International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 07 | July 2018 www.irjet.net p-ISSN: 2395-0072

Deep Incremental Statistical Closeness Factor Based Algorithm (DIS-CFBA) to assess Diabetes Mellitus

Rahul Raghvendra Joshi[1], Dr. Preeti Mulay[2]

[1,2]Symbiosis Institute of Technology (SIT), Pune, Affiliated to Symbiosis International Deemed University (SIDU), Pune, India

[1]Assistant Professor, Research Scholar & [2]Associate Professor, Phd Guide

---

Abstract - *Diabetes Mellitus (DM) is potential epidemic in India. Millions of individual diagnosed with this disease. Different types and number of lab tests used to diagnose it. This paper presents a novel tactic to assess DM based on CFBA and statistical methods. DM parameters assessed and categorized into significant and insignificant parameters. DM dataset processed using cluster analysis and principal component analysis (PCA). Cluster analysis creates fifteen distinct clusters. PCA applied to find out variability and to do categorization of parameters. Thus, this paper illustrates utility of statistical clustering for effective DM management.*

Key Words: Deep Incremental; Statistical; CFBA; Diabetes Mellitus etc.

1. INTRODUCTION

DM is an unending, enduring situation. It affects body's capacity to use the food energy. These day's it has treated as a major nuisance in health care industry. To develop evolved systems, incremental clustering acts as a key. It can help in effective knowledge augmentation. It in turn leads to effectual knowledge management [1, 2, 5 and 13-28]. Incremental learning can also be achieved through web based interactive data mining tools [4]. Data science is a booming field. Its innovative usage can reduce tests and trouble to patients during DM diagnosis. Data science coupled with Knowledge Management System (KMS) can alter traditional DM dealings pattern. Also, recommendations presented through this coupled system can act as preventive measures [3]. In this paper, a novel DIS-CFBA proposed and applied on different parameter values of patients.

Multivariate statistical approach used for different purposes like water quality assessment of river, assessment of trace elements levels in patients with type 2 diabetes, diabetes classification, identification of cigarette design influencing features and name a few [6-12]. In a similar way, multivariate statistical analysis used here to identify significant DM parameters.

Section 2 presents methods used. Section 3 throws light on results and related discussion followed by outlook in section 4. The references referred are listed at the end of this paper.

Table -1: DM ATTRIBUTES RANGE [6]

Sr. No.	Name of attribute	Range of attributes in mg/dl
1	BLOOD GLUCOSE FASTING	115-210
2	BLOOD GLUCOSE PP	140-250
3	CHOLESTROL	140-250
4	TRIGLYCERIDES	140-300
5	HDL CHOLESTROL	40-60
6	VLDL	20-60
7	LDL CHOLESTROL	60-115
8	NON HDL CHOLESTROL	120-170

2. METHODS USED

2.1 Cluster Analysis through DIS-CFBA

Input: Raw Dataset, Data Series (DS)

Output: Parameters categorization as significant and insignificant one

Closeness Factor Based Algorithm (CFBA)

1. Initial count of clusters K = 0
2. Calculate CF for DS (i)
3. Calculate CF for DS (i++)
4. Clusters formed on the basis of calculated CF
5. If (~Processed _Flag) then
   CF (added new cluster) = xi
   insert_counter (added new cluster) = 1 &
   CFBA_Clusters ← Clusters ∪ added new cluster
6. for all xi ∈ I
7. As Processed _Flag = False
8. For all clusters ∈ clusters do
9. if || xi || < = CF then
10. As Processed _Flag = False
11. Update (cluster)
12. insert_counter (cluster)
13. Processed _Flag = True
14. Exit loop

15. end if
16. end for
17. end if

DIS - CFBA

18. if (~Processed _Flag) then
19. if (Required_Parameter_Range (added new cluster)) == (Required_Parameter_Range (CFBA_Clusters))
20. Update Required_Parameter_Range (CFBA_Clusters)
21. Else discard (added new cluster) & Processed_Flag = True
22. end if
23. end if
24. Correlation, KMO & Bartlett's test & Factor Loadings on required_Parameter_Range (CFBA_Clusters)
25. Parameters_Categorization as significant and insignificant one
26. Exit

2.2 Sampling

The DM data set of working adults whose age is between 35 to 45 years considered for 2016 -17 [6]. Sample analysis takes place through below mentioned principal activities.

1. Table 1 shows DM parameters range based on pathology reports.
2. After application of DIS-CFBA fifteen distinct clusters obtained.
3. Required parameters range analyzed on the basis of obtained clusters.
4. Statistical methods applied on this analysis does categorization of parameters.

2.3 Statistical Data Treatment

DM data generated through clustering normalized by log normal transformation. The fitness of data for PCA verified through Kaiser-Mayer-Olkin (KMO) and Barlett's tests. The DM management subjected to two major approaches viz., cluster analysis and PCA. All statistical calculations carried out through Minitab 17.0 software. KMO and Barlett's tests performed through R programming language.

2.4 Principal Component Analysis (PCA)

Two Key aspects of PCA used viz. Data reduction and summarization. It analyzes interrelationships among attributes with their common causal dimensions, branded as parameters [9]. PCA performed on normalized parameters to get significant Principal Components (PCs). These PCs further lessen the contribution of parameters with minor significance. These PCs subjected to varimax rotation generates Vari-Factors (VFs) [9]. PCs defined as parameters when their variance is greater than 1 [9]. The standard test score of any parameter must be more than variance of any single parameter [9]. Hence, PCA coupled with varimax rotation along with Kaiser Normalization applied. The extracted Eigen values from correlation matrix, significant factors and variance percent presented.

3. Results and Discussion

Statistical summary of extracted DM parameters through clustering observed at developed clusters. Table 2 illustrates the same. The values of Triglycerides vary from 80 to 530 which show its importance in relation to DM. Also, the values of Blood Glucose vary from 135 to 350 which is second important attribute in relation to DM. HDL Cholesterol, VLDL, LDL Cholesterol and Non HDL Cholesterol have lower ranges which implies their DM specific importance.
Correlation coefficient of DM parameters presented in Table 3. There is strong positive correlation between Blood Glucose Fasting and Triglycerides (r=0.41). Also, significant correlation has found between Blood Glucose and Triglycerides (r=0.36). Moderate correlation is there between Blood Glucose PP and Cholesterol (r=0.33).

Table -2: Summary of DM Parameters of Generated Clusters

Sr. No.	Parameters	Min	Max	Mean	Std. Dev.
I	Blood Glucose Fasting (mg/dl)	100	260	144.77	21.41
II	Blood Glucose PP (mg/dl)	135	350	200.6	24.82
III	Cholesterol (mg/dl)	130	240	179.66	25.03
IV	Triglycerides (mg/dl)	80	530	231.3	26.67
V	HDL Cholesterol (mg/dl)	30	59	49.49	5.58
VI	VLDL (mg/dl)	15	106	38.51	10.26
VII	LDL Cholesterol (mg/dl)	57	143	88.08	14.01
VIII	Non HDL Cholesterol (mg/dl)	90	185	141.56	12.74

Before PCA, dataset normalized using log normal transformation and Kolmogorove Smirnov (K-S) statistics test. These tests verify the goodness of fit of the data to log-normal distribution [8]. The results of K-S test shows that all attributes follows the log normal distribution. To investigate suitability of data for PCA, KMO and Barlett's test of sphericity performed [7]. The significance value of 0.62 (Table 4) indicates dataset is fit for PCA. Barlett's test of

sphericity test used to check relatedness of parameters. Its 0.57 Significance level indicates correlation matrix is an identity matrix. So, parameters are related to each other. PCA applied on obtained clusters using Minitab 17.0 software. PCA yields correlation matrix for parameters I to VIII (TABLE 3) and factors extracted through centroid method of varimax rotation. Eigen value indicates the significance of the parameter. Eigen value more than one treated as significant one [10, 11, 12].

Table -3: CORRELATION MATRIX (PEARSON (n))

	I	II	III	IV	V	VI	VII	VIII
I	1							
II	0.02	1						
III	0.2	0.33	1					
IV	0.41	0.36	0.28	1				
V	0	0.11	0.09	-0.02	1			
VI	0.09	0.05	0.03	0.11	0.06	1		
VII	0.08	0.14	0.01	0.24	0.02	0.09	1	
VIII	0.11	0.14	-0.03	0.2	0.01	0.08	0.03	1

Table -4: KMO AND BARLETT'S TEST

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.62
Approx. Chi-Square		5.69
Bartlett's Test of Sphericity	df	7
	Sig.	0.57

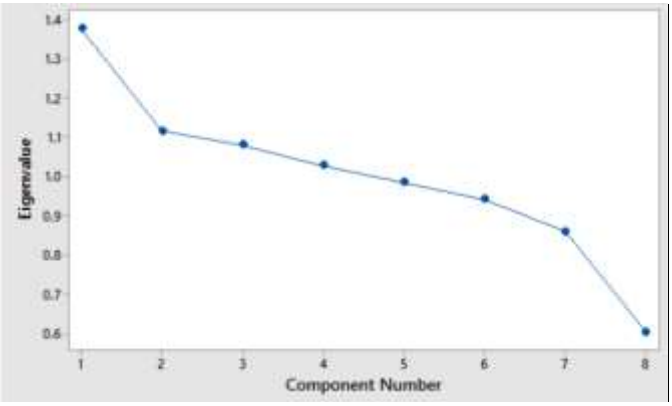

Figure -1: SCREE PLOT FOR PARAMETERS

PCA result shows first four Eigen values are higher than one and considered as significant. The scree plot (Fig. 1) shows greater part of variance in the original data covered by first four parameters. These four PCs accounts for 57.6% of the total variance associated with all 08 parameters. Table 5 shows variance of first four parameters for rotated and unrotated factor loadings. PC1 has 16% of total variance. It has strong positive loading of Triglycerides, negative loading of blood glucose. PC2 handles 14% of total variance, has strong positive loading of LDL Cholesterol. This factor thus acts as a reactive component for DM. Cholesterol reveals its DM specific influence in relation to obtained loadings. PC3 has a strong loading of blood glucose fasting and accounts for 13% of variance. PC4 exhibited again 13% of total variance within the data set. It has strong positive loading of VLDL and HDL Cholesterol. VLDL and HDL Cholesterol are indicative details of PC4. Table 6 represents the factor loadings of various parameters.

Table -5: VARIANCE

Parameters	Principal Components			
	1	2	3	4
Blood Glucose Fasting	0.013	0.029	-0.906	0.016
Blood Glucose PP	-0.812	0.091	0.139	-0.068
Cholesterol	0.065	-0.769	0.319	0.027
Triglycerides	0.811	0.070	0.134	-0.074
HDL Cholesterol	0.044	-0.099	-0.044	0.406
VLDL	-0.123	-0.142	-0.055	0.661
LDL Cholesterol	0.027	0.546	0.284	-0.181
Non HDL Cholesterol	0.077	0.422	0.172	0.654

Table -6: FACTOR LOADINGS

PCs	Sum of squared loadings		Rotated sum of squared loadings	
	Variance	% of Variance	Variance	% of variance
1	1.35	0.16	1.34	0.16
2	1.11	0.14	1.10	0.14
3	1.08	0.135	1.075	0.13
4	1.05	0.132	1.072	0.12

Hierarchical cluster analysis used to club DM attributes with certain similarity. The dendogram (Fig. 2) categorizes eight parameters of DM in two clusters on the basis of similarity between their characteristics. Cluster 1 consists of first five attributes while cluster 2 consists of remaining three. Cluster 1 and cluster 2 corresponded to high and low significant DM parameters. The patient's data among these clusters can attribute to difference in their influence to DM.

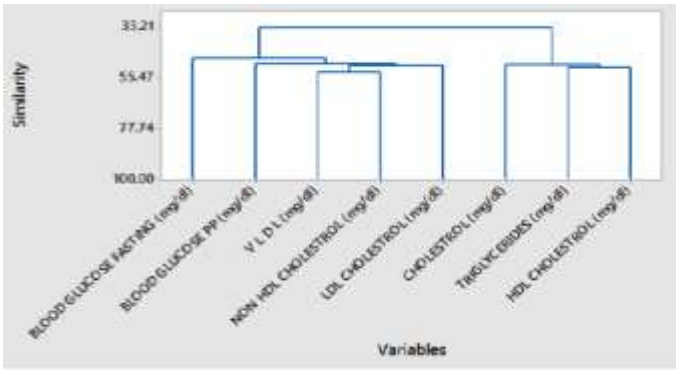

Figure -2: DENDOGRAM FOR DM PARAMETERS

4. CONCLUSION

Statistical techniques powered by clustering and PCA applied on 5K dataset for 2016-2017. Blood Glucose Fasting, Blood Glucose PP, VLDL and Non HDL cholesterol are significant parameters. This illustrates utility of statistical techniques in DM assessment. So, cut down of these parameters can improve DM management. This information can also alter pathology test pattern for DM.

REFERENCES

[1] Mulay, P. (2016). Threshold Computation to Discover Cluster Structure: A New Approach. International Journal of Electrical and Computer Engineering (IJECE), 6(1), 275-282.

[2] Karuna, P., & Preeti, M. (2016). Global Plagiarism Management Through Intelligence of Hawk Eye. *Indian Journal of Science and Technology*, *9*(15).

[3] Mulay, P., & Mulatu, S. (2016). What You Eat Matters Road Safety: A Data Mining Approach. *Indian Journal of Science and Technology*, *9*(15).

[4] Borhade, M., & Mulay, P. (2015). Online interactive data mining tool. *Procedia Computer Science*, *50*, 335-340.

[5] Mulay, P., & Kulkarni, P. A. (2013). Knowledge augmentation via incremental clustering: new technology for effective knowledge management. *International Journal of Business Information Systems*, *12*(1), 68-87.

[6] Mulay, P., Joshi, R. R., Anguria, A. K., Gonsalves, A., Deepankar, D., & Ghosh, D. (2017). Threshold Based Clustering Algorithm Analyzes Diabetic Mellitus. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications (pp. 27-33). Springer, Singapore.

[7] Rizvi, N., Katyal, D., & Joshi, V. (2015). A Multivariate Statistical Approach for Water Quality Assessment of River Hindon, India. World Academy of Science, Engineering and Technology, International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering, 10(1), 6-11.

[8] Badran, M., Morsy, R., Soliman, H., & Elnimr, T. (2016). Assessment of trace elements levels in patients with Type 2 diabetes using multivariate statistical analysis. Journal of Trace Elements in Medicine and Biology, 33, 114-119.

[9] Nilashi, M., Ibrahim, O. B., Mardani, A., Ahani, A., & Jusoh, A. (2016). A soft computing approach for diabetes disease classification. Health Informatics Journal, 1, 15.

[10] Agnew-Heard, K. A., Lancaster, V. A., Bravo, R., Watson, C. H., Walters, M. J., & Holman, M. R. (2016). Multivariate Statistical Analysis of Cigarette Design Features Influence on ISO TNCO Yields. Chemical Research in Toxicology.

