Supplementary material

Proof of the problem P1

If we set Eq. 11 to zero, we have

$$\sum_{i=1}^{n_k} u_{gip}[(x_{ij} - z_{gpj})(z_{gpj} - z_{gGj})^2 + (z_{gpj} - z_{gGj})(x_{ij} - z_{gpj})^2] = 0 \quad (27)$$

Which gives:

$$\sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gpj})(z_{gpj} - z_{gGj})(\cancel{z_{gpj}} - z_{gGj} + x_{ij} - \cancel{z_{gpj}}) = 0 \quad (28)$$

If $z_{gpj} \neq z_{gGj}$, we have:

$$z_{kpj} = \frac{\sum_{i=1}^{n_g} u_{gip} x_{ij}(x_{ij} - z_{gGj})}{\sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gGj})} \quad (29)$$

Constantly necessary and sufficient condition for this equality to be realizable is when:

$$x_{min} \leq z_{gpj} \leq x_{max} \tag{30}$$

where:

$x_{min} = \min_{i=1,\ldots,n} u_{gip}x_{ij}$ and $x_{max} = \max_{i=1,\ldots,n} u_{gip}x_{ij}$ for the subgoup p from the apriori group K

Suppose that $\sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gGj}) > 0$ the inequality becomes:

$$x_{min} \sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gGj}) \underset{(1)}{\leq} \sum_{i=1}^{n_g} u_{gip}x_{ij}(x_{ij} - z_{gGj}) \underset{(2)}{\leq} x_{max} \sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gGj}) \tag{31}$$

For the inequality (1):

$$\sum_{i=1}^{n_g} u_{kip}x_{ij}(x_{ij} - z_{gGj}) \geq x_{min} \sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gGj}) \tag{32}$$

We get:

$$z_{kGj} \leq \frac{\sum_{i=1}^{n_g} u_{gip}x_{ij}^2 - x_{min} \sum_{i=1}^{n_g} u_{gip}x_{ij}}{\sum_{i=1}^{n_g} u_{gip}x_{ij} - x_{min}n_g} \tag{33}$$

$$z_{gGj} \leq \frac{\sum_{i=1}^{n_g} u_{gip}x_{ij}^2 - x_{min} \sum_{i=1}^{n_g} u_{gip}x_{ij}}{n_g(\bar{X}_{gp} - x_{min})} \tag{34}$$

For the inequality (2):

$$\sum_{i=1}^{n_g} u_{gip}x_{ij}(x_{ij} - z_{gGj}) \leq x_{max} \sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gGj}) \tag{35}$$

We get:

$$z_{gGj} \leq \frac{x_{max}n_g\bar{X}_{gp} - \sum_{i=1}^{n_g} u_{gip}x_{ij}^2}{n_g(x_{max} - \bar{X}_{gp})} \tag{36}$$

So:

$$z_{gGj} \leq \min\{\frac{\sum_{i=1}^{n_g} u_{gip}x_{ij}^2 - x_{min}\sum_{i=1}^{n_g} u_{gip}x_{ij}}{n_g(\bar{X}_{gp} - x_{min})}, \frac{x_{max}n_g\bar{X}_{gp} - \sum_{i=1}^{n_g} u_{gip}x_{ij}^2}{n_g(x_{max} - \bar{X}_{gp})}\} \tag{37}$$

Suppose that $\sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gGj}) < 0$ the inequality becomes:

$$x_{max}\sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gGj}) \underset{(1)}{\leq} \sum_{i=1}^{n_g} u_{gip}x_{ij}(x_{ij} - z_{gGj}) \underset{(2)}{\leq} x_{min}\sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gGj}) \tag{38}$$

For the inequality (1):

$$\sum_{i=1}^{n_g} u_{gip}x_{ij}(x_{ij} - z_{gGj}) \geq x_{max}\sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gGj}) \tag{39}$$

We get:

$$z_{gGj} \geq \frac{x_{max}n_g\bar{X}_{gp} - \sum_{i=1}^{n_g} u_{gip}x_{ij}^2}{n_g(x_{max} - \bar{X}_{gp})} \tag{40}$$

For the inequality (2):

$$\sum_{i=1}^{n_g} u_{gip}x_{ij}(x_{ij} - z_{gGj}) \leq x_{min}\sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gGj}) \tag{41}$$

