A Multilevel Clustering Approach Driven by Continuous Glucose Monitoring Data for Further Classification of Type 2 Diabetes

Rui Tao[#1], Xia Yu[#1], Jingyi Lu[#2], Yun Shen[2], Wei Lu[2], Wei Zhu[2], Yuqian Bao[2],

Hongru Li[1], Jian Zhou[*2]

[1] College of Information Sciences and Engineering, Northeastern University, Shenyang, 110819, China

[2] Department of Endocrinology and Metabolism, Shanghai Clinical Center for Diabetes, Shanghai Diabetes Institute, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai 200233, China

Keywords: Type 2 Diabetes, multilevel cluster analysis, knowledge mining, CGM data, clinical phenotypes

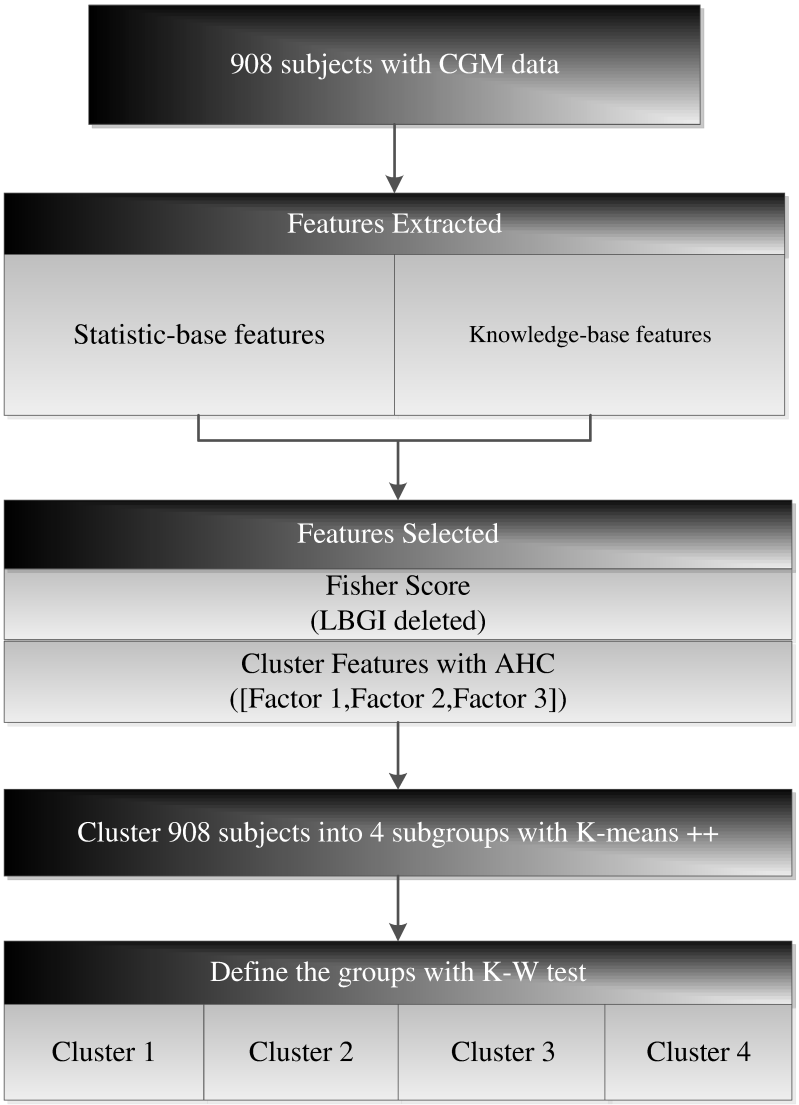
# These three authors contributed equally to the paper.

* Corresponding author.

