The development of a grey clustering toolbox and its application to test score clustering

Kun-Li Wen† & Yi-Fung Huang‡

Chienkuo Technology University, Changhua, Taiwan†
Tunghai University, Taichung, Taiwan‡

ABSTRACT: The theme of this article is to design a *grey clustering toolbox* via C++ in the GM model to facilitate easy studying and learning of grey system theory. In the past, research about the grey model toolbox was focused on a case-by-case basis. Although all studies have shown good results, only individual case results could be presented. The full details were still found to be lacking. Therefore, in this study, C++ language is first used in order to develop a toolbox for a grey clustering model. Secondly, an example in test score clustering is given to implement the feasibility of this toolbox. Through the results, the grey clustering toolbox in the GM field not only presents many figures in order to make the results clearer, but also upgrades the analytical level in grey system theory during the whole implementation of this approach.

INTRODUCTION

The main study in this article is based on grey system theory. Much research on grey clustering has been published over the past five years, although it has been somewhat uneven [1-15]. Hence, the authors have developed a grey clustering toolbox; however, the authors do not intend to replace the traditional method; they seek only to complement and enhance the traditional method.

The mathematical model of grey clustering is also presented [16]. The example is given of test score clustering from third grade students at a Changhua senior high girl school [17]. The main reason for this is to test the implementation and feasibility of the proposed toolbox. Also, C++ is utilised to develop the computer software toolbox required [18]. Past studies, such as Fuzzy, GA and Rough set, have all had their own toolboxes [19]. Software plays an important role in this system because it not only causes significant effects to the operation of the engineering system, but also makes the result on the real system more convincing and practical. Further, this is based on the software training of researchers at the University. In this article, C++ is used to create the toolbox. The main effect is that students understand the computer software more easily, as well as apply it to grey system theory. Also, past research in this field has focused on a main method to solve this kind of problem, namely packaged software like GM(1,1), grey entropy and *Matlab*/Mathematics toolbox [20-22]. Some advantages and disadvantages are presented and some suggestions towards further research are offered.

GREY CLUSTERING

Grey Whiteness Function

Assume $f(x)$ is a linear monotonic function of x, x is the grey number, and $f(x) \in [0,1]$. Then $f(x)$ is called the grey whiteness function, where $f_{max} = 1$, and divided into high, middle and low levels (see Figure 1). Based on grey system theory, the values in grey whiteness function are given objectively [16].
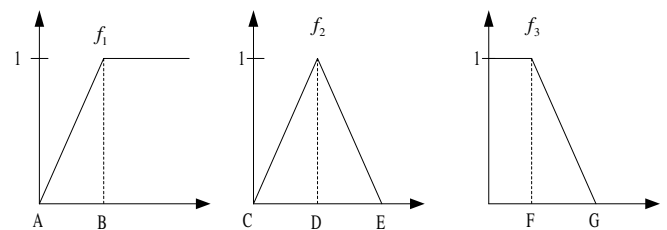

Figure 1: The whiteness function (high, middle and low).

The Definition of Grey Clustering

Definition 1: $r_1, r_2, r_3, \ldots, r_m$ are the clustering objects. $p_1, p_2, p_3, \ldots, p_n$ are the clustering indexes. $s_1, s_2, s_3, \ldots, s_l$ are the grey clustering, and d_{ij} are the sample values for the ith objects that correspond to the jth index, which is $1 \le i \le m$, $1 \le j \le n$.

Definition 2: D is the matrix, which contains d_{ij} elements:

$$D = \begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1n} \\ d_{21} & d_{22} & \ldots & d_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ d_{m1} & d_{m2} & \ldots & d_{mn} \end{bmatrix} \quad (1)$$

Definition 3: F is a mapping, and $op[f_k(d_{ij})]$ is the operation of d_{ij} corresponding to k clustering, which is $1 \le k \le l$, $1 \le i \le m$, $1 \le j \le n$, and F being called grey clustering.

