Grey Systems, Grey Models and Their Roles in Data Analytics

Yingjie Yang

School of Computer Science and Informatics
De Montfort University
Leicester, UK
yyang@dmu.ac.uk

Sifeng Liu

Institute for Grey Systems Studies
Nanjing University of Aeronautics and Astronautics
Nanjing, P.R. China
sfliu@nuaa.edu.cn

***Abstract* - This paper gives an introduction to the basic concepts of grey systems, grey numbers and grey models, and discusses their roles in data analysis of data science. The necessity of small data models is demonstrated, and their complementary functions to Big Data models are investigated. Based on these investigations, a novel framework for combining Big Data models with grey models is proposed.**

Keywords - small data; grey systems; grey numbers; grey models

I. INTRODUCTION

With the increased availability of data from the Internet and Internet of Things, data analytics targeting on Big Data is becoming the major focus in data science. The highlighted topics are those related with large volume, high velocity and variety [1]. Many companies are establishing their data analytics teams focusing mainly on Big Data and related data analytics, and small data analytics is largely ignored. However, the coming of the Big Data era does not necessarily bring an end to small data problems, and there are still situations where Big Data cannot solve the problems [2]–[4]. For example, traffic volume can be recorded as a Big Data volume and good predictions can be easily achieved by Big Data models for average estimation in a year or a month in a large area. However, an hourly or even finer prediction for a specific location will be difficult for a Big Data model as it is more related with the local and recent data for that period and location. In that case, a small data model may perform better.

In real world, there are situations where data could be extremely limited and uncertain. For example, yearly economic evaluation may have to rely on a few data samples only (one for each year), and aggregation from Big Data can also lead to extreme limited data. In this case, Big Data technology cannot help, and neither can the traditional statistics due to the extremely low cardinality of data samples. However, the grey models in grey systems [5]–[8] are exactly designed for these situations. In grey systems, systems are classified into three different categories: white systems where everything is known, black systems where nothing is known and grey systems where only partial information is available. Due to the partial information, grey models consider the available information as a partial reflection of the ground truth, hence a series of grey operators are defined to maximise the usage of the available information and minimise the impact of this unknown information. In this way, grey models have the ability to establish a feasible model using an extremely limited number of data samples. It is obvious that grey models provide a complementary capability which Big Data models do not have. In real world applications, it is possible that both Big Data models and small data models could prove useful in covering a wider scope of applications in different scales as opposed to their competitors.

However, so far, grey models and Big Data models are completely separated from each other. Researchers in the two different models are working in isolation and have not yet realised their complementary roles to each other. In this paper, we will try to draw a link between the two different models so as to connect them together to provide a more robust data analytical tool for data science.

II. GREY SYSTEMS AND GREY NUMBERS

Grey systems were firstly proposed by Professor J. Deng in 1982 [9]. As aforementioned, the information is classified into three categories: white with completely certain information, grey with insufficient information, and black with totally unknown information. Grey systems are concerned with, in particular, the information belonging to the grey category. Because of insufficient information, most of the statistical characteristics of the system may not be clearly identified. However, the data available may reveal the range of information. We now provide some basic definitions.

Definition 1 (Grey numbers [10]): Let Ω ⊂ R be the universe, g± ∈ Ω be an unknown real number within a union set of closed or open intervals

$$g^{\pm} \in \bigcup_{i=1}^{n} [a_i^-, a_i^+]$$

$i = 1, 2, \cdots, n$, n is an integer and $0 < n < \infty$, a_i^-, $a_i^+ \in \Omega$ and $a_{i-1}^+ < a_i^- < a_i^+ < a_{i+1}^-$. For any interval $[a_i^-, a_i^+] \subseteq \bigcup_{i=1}^{n}[a_i^-, a_i^+] \subseteq \Omega$, p_i is the probability for $g^{\pm} \in [a_i^-, a_i^+]$. If the following conditions hold

- $p_i > 0$
- $\sum_{i=1}^{n} p_i = 1$

Then we call $g^{\pm}$ a grey number. $g^- = \inf a_i^-$ and $g^+ = \sup a_i^+$ are called the lower and upper limits of $g^{\pm}$.