E-mail address: zhoujian@sjtu.edu.cn (Jian Zhou)

1. The overview of multilevel clustering of CGM data:

The multilevel clustering method as a data-mining model contains four steps. The first step extract features from CGM data. Statistic-base features like mean values (mean sensor glucose, MSG), standard deviation of the values (SD of the sensor glucose, SDBG) and coefficient of variation (CV) were calculated. In addition, some knowledge-base features which be used in clinical also extracted. Globally, a 15-dimensional group of feature variables extracted from time-series CGM data was used to characterize the patient for future analysis. Then we designed a step for selecting features in the feature group. In this step, an effective feature selection method based on Fisher score and agglomerative hierarchical clustering was applied to find the pattern similarities of CGM features and reducing features. Through this step, the original 15-dimensional features could be further reduced to 3-dimensional features with clinical significance. In the third step, we used the 3-dimensional features in the second step to characterize the participants. And subgroups of the participants were found by K-means++ method. At the last step, four subgroups were identified and their clinical phenotypes were compared. The overall flow chart is shown in **Supplementary Figure S1**. Furthermore, the details of the method are shown in **Supplementary Table S1**.



Supplementary Figures S1: The flow chart of the cluster method.
Abbreviations: LBGI-low blood glucose indices, K-W test-Kruskal Wails test, AHC-agglomerative hierarchical clustering

Supplementary Table S1: The iterative procedure of Multilevel cluster for CGM data

Algorithm Multilevel cluster for CGM data (MLC-CGM)
Input: CGM datasets ***Output:*** the groups' labels of each subject ***Step 1: Feature Extraction*** ***Step 1.1:*** Calculate ***knowledge-based features*** ***Step 1.2:*** Calculate ***statistical-base features*** ***Step 1.3:*** Initialize the set of features ***S*** ***Step 2: Feature Selection*** ***Step 2.1:*** Build a simple clustering using ***S*** as input ***Step 2.2:*** For every feature f_i in ***S***: ***Step 2.2.1:*** Compute the Fisher Score (***FS***). ***Step 2.2.2:*** Find $f^* = f_i$ if ***FS>threshold*** such that adding it to ***S set=S*** ***Step 2.3:*** Update the ***S* set*** ***Step 2.4:*** Cluster features in ***S* set*** ***Step 3: Subjects Cluster*** ***Step 3.1:*** Choose the optimal clustering number ***n*** by elbow method ***Step 3.2:*** Cluster patients into n subgroups among ***S* set*** ***Step 4:*** Definite the subgroups using ***Kruskal-Wallis test***

2. Fisher score

As a state-of-the-art feature selection, the main idea of the Fisher score is that the features with strong discrimination performance keep the distance within clusters as small as possible and keep the distance between clusters as large as possible. Compared to Mutual information (MI),[1] Relieff,[2] Laplacian Score[3] and Trace ratio criterion,[4] Fisher Score (FS) is one of the most widely used criteria for filter-based feature selection [5]. Here, the Fisher score of a single feature, used as the criterion for selecting the CGM indices can be expressed by the following formulas:

$$J_{\text{fisher}}(k) = S_B(k) / S_w(k)$$
$$S_B(k) = \sum_{i=1}^{c} \frac{n_i}{n} \left(m_i(k) - m(k) \right)^2 \quad (1)$$
$$S_w(k) = \sum_{i=1}^{c} \sum_{x \in w_i} \left(x(k) - m_i(k) \right)^2$$

where there are n samples in the data set that belong to C clusters, each cluster contains n_i samples respectively. $x(k)$, $m_i(k)$, $m(k)$ means the sample x, mean values of the i cluster, and all samples' mean values on the k features. The Fisher score under the threshold will be removed [6]. Here, we set the threshold value to 0.5. And the Fisher score of the CGM features are shown in **Supplementary Tables S2.**

Supplementary Table S2. The fisher score results in the feature selection

Feature	FS values	Features	FS values	Features	FS values
MSG	3.23	CV	0.92	LBGI	0.22
MPSG	2.53	LAGE	1.44	HBGI	3.75
TIR	3.44	M value	2.98	ADRR	2.08
TOR	3.44	MAGE	1.00	GRADE	2.37
SDSG	1.46	J-index	3.56	LI	0.57

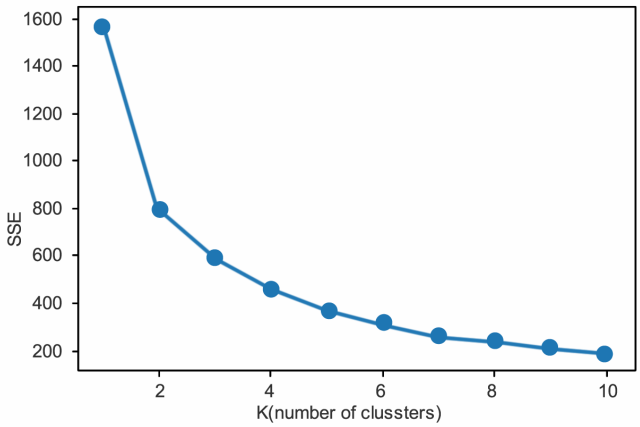
Abbreviations: MSG-mean sensor glucose; MPSG-mean postprandial sensor glucose; TIR-percentages of values within the target range ([3.9-10] mmol/L); TOR-percentages of values out of the target range (<3.9 or >10 mmol/L); SDSG-standard deviation of the sensor glucose, CV-coefficient of variation; LAGE-largest amplitude of glycemic excursions; MAGE-mean amplitude of glycemic excursions; HBGI,LBGI-high or low blood glucose indices; GRADE-glycemic risk assessment diabetes equation score; LI-liability index; ADRR-average daily risk range.

3. Elbow method

To properly choose the optimal number of the K-means++ on the basis of the elbow method [7 8], we divided the data into categories 1 to 11. Then the error sum of the square distance (SSE) between the particle of each cluster and the sample point in the cluster were calculated as:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - u_i\|^2$$
$$u_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \quad (2)$$

where u_i is the center of cluster i, x is the sample point, and k is the value of K-means++. Distortions of the patient's matrix at 4 will be the most appropriate values as shown in **Supplementary Figures S2.**



Supplementary Figure S2.the SSE line chart of the elbow method. The curve reached the break point at k=4. As k increases, SSE decreases slowly.

Supplementary Table S3. Two sample Kolmogorov-Smirnov test results

SEX	AGE	BMI	Duration	SBP	DBP
	0.000	0.001	0.109	0.067	0.158
SEX	HbA1c	HDL	LDL	FCP	CP2h
	0.003	0.000	0.025	0.145	0.175

A two-tailed P value of <0.010 was considered statistically significant. As shown in table, most adjusted P values above 0.010. These results verified the similar distributions at characteristics in men and women. So, sex could not be considered in the analysis.

Abbreviations: FCP -C peptide variables as fasting C peptide; CP2h-C peptide levels at 2 hours; ΔCP120min- the increments of C peptide in plasma levels 90 minutes; SBP-systolic blood pressure; DBP-diastolic blood pressure; BMI -body mass index; Duration-duration of diabetes; HDL-high density lipoprotein; LDL-low density lipoprotein.

Supplementary Table S4. Glycemic level factor (GLF), the glycemic fluctuation factor (GFF) and negative glycemic factor (NGF) correlation analysis

	MSG	HBGI	CV	TIR
GLF	0.988	0.993	0.120	-0.925
GFF	0.365	0.530	0.899	-0.553
NGF	-0.986	-0.980	-0.047*	0.935

A two-tailed P value of <0.010 was considered statistically significant. And the number with a * means its P value above 0.010. As shown in the table (Table S2), GLF had a higher correlation with MSG and HBGI, and GFF had the highest correlation with CV. NGF had the negative correlation with the indices unlike Factor 1 and Factor 2.

Abbreviations: MSG-mean sensor glucose; HBGI- high blood glucose indices; CV-coefficient variation; TIR-time in range **[3.9-10] mmol/L**

References:

1. Koller D, Sahami M. Toward optimal feature selection: Stanford InfoLab, 1996.
2. Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. *J Machine learning* 2003;53(1-2):23-69.
3. X He, D Cai, P Niyogi P. Laplacian score for feature selection. Advances in neural information processing systems; 2006.
4. Nie F, Xiang S, Jia Y. et.al.: Trace ratio criterion for feature selection. AAAI 2008; 2:671-676.
5. Tang JA, Salem, Liu H. Feature selection for classification: A review. *Data classification: Algorithms applications* 2014:37.
6. Zhou M. Hybrid feature selection method based on fisher score and genetic algorithm. *J Journal of Mathematical Sciences: Advances Applications* 2016;37(37):51-78.
7. Pham DT, Dimov SS, Nguyen CD. Selection of K in K-means clustering. *Journal of Mechanical Engineering Science* 2005;219(1):103-19.
8. Bholowalia PK, Arvind EBK-means: A clustering technique based on elbow method and k-means in WSN. *J International Journal of Computer Applications* 2014;105(9)