Operational Steps of Grey Clustering

- Give the whiteness function f_1, f_2, f_3,…, f_l objectivity (see Figures 2-4).
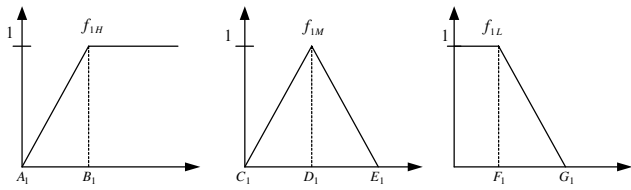

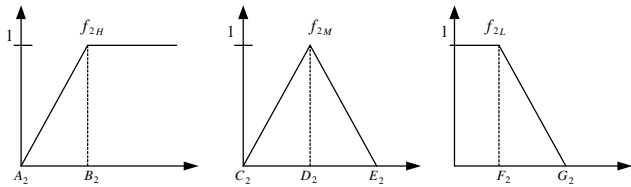Figure 2: The whiteness function for clustering 1.

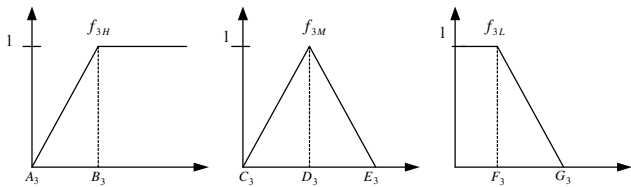Figure 3: The whiteness function for clustering 2.

Figure 4: The whiteness function for clustering 3.

- Calculate the values of ith clustering index for high level:

$$\eta_{1H} = \frac{B_1}{B_1 + B_2 + B_3 + ...}, \eta_{2H} = \frac{B_2}{B_1 + B_2 + B_3 + ...},$$
$$\eta_{3H} = \frac{B_3}{B_1 + B_2 + B_3 + ...},… \quad (2)$$

- Calculate the values of ith clustering index corresponding to η_{kH}, which $\sigma_{iH} = \sum_{k=1}^{l} f_{kH}(d_{ij})\eta_{kH}$, $1 \le k, j \le l$

$$\sigma_{1H} = f_{1H}(d_{11})\eta_{1H} + f_{2H}(d_{12})\eta_{2H} + f_{3H}(d_{13})\eta_{3H} + ... + f_{lH}(d_{1l})\eta_{lH}$$
$$\sigma_{2H} = f_{1H}(d_{21})\eta_{1H} + f_{2H}(d_{22})\eta_{2H} + f_{3H}(d_{23})\eta_{3H} + ... + f_{lH}(d_{2l})\eta_{lH}$$
$$\sigma_{3H} = f_{1H}(d_{31})\eta_{1H} + f_{2H}(d_{32})\eta_{2H} + f_{3H}(d_{33})\eta_{3H} + ... + f_{lH}(d_{3l})\eta_{lH} \quad (3)$$
$$………………………………………….$$
$$\sigma_{iH} = f_{1H}(d_{i1})\eta_{1H} + f_{2H}(d_{i2})\eta_{2H} + f_{3H}(d_{i3})\eta_{3H} + ... + f_{lH}(d_{il})\eta_{lH}$$

- Give a threshold value σ_{THH} objectively, and compare it with the values of σ_{kH} to determine the clustering level.
- Calculate the values of the ith clustering index for the middle level:

$$\eta_{1M} = \frac{D_1}{D_1 + D_2 + D_3 + ...}, \eta_{2M} = \frac{D_2}{D_1 + D_2 + D_3 + ...},$$
$$\eta_{3M} = \frac{D_3}{D_1 + D_2 + D_3 + ...},… \quad (4)$$

- Calculate the ith clustering index values that correspond to η_{kM}, in which $\sigma_{iH} = \sum_{k=1}^{l} f_{kM}(d_{ij})\eta_{kM}$ $1 \le k, j \le l$

$$\sigma_{1M} = f_{1M}(d_{11})\eta_{1M} + f_{2M}(d_{12})\eta_{2M} + f_{3M}(d_{13})\eta_{3M} + ... + f_{lM}(d_{1l})\eta_{lM}$$
$$\sigma_{2M} = f_{1M}(d_{21})\eta_{1M} + f_{2M}(d_{22})\eta_{2M} + f_{3M}(d_{23})\eta_{3M} + ... + f_{lM}(d_{2l})\eta_{lM}$$
$$\sigma_{3M} = f_{1M}(d_{31})\eta_{1M} + f_{2M}(d_{32})\eta_{2M} + f_{3M}(d_{33})\eta_{3M} + ... + f_{lM}(d_{3l})\eta_{lM} \quad (5)$$
$$………………………………………….$$
$$\sigma_{iM} = f_{1M}(d_{i1})\eta_{1M} + f_{2M}(d_{i2})\eta_{2M} + f_{3M}(d_{i3})\eta_{3M} + ... + f_{lM}(d_{il})\eta_{lM}$$