If $g^- = g^+$, $g^{\pm}$ has no uncertainty at all and is called a white number; on the contrary, if $\bigcup_{i=1}^{n}[a_i^-, a_i^+] = \Omega$, there

is nothing known about $g^{\pm}$ and it is called a black number.

The degree of greyness of a grey number measures the significance of uncertainty in a grey number. For example, three different definitions for the degree of greyness of a generalised grey number have been proposed [7], [10], [11].

Definition 2 (Degree of greyness of a grey number [10]):

Let $\Omega \subset R$ be the universe and $g^{\pm} \in \bigcup_{i=1}^{n}[a_i^-, a_i^+] \subseteq \Omega$, $d_{min}, d_{max} \in \Omega$ are the minimum and maximum values of Ω. μ is a measurement defined on Ω. The degree of greyness of $g^{\pm}$ is defined as

$$g^{\circ}(g^{\pm}) = \frac{\mu(g^{\pm})}{\mu(\Omega)} \tag{1}$$

Similar to grey numbers, we can classify sets into three different categories [10]:

Definition 3 (White sets [10]): For a set A ⊆ U, if its characteristic function value of each x with respect to A can be expressed with a single white number $v \in [0; 1]$:

$$\chi_A : U \rightarrow [0, 1] \tag{2}$$

Now A can be considered a white set.

In fact, a type-1 fuzzy set can be considered as a special case of a white set. A crisp set is clearly a white set and it is not fuzzy at all, but a type-1 fuzzy set is still a white set although it is fuzzy compared with a crisp set in that it has a single white number as its characteristic function value.

Definition 4 (Black sets [10]): For a set A ⊆ U, if its characteristic function value of each x with respect to A can be expressed with a black number, then A is a black set. An object contained within a black set has a completely unknown characteristic function value, and it is opposite to a white set where we have complete knowledge about the characteristic function value. Between the two extremes, a set with incomplete information about its characteristic function values is defined as a grey set [10]:

Definition 5 (Grey sets [10]): For a set A ⊆ U, if the characteristic function value of x with respect to A can be

expressed with a grey number $g_A^{\pm}(x) \in \bigcup_{i=1}^{n}[a_i^-, a_i^+] \in D[0,1]$:

$$\chi_A : U \rightarrow D[0, 1]^{\pm}$$

then A is a grey set.

Here, D[0; 1]± refers to the set of all grey numbers within the interval [0,1]. Similar to the expression of a fuzzy set, a grey set A is represented with its relevant objects and their associated grey numbers for its characteristic function:

$$A = g^{\pm}(x_1)/x_1 + g^{\pm}(x_2)/x_2 + \ldots + g^{\pm}(x_n)/x_n$$

Because of the existence of grey and black objects, the relationships between some objects and a grey set may not be completely known. As a result, the value for its corresponding characteristic function can only be expressed as a grey number. This is due to the incomplete information of this object. Similar to the case for a grey number, the uncertainty caused by the information incompleteness can be measured using a degree of greyness [10].

Definition 6 (Degree of greyness for an object [10]): Let U be the finite universe of discourse, x ∈ U. For a grey set

A ⊆ U, the characteristic function value of x with respect

to A is $g_A^{\pm}(x) \in D[0,1]^{\pm}$. The degree of greyness $g_A^{\pm}(x)$ of object x for set A is expressed as

$$g_A^{\pm}(x) = |g^+ - g^-| \tag{3}$$

Based on the degree of greyness for an object, a degree of greyness for a set is defined as follows:

Definition 7 (Degree of greyness for a set [10]): Let U be the finite universe of discourse, and let A be a grey set and

A ⊆ U. