Performance of clustering methods after applying optimization

Method	Iterations	Time (s)	SSDE	Clusters Identified
K-means	9	0.01	38.33	2
DBSCAN	10	0.09	25.02	2
EM	100	6.06	30.23	2
Farthest First	8	0.01	20.01	2

After applying the numerical methods based optimization the performance of some methods have improved which are comparative to the without optimization results for given dataset.

V. GRAPHICAL REPRESENTATION OF RESULTS

The graphical representation of the result of the clustering is shown in the below given figures. Fig.1 shows the K-means clustering result without applying optimization

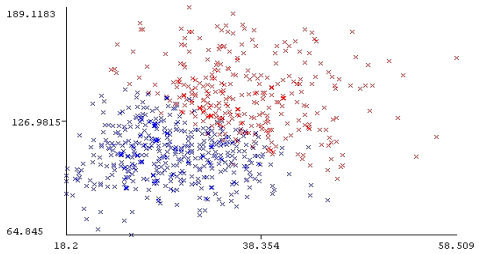

Figure 1 K-means created clusters

Fig.2 shows the DBSCAN clustering result without applying optimization we can see that only one cluster is generated as DBSCAN automatically identifies the number of clusters.

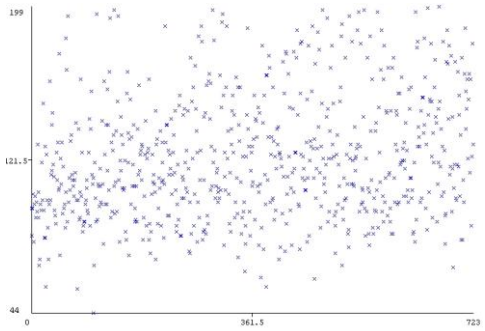

Figure 2 DBSCAN Created clusters

Fig.3 shows the EM clustering result without applying optimization

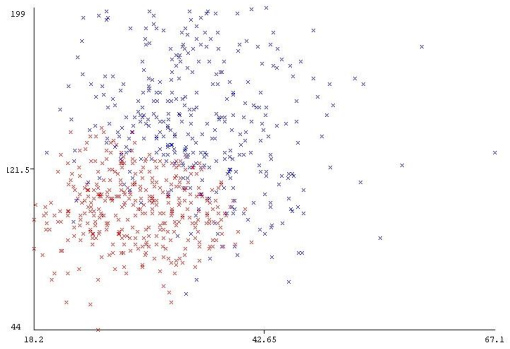

Figure 3 EM Created Clusters

Fig. 4 shows the Farthest First clustering result without applying optimization

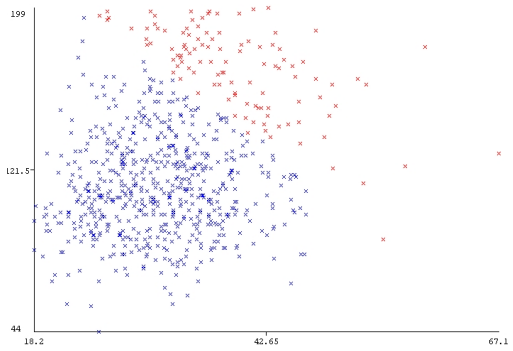

Figure 4 Farthest First Created Clusters

Fig.5 shows the K-means clustering result with optimization

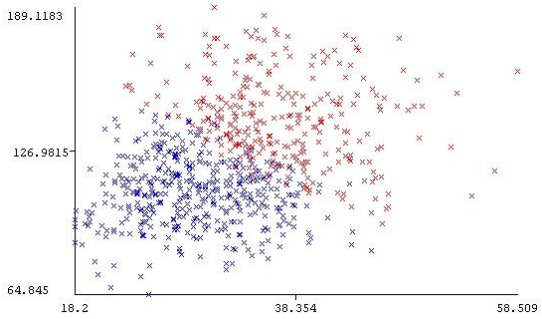

Figure 5 K-means created clusters with optimization

Fig.6 shows the DBSCAN clustering result with optimization

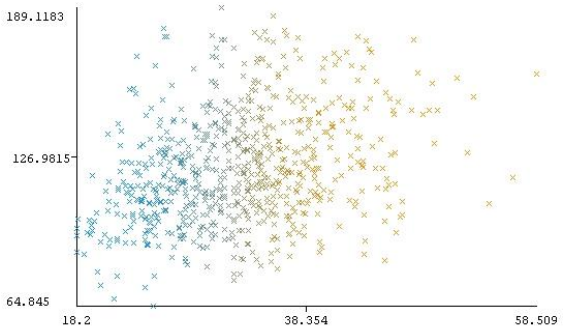

Figure 6 DBSCAN Created clusters with optimization

Fig.7 shows the EM clustering result with optimization

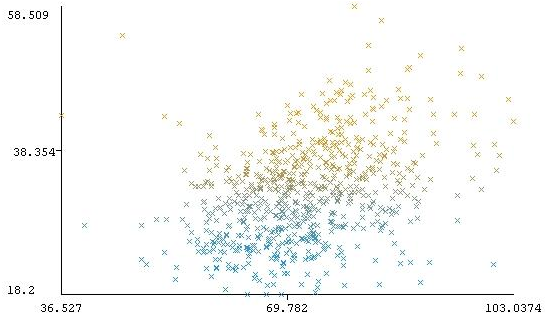

Figure 7 EM Created clusters with optimization

Fig.8 shows the Farthest First clustering result with optimization

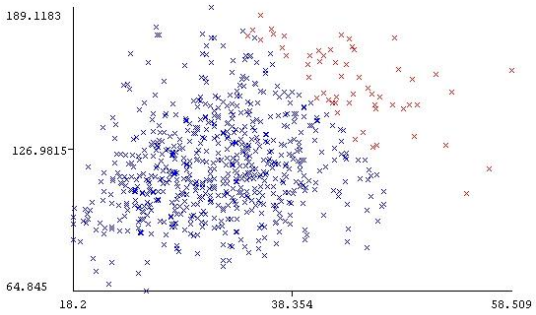

Figure 8 Farthest First Created clusters with optimization

The above figures show the improvements in cluster assignment the blue color shows 1st cluster points and red shows 2nd cluster points.

VI. CONCLUSION

Data analysis is a key for the diagnosis of disease in medical science specially when periodically exploding new viruses in this century. The disease of diabetes is horribly expending day to day in India as WHO is alarming that nearly 61% of Indians would be diabetic in 2020.

Correct prediction using available datasets is preferable to timely diagnosis and cure. In this study, it is depicted that if we use the proposed approach of optimization using functional approximation and curve fitting, can give better improved performance compared to existing clustering algorithms. Although no clustering technique can be applied in every dataset, we can only find the appropriate method by using them all. So the performance and feasibility of the clustering methods always depends upon what we want in data.

REFERENCES

1. Anuradha, S., Jyothirmai, P.,Tirumala, Y., Goutham, S., Hariprasad, V., (2014), "Comparative Study of Clustering Algorithms on Diabetes Data", International Journal of Engineering Research & Technology (IJERT) Volume 03, Issue 06.
2. A. K. Jain, M. N. Murty, P. J. Flynn (1999), "Data clustering: a review," ACM Computing Surveys, 31.