- Give a threshold value σ_{THM} objectively, and compare it with the values of σ_{kM} to decide the clustering level.
- Calculate the values of ith clustering for the low level:

$$\eta_{1L} = \frac{F_1}{F_1 + F_2 + F_3 + ...}, \eta_{2L} = \frac{F_2}{F_1 + F_2 + F_3 + ...},$$
$$\eta_{3L} = \frac{F_3}{F_1 + F_2 + F_3 + ...},… \quad (6)$$

- Calculate the ith clustering index values that correspond with η_{kL}, which $\sigma_{iL} = \sum_{k=1}^{l} f_{kL}(d_{ij})\eta_{kL}$, $1 \le k, j \le l$

$$\sigma_{1L} = f_{1L}(d_{11})\eta_{1L} + f_{2L}(d_{12})\eta_{2L} + f_{3L}(d_{13})\eta_{3L} + ... + f_{lL}(d_{1l})\eta_{lL}$$
$$\sigma_{2L} = f_{1L}(d_{21})\eta_{1L} + f_{2L}(d_{22})\eta_{2L} + f_{3L}(d_{23})\eta_{3L} + ... + f_{lL}(d_{2l})\eta_{lL}$$
$$\sigma_{3L} = f_{1L}(d_{31})\eta_{1L} + f_{2L}(d_{32})\eta_{2L} + f_{3L}(d_{33})\eta_{3L} + ... + f_{lL}(d_{3l})\eta_{lL} \quad (7)$$
$$………………………………………….$$
$$\sigma_{iL} = f_{1L}(d_{i1})\eta_{1L} + f_{2L}(d_{i2})\eta_{2L} + f_{3L}(d_{i3})\eta_{3L} + ... + f_{iL}(d_{il})\eta_{lL}$$

- Give a threshold value σ_{THL} objectively, and compare it with the values of σ_{kL} to determine the clustering level.

STUDENTS' TEST SCORES

The main reason for presenting this example is to implement the feasibility the toolbox and the new approach to grey clustering [17].

The pre-assumptions of the students' test score are as follows:

- The statistical objects are the third grade students (in the medicine and agriculture fields): class 302 (47students), class 303 (47students) and class 304 (46students), yielding a total of three classes and 140 students.
- The statistical indices are the average score of the 1st examination in three courses, namely: Chinese, English and mathematics, for the first semester of 2003.
- The statistical clustering is at the high, middle and low levels.

Based on the education concept, the grey whiteness function can be divided into three parts, as shown in Figure 5.
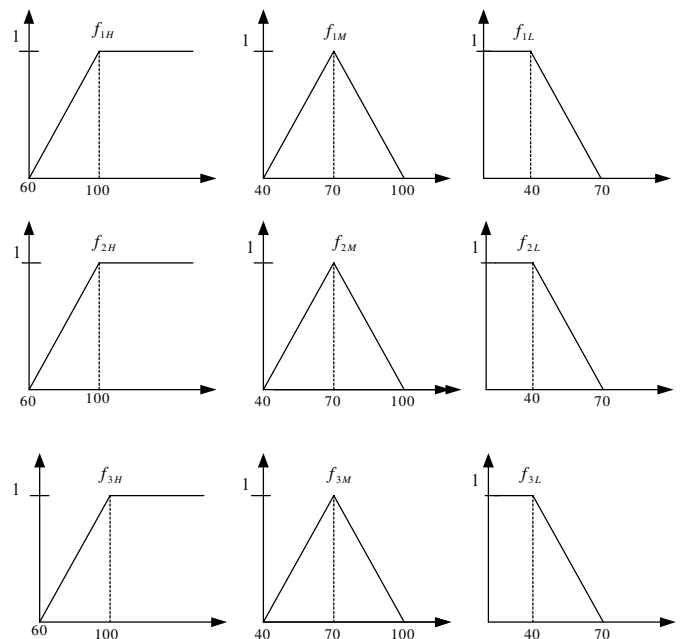

Figure 5: The whiteness function for the example.