[11] Shyamala, G., & Jeyanthi, J. (2016). Application of Integrated Hydrochemical Model and Cluster Analysis in Assessing Groundwater Quality. International Journal of Ecology & Development™, 31(4), 34-45.

[12] Jung, K. Y., Lee, K. L., Im, T. H., Lee, I. J., Kim, S., Han, K. Y., & Ahn, J. M. (2016). Evaluation of water quality for the Nakdong River watershed using multivariate analysis. Environmental Technology & Innovation, 5, 67-82.

[13] Mulay, P., & Kulkarni, P. (2008). An Automated Forecasting Tool (AFT) achieved by clustering Entity Relationship Model. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY, 8(12), 371-381.

[14] Mulay, Preeti, and Parag Kulkarni. "Support Vector Machine based, project simulation with focus on Security in software development Introducing Safe Software Development Life Cycle (SSDLC) model." IJCSNS 8.11 (2008): 393.

[15] Mulay, P., & Shinde, K. (2019). Personalized Diabetes Analysis Using Correlation-Based Incremental Clustering Algorithm. In Big Data Processing Using Spark in Cloud (pp. 167-193). Springer, Singapore.

[16] Khobragade, S., & Mulay, P. (2018). Enhance Incremental Clustering for Time Series Datasets Using Distance Measures. In International Conference on Intelligent Computing and Applications (pp. 543-556). Springer, Singapore.

[17] Mulay, P., & Ahire, P. (2017). Knowledge Management Academic Research:"NUMPATIBILITY"–Numeral Era of Compatibility. In Enhancing Academic Research With Knowledge Management Principles (pp. 45-91). IGI Global.

[18] Mulay, P., Patel, K., & Gauchia, H. G. (2017). Distributed System Implementation Based on "Ants Feeding Birds" Algorithm: Electronics Transformation via Animals and Human. In Detecting and Mitigating Robotic Cyber Security Risks (pp. 51-85). IGI Global.

[19] Alehegn, M., & Mulay, R. J. D. P. Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm.

[20] Shinde, K., & Mulay, P. (2017, April). Cbica: Correlation based incremental clustering algorithm, a new approach. In Convergence in Technology (I2CT), 2017 2nd International Conference for (pp. 291-296). IEEE.

[21] Gaikwad, S. M., Joshi, R., & Gaikwad, S. M. (2016, March). Modified analytical hierarchy process to recommend an ice cream to a diabetic patient. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (p. 138). ACM.

[22] Ahire, P. R., & Mulay, P. (2016). Discover compatibility: Machine learning way. Journal of Theoretical & Applied Information Technology, 86(3).

[23] Laddha, A. R., Joshi, R. R., & Mulay, P. (2016). ENRICHING PROCESS OF ICE-CREAM RECOMMENDATION USING COMBINATORIAL RANKING OF AHP AND MONTE CARLO AHP. Journal of Theoretical & Applied Information Technology, 85(3).

[24] Gaikwad, S. M., Joshi, R. R., & Mulay, P. (2015). Attribute visualization and cluster mapping with the help of new proposed algorithm and modified cluster formation algorithm to recommend an ice cream to the diabetic patient based on sugar contain in it. International Journal of Applied Engineering Research, 10.

[25] Kulkarni, P. A., & Mulay, P. (2013). Evolve systems using incremental clustering approach. Evolving Systems, 4(2), 71-85.

[26] Gaikwad, S. M. (2015). Cluster mapping with the help of new proposed algorithm and MCF algorithm to recommend an ice cream to the diabetic Patient. International Journal of Applied Engineering Research, 10(8), 21259-21266.

[27] Gaikwad, S. M., Mulay, P., & Joshi, R. R. (2015). Mapping with the help of new Proposed Algorithm and Modified Cluster Formation Algorithm to recommend an Ice Cream to the Diabetic Patient based on Sugar Contain in it. International Journal of Students' Research in Technology & Management, 3(6), 410-412.

[28] Mulay, M. P. Incremental Learning.