3. P. Padmaja, V. Srikanth, N. Siddiqui, D. Praveen, B. Ambica, V. B. V. E. Venkata Rao, and V.J.P. Raju Rudraraju (2008), Characteristic evaluation of diabetes data using clustering techniques, journal No.11.
4. Ramsay, J. (1982), "When the Data are Functions," Psychometrika, 47, 379396.
5. Huang,S.(2017).Functional Data Smoothing Methods and Their Applications.(Doctoraldissertation).Retrieved from https://scholarcommons.sc.edu/etd/4345
6. Hitchcock, D., Casella, G., and Booth, J. (2006), "Improved Estimation of Dissimilarities by Pre-Smoothing Functional Data". Journal of the American
Statistical Association, 101, 211-222.
7. Hitchcock, D., Booth, J., and Casella, G. (2007), "The Effect of Pre-Smoothing Functional Data on Cluster Analysis," Journal of Statistical Computation and Simulation, 77, 1043-1055.
8. Ramsay, J., and Silverman, B. (2005), Functional Data Analysis (2nd ed.), New York, NY: Springer-Verlag.
9. Vincent Sigillito (1990), Pima Indians Diabetes Database, retrieved from
https://datahub.io/machine-learning/diabetes#data-cli
10. Liyanapathirana, L. (2018). Machine Learning Workflow on Diabetes Data: Part 01. Towards Data Science. Available at: https://towardsdatascience.com/machine-learning-workflow-on-diabetes-data-part-01-573864fcc6b8 [Accessed 14 Jun. 2019].
11. Martin Ester, Hans-Peter Kriegel, Joerg Sander, Xiaowei Xu (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Institute for Computer Science, University of Munich. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)
12. Sthda.com. (2019). DBSCAN: density-based clustering for discovering clusters in large datasets with noise - Unsupervised Machine Learning - Easy Guides - Wiki - STHDA. Available at: http://www.sthda.com/english/wiki/wiki.php?id_contents=7940 [Accessed 14 Jun. 2019].
13. Nicolas, A. (2019). Fitting exponential curve to data. Mathematics Stack Exchange. Available at: https://math.stackexchange.com/questions/350754/fitting-exponential-curve-to-data [Accessed 20 Jun. 2019].
14. Dundas.com. (2019). Weighted Moving Average Learning. Available at: https://www.dundas.com/support/learning/documentation/design-view/formulas/standard/weighted-moving-average [Accessed 20 Jun. 2019].

AUTHORS PROFILE

Satish Kumar Soni, PhD (pursuing) in Computer Sciences, Barkatullah University, Bhopal. MCA from RGPV University, Bhopal, BSc (Statistics, Mathematics, Physics) from APS University, Rewa, currently teaching MSc Students of CS&IT in Computer Science & Application Dept. B.U., Bhopal. Research areas are Data Mining, Machine Learning, Pattern Recognition and Mathematical Modeling.

Dr. Ramjeevan Singh Thakur, Associate Professor(MCA), MANIT, Bhopal, Teacher, Researcher and consultant in the field of Computer Science and Information Technology, Ph.D. (Computer Science) from RGPV Bhopal, areas of interest include Data Mining, Data Warehousing, Web Mining, Text Mining, and Natural Language Processing, member of the CSI, IEEE, ACM, IAENG, ISTE, GAMS and IACSIT.

Dr. Anil Kumar Gupta, Head, Computer Science & Applications Dept., Barkatullah University, Bhopal, PhD in Computer Science, Active Researcher in the areas of Data Mining, Artificial Intelligence and Machine Learning.