112

The D Matrix

The D matrix of the examinations is $D = \begin{bmatrix} 70.1 & 62.5 & 76.6 \\ 69.3 & 59.1 & 51.4 \\ 67.9 & 60.9 & 56.5 \end{bmatrix}$

Finding the high level of clustering for class 302, class 303 and class 304 entails the following:

- Calculate the ith clustering index value at the high level:

$$\eta_{1H} = \frac{100}{100+100+100} = \frac{1}{3}, \eta_{2H} = \frac{100}{100+100+100} = \frac{1}{3},$$

$$\eta_{3H} = \frac{100}{100+100+100} = \frac{1}{3}$$

- Calculate the value of the ith clustering index that corresponds to the value of η_{kH}:

$$\sigma_{1H} = f_{1H}(70.1) \times \frac{1}{3} + f_{2H}(62.5) \times \frac{1}{3} + f_{3H}(76.6) \times \frac{1}{3}$$

$$= 0.2525 \times \frac{1}{3} + 0.0625 \times \frac{1}{3} + 0.415 \times \frac{1}{3} = 0.2433$$

$$\sigma_{2H} = f_{1H}(69.3) \times \frac{1}{3} + f_{2H}(59.1) \times \frac{1}{3} + f_{3H}(51.4) \times \frac{1}{3}$$

$$= 0.2325 \times \frac{1}{3} + 0.0000 \times \frac{1}{3} + 0.0000 \times \frac{1}{3} = 0.0775$$

$$\sigma_{3H} = f_{1H}(67.9) \times \frac{1}{3} + f_{2H}(60.9) \times \frac{1}{3} + f_{3H}(56.5) \times \frac{1}{3}$$

$$= 0.1975 \times \frac{1}{3} + 0.0225 \times \frac{1}{3} + 0.0000 \times \frac{1}{3} = 0.0733$$

- Set a threshold value $\sigma_{THH} = 0.5$, and compare it with the values of σ_{kH} to determine the clustering level:

$$\sigma_{1H} = 0.2433 < 0.5, \ \sigma_{2H} = 0.0775 < 0.5, \ \sigma_{3H} = 0.0733 < 0.5.$$

Hence, class 302, class 303 and class 304 are the same clustering, and are not high level.

Ascertaining the middle level of clustering for class 302, class 303 and class 304 comprises the following:

- Calculate the values of ith clustering index for the middle level:

$$\eta_{1M} = \frac{70}{70+70+70} = \frac{1}{3}, \eta_{2M} = \frac{70}{70+70+70} = \frac{1}{3},$$

$$\eta_{3M} = \frac{70}{70+70+70} = \frac{1}{3}$$

- Calculate the value of the ith clustering index that corresponds with the value of η_{kM}:

$$\sigma_{1M} = f_{1M}(70.1) \times \frac{1}{3} + f_{2M}(62.5) \times \frac{1}{3} + f_{3M}(76.6) \times \frac{1}{3}$$

$$= 0.9967 \times \frac{1}{3} + 0.7500 \times \frac{1}{3} + 0.7800 \times \frac{1}{3} = 0.8422$$

$$\sigma_{2M} = f_{1M}(69.3) \times \frac{1}{3} + f_{2M}(59.1) \times \frac{1}{3} + f_{3M}(51.4) \times \frac{1}{3}$$

$$= 0.9767 \times \frac{1}{3} + 0.6367 \times \frac{1}{3} + 0.3800 \times \frac{1}{3} = 0.6644$$

$$\sigma_{3M} = f_{1M}(67.9) \times \frac{1}{3} + f_{2M}(60.9) \times \frac{1}{3} + f_{3M}(56.6) \times \frac{1}{3}$$

$$= 0.9300 \times \frac{1}{3} + 0.6966 \times \frac{1}{3} + 0.5500 \times \frac{1}{3} = 0.7256$$

- Set a threshold value $\sigma_{THM} = 0.5$, and compare it with the values of σ_{kM}, to determine the clustering level:

$$\sigma_{1M} = 0.8422 > 0.5, \ \sigma_{2M} = 0.6644 > 0.5, \ \sigma_{3M} = 0.7256 > 0.5.$$

Hence, class 302, class 303 and class 304 are the same clustering, and are all middle level.

Finding the low level of clustering for class 302, class 303 and class 304 involves the following:

- Calculate the values of the ith clustering for the low level:

$$\eta_{1L} = \frac{40}{40+40+40} = \frac{1}{3}, \eta_{2L} = \frac{40}{40+40+40} = \frac{1}{3},$$

$$\eta_{3L} = \frac{40}{40+40+40} = \frac{1}{3}$$

- Calculate the value of the ith clustering index that corresponds to the value of η_{kL}:

$$\sigma_{1L} = f_{1L}(70.1) \times \frac{1}{3} + f_{2L}(62.5) \times \frac{1}{3} + f_{3L}(76.6) \times \frac{1}{3}$$

$$= 0.0000 \times \frac{1}{3} + 0.2500 \times \frac{1}{3} + 0.0000 \times \frac{1}{3} = 0.0833$$

$$\sigma_{2L} = f_{1L}(69.3) \times \frac{1}{3} + f_{2L}(59.1) \times \frac{1}{3} + f_{3L}(51.4) \times \frac{1}{3}$$

$$= 0.0233 \times \frac{1}{3} + 0.3633 \times \frac{1}{3} + 0.6200 \times \frac{1}{3} = 0.3356$$

$$\sigma_{3L} = f_{1L}(67.9) \times \frac{1}{3} + f_{2L}(60.9) \times \frac{1}{3} + f_{3L}(56.6) \times \frac{1}{3}$$

$$= 0.0700 \times \frac{1}{3} + 0.3033 \times \frac{1}{3} + 0.4500 \times \frac{1}{3} = 0.2744$$

- Set a threshold value $\sigma_{THL} = 0.5$, and compare it with the values of σ_{kL} to determine the clustering level:

$$\sigma_{1L} = 0.0833 < 0.5, \ \sigma_{2L} = 0.3356 < 0.5, \ \sigma_{3L} = 0.2744 < 0.5.$$

Hence, class 302, class 303 and class 304 are the same clustering, but are not low level

THE GREY CLUSTERING TOOLBOX

The input/output interface in the grey clustering toolbox is based on a Microsoft-style structure. According to the characteristics of the grey statistics, the multi-page, multi-function and menu-bar toolbox have been created to make this toolbox more user friendly. This is the first major characteristic of the toolbox. In grey statistics and grey clustering, when the number of the sequence *m* is large, the operation time will increase very quickly. If the traditional method is used to find the whole corresponding value, it will waste a lot of time. Therefore, in order to increase the operational speed and decrease calculation errors, a look-up table is utilised as a mathematical method to find the accurate value; this is the second major characteristic of the toolbox. Also, in the *history page*, the entire processing of the operation will be recorded; the main purpose for this is not only modification for the user, but also to make the results on the analysis system more convincing and practical; this function is the third major characteristic of the toolbox.

Software requirements cover Windows 2000 or an upgraded version, as well as a screen resolution of at least 1024×768.

The operation of the grey clustering toolbox is as follows:

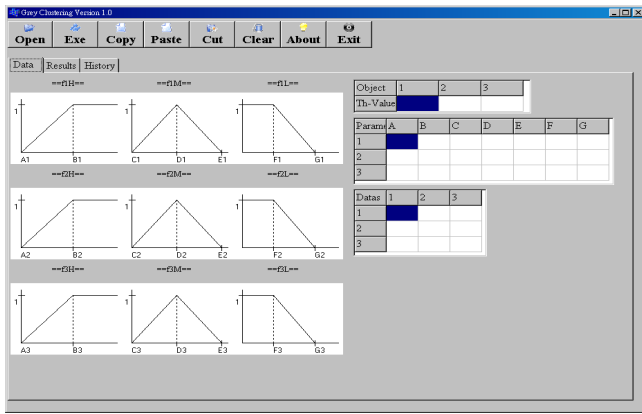- Start the toolbox:

    - Main screen (Figure 6);

Figure 6:The main screen of grey clustering.

- Open the analysis data (*.txt type) (Figure 7);

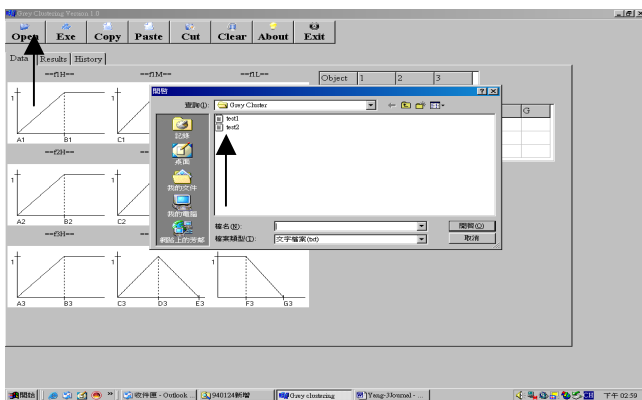Figure 7: Open the data (*.txt type).
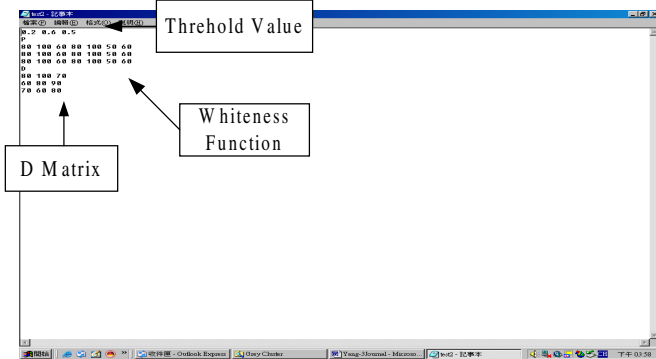
- Format of the data (Figure 8);

Figure 8: Format of the data (*.txt type).

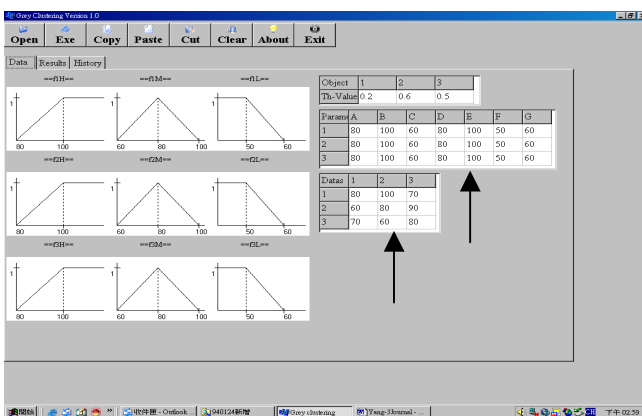- Data on the screen (Figure 9);

Figure 9: The analysis data on the screen.