We get:

$$z_{gGj} \geq \frac{\sum_{i=1}^{n_g} u_{kip}x_{ij}^2 - x_{min}\sum_{i=1}^{n_g} u_{gip}x_{ij}}{\sum_{i=1}^{n_g} u_{gip}x_{ij} - x_{min}n_g} \tag{42}$$

$$z_{gGj} \geq \frac{\sum_{i=1}^{n_g} u_{gip}x_{ij}^2 - x_{min}\sum_{i=1}^{n_g} u_{gip}x_{ij}}{n_g(\bar{X}_{gp} - x_{min})} \tag{43}$$

So:

$$z_{gGj} \geq \max\left\{ \frac{x_{max} n_g \bar{X}_{gp} - \sum_{i=1}^{n_g} u_{gip} x_{ij}^2}{n_g (x_{max} - \bar{X}_{gp})}, \frac{\sum_{i=1}^{n_g} u_{gip} x_{ij}^2 - x_{min} \sum_{i=1}^{n_g} u_{gip} x_{ij}}{n_g (\bar{X}_{gp} - x_{min})} \right\} \tag{44}$$

To assure that z_{gGj} is a solution for Eq. 11 it suffices to verify that z_{gGj} meets the following conditions:

$$\begin{cases} z_{gGj} \leq \min\left\{ \frac{\sum_{i=1}^{n_g} u_{gip} x_{ij}^2 - x_{min} \sum_{i=1}^{n_g} u_{gip} x_{ij}}{n_g (\bar{X}_{gp} - x_{min})}, \frac{x_{max} n_g \bar{X}_{gp} - \sum_{i=1}^{n_g} u_{gip} x_{ij}^2}{n_g (x_{max} - \bar{X}_{gp})} \right\} & \text{if} \\ \displaystyle\sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gGj}) > 0 & \\ z_{gGj} \geq \max\left\{ \frac{x_{max} n_g \bar{X}_{gp} - \sum_{i=1}^{n_g} u_{gip} x_{ij}^2}{n_g (x_{max} - \bar{X}_{gp})}, \frac{\sum_{i=1}^{n_g} u_{gip} x_{ij}^2 - x_{min} \sum_{i=1}^{n_g} u_{gip} x_{ij}}{n_g (\bar{X}_{gp} - x_{min})} \right\} & \text{if} \\ \displaystyle\sum_{i=1}^{n_g} u_{gip}(x_{ij} - z_{gGj}) < 0 & \end{cases} \tag{45}$$

Description of the Data Generation Processes

This section aims to offer precise and clear definitions for each DGP, ensuring understanding and clarity. It is essential to ensure a comprehensive understanding of each DGP in order to facilitate accurate interpretation and analysis of the generated data:

DGP 1: Two clusters were created by utilizing independent Gaussian random variables. Cluster 1 has a mean value of 0 and a covariance matrix of $0.8^2 I_{40}$, where I_{40} represents the identity matrix of size 40. Cluster 2 shows a mean of 0.1 and a covariance matrix of $0.9^2 I_{40}$.

DGP 2: Two distinct clusters were created using independent random variables that had different distributions. Cluster 1 is formed from a

Lognormal distribution that produces samples from a conventional normal distribution with a mean of 0 and a covariance matrix of I_{45}. Cluster 2 is formed by applying the Pareto distribution with a shape parameter of 2.62.

DGP 3: Three clusters were created by utilizing independent Gaussian random variables. Cluster 1 has a mean value of 0 and a covariance matrix of $0.7^2 I_{40}$. Cluster 2 has a mean of 0.1 and a covariance matrix of $0.6^2 I_{40}$, while cluster 3 has a mean of 0.7 and a covariance matrix of $0.5^2 I_{40}$.

DGP 4: Three distinct clusters were generated by using independent random variables that had different distributions. Cluster 1 is generated when samples from a Chi-Square distribution with a degree of freedom of 2 are drawn from a Lognormal distribution. Cluster 2 is produced when an exponential distribution is utilized to represent the intervals between events in a Poisson process with a scale of 1. Conversely, cluster 3 is generated using a uniform distribution, which yields values that are uniformly distributed within the range of 0 to 1.5.

DGP 5: Four clusters were created by utilizing independent Gaussian random variables. Cluster 1 has a mean value of 0 and a covariance matrix of I_{40}. Cluster 2 has a mean of 0.5 and a covariance matrix of the identity matrix with dimensions 40. Cluster 3 has a mean of 1 and a covariance matrix of the same identity matrix. Cluster 4 has a mean of 1.5 and a covariance matrix of $1.5^2 I_{40}$.

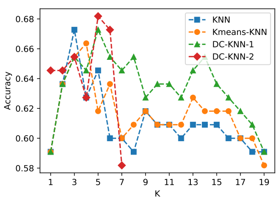
DGP 6: Five distinct clusters were created using independently generated random variables with different distributions. Specifically, the first cluster was derived from non-central t-distributions with 25 degrees of freedom and a non-centrality parameter of 1.5. The second cluster was produced from gamma distributions, represented by the symbol $Gam(3,\ 1.2)$, where 3 and 1.2 indicate the shape and rate parameters, respectively. The third cluster was formed using a uniform distribution over the continuous interval between 1 and 5. The fourth cluster was composed of independent Gaussian random variables with a mean of -1 and a covariance matrix of $1.5^2 I_{30}$. Finally, the fifth cluster formed from the Gaussian distribution of mean 2 and a covariance matrix of $2^2 I_{30}$.

Table 4: Specifications of DGPs in the Simulation Study. N: denotes the number of classes, p: denotes the number of variables, and n: denotes the number of observations.

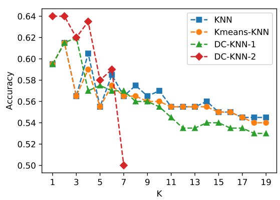
	N	p	Distributions	Clusters size	n
DGP1	2	40	Gaussians	(300, 250)	550
DGP2	2	45	Lognormal, Pareto	(500, 500)	1000
DGP3	3	30	Gaussians	(200, 400, 300)	900
DGP4	3	45	Chi-Square, Exponential, Uniform	(500, 500, 500)	1500
DGP5	4	27	Multi-variate Gaussian distributions	(350, 200, 300, 250)	1100
DGP6	5	30	Non-central t-distributions, uniform, Gamma and Gaussians	(250, 200, 250, 200, 200)	1100

Analysis of classification accuracy on simulated data for different values of *K*

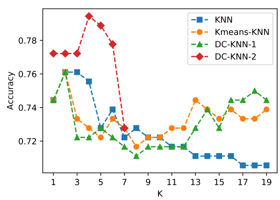
In our main paper, we examine the classification effectiveness of the DC-KNN techniques by varying the value of K, particularly focusing on the classification accuracy with simulated datasets. As detailed in Subsection 4.2 of the main paper, K ranges from 1 to 20, increasing incrementally by 1. We discuss the classification accuracy for various K values and illustrate these findings in Figure 4. Our analysis shows that DC-KNN2's classification accuracy is notably impacted by the different K values, achieving better results with smaller K values, which remain constant for larger K. This behavior aligns with the phenomenon where K exceeds the sum of apriori class subgroups. Moreover, we found that DC-KNN1 consistently outperforms traditional KNN and Kmeans-KNN across all K values, underscoring the beneficial impact of DC's new objective function on classifier performance. The determination of optimal subgroup numbers in our proposed method is data-dependent, as further elaborated in our study.



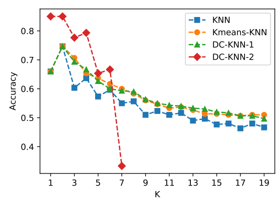
(a) DGP 1



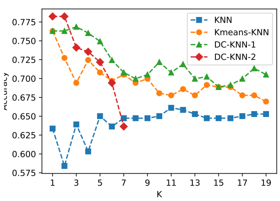
(b) DGP 2



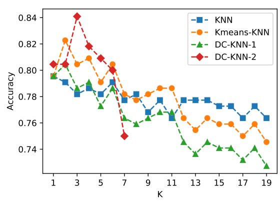
(c) DGP 3



(d) DGP 4



(e) DGP 5



(f) DGP 6

Fig. 4: The classification accuracy for each method on the simulated datasets with varying K values