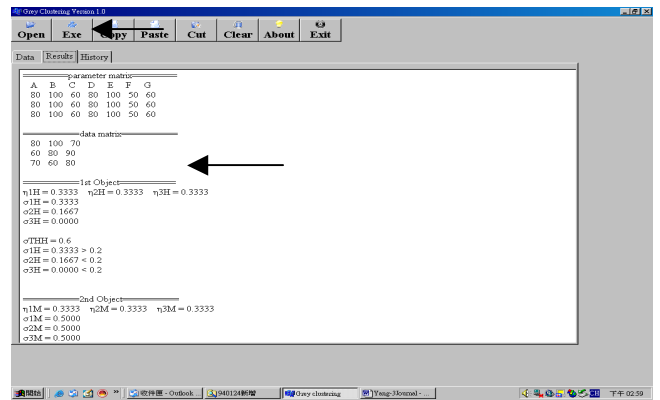- Click the *Exe* icon to execute the toolbox (Figures 10-11);
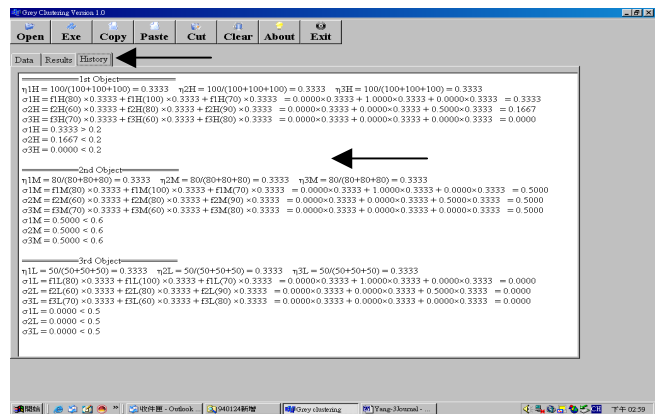

Figure 10: The execution results.

Figure 11: The history of processing.

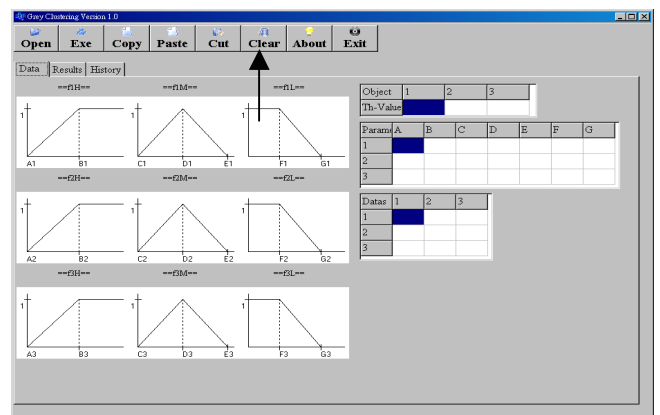- Click the *Clear* icon to clear the data (Figure 12);

Figure 12: Click the *Clear* icon to clear whole data.

- The others functions are the same as for Windows (ie Copy, Cut, Paste and Exit).

CONCLUSIONS

In previous development of grey system theory software design, the focus was on a case study. However, grey system theory often contains complex operations and plenty of graphic demonstrations; this poses certain difficulties for software design. Therefore, in view of grey systems, the C^{++} language has been utilised to design a grey system toolbox. It is not only original in grey system theory, but it also has the following academic implications:

- The model facilitates a thorough understanding of the GM model and the essential meanings involved in grey clustering, especially with regard to the original grey clustering toolbox involving grey system theory.
- The writing is easy and, in fact, the grammar of the C^{++} language does not present any difficulties, as it is portable. In addition, it is user-friendly for human interface.
- It also extends to economic benefits. One ordinary set of C^{++} is priced at approximately \$US70 (in 2005), making it very inexpensive, and the total required memory space used is small (740kB).
- Furthermore, the value of the grey whiteness function is given objectively; this is the same as the fuzzy membership function. However, the most obvious difference is that the value of the grey whiteness function can be allocated arbitrarily, and the value of the fuzzy membership must be given a cardinal.

If the traditional method is used for the example presented in this article, only the rank of each score and the grade can be obtained; it cannot give the required detail.

Table 1 summarises the analysis of the results of the first examination using the traditional summation method.

Table 1: Analysis of the results of the first examination using the traditional summation method.

Course	The First Examination	Grade
Chinese	$\dfrac{(70.1+69.3+67.9)}{3}=69.10$	C
English	$\dfrac{(62.5+59.1+60.9)}{3}=60.83$	C
Mathematics	$\dfrac{(55.7+51.4+56.5)}{3}=54.53$	D

To sum up, a grey clustering toolbox has been created that presents serves to elaborate on grey system theory. This study presents both the theoretical and practical significances. However, several weaknesses still exist, namely:

- The combination of the *Matlab* and mathematics toolbox with the aforementioned approach;
- A comparison of the different points of the membership function in the fuzzy system;
- In the proposed toolbox, the whiteness function only has three levels, hence, the level can be extended in the future.

ACKNOWLEDGEMENTS

REFERENCES

1. Fang, P., Wu, G.P. and Xu, Z.G., Grey clustering with programming: assessing the caprock of oil and gas. *J. of Grey System*, 12, **2**, 117-120 (2000).
2. Wu, G.P., Zhang, X.H. and Xu, Z.X., Grey ART clustering for oil and gas. *J. of Grey System*, 12, **2**, 199-201 (2000).
3. Wu, G.P., Zhang, X.H. and Xu, Z.X., Line-Point grey clustering for blind oil-gas trap. *J. of Grey System*, 12, **3**, 293-298 (2000).
4. Chen, X.Y. and Chen, Z.X., Grey clustering the sportswear style. *J. of Grey System*, 12, **4**, 409-414 (2000).
5. Jing, G.X., Yao, R., Zhang, F.R. and Leng, X.Y., Grey clustering the mine intrinsic fire. *J. of Grey System*, 13, **4**, 403-406 (2001).
6. Cai, H.Q. and Li, Z.C., Assessment of management stuff via grey clustering. *J. of Grey System*, 14, **1**, 97-100 (2002).
7. Wan, K., Li, C. and Zhang, J.L., Grey clustering the high technology competence. *J. of Grey System*, 14, **2**, 209-212 (2002).
8. Xie, J.P., Chen, R.Q. and Kong, L.C., Grey assessing ERP software via grey clustering. *J. of Grey System*, 15, **1**, 87-92 (2003).
9. Ma, B.G., Li, L. and Buo, C.F., Grey clustering of air materiel cannibalization feasibility. *J. of Grey System*, 15, **2**, 181-184 (2003).
10. Su, X.B., Liao, J.Q. and Gong, Y.M., Grey clustering analyzing the monger's work stress in state-owned enterprises. *J. of Grey System*, 15, **3**, 279-282 (2003).
11. Zhao, C.L. and Li, Z.C., Application of grey clustering to cash crops distribution. *J. of Grey Sy*stem, 15, **3**, 379-382 (2003).
12. Wang, Y.Z. and Lee, F.M., A study on the relationship between schooling performance and occupational classification of the graduates for the vocational schools. *J. of the Chinese Grey System*, 4, **2**, 119-130 (2001).
13. Yeh, M.F., Chang, J.C. and Lu, H.C., Unsupervised clustering algorithm via grey relational pattern analysis. *J. of The Chinese Grey System*, 5, **1**, 16-22 (2002).
14. Lam, K.T., A grey clustering approach for machine cell formation. *J. of the Chinese Grey System*, 5, **2**, 107-116 (2002).
15. Ken, M.L. and Lee, Y.B., The pattern identification via grey clustering method. *J. of Grey System*, 7, **2**, 83-90 (2004).
16. Wen, K.L. and Huang, Y.F., The development of grey statistic toolbox and its application in the clustering of student's test score. *J. of Quantitative Management*, 1, **2**, 219-238 (2004).
17. National Changhua Girl Senior High School, The Examination Score of The Third Grade. Changhua: National Changhua Girl Senior High School (2003).
18. Chiang, C.H., *The Art of C^{++} Program Design.* Taipei: CWC Publishers (2004).
19. Wen, K.L. and Huang, Y.F., The development of grey Verhulst toolbox and the analysis of population saturation state in Taiwan-Fukien. *Proc. IEEE SMC'2004 Conf.*, Hague, The Netherlands, 5007-5012 (2004).
20. Huang, Y.F., Wang, L.K. and Wen, K.L., GM(1,1) toolbox for engineering. *Proc. 8th National Conf. of Grey Theory and Applications*, Taipei, Taiwan, C017 (2003).
21. Wen, K.L., Huang, Y.F. and You, M.L., Grey entropy toolbox in weighting analysis. *Proc. 8th National Conf. of Grey Theory and Applications*, Taipei, Taiwan, C056, (2003).
22. Wen, K.L. and Chang, T.C., The research and development of completed GM(1,1) model toolbox using *Matlab. Inter. J. of Computational Cognition*, 3, **2**, 41-47 (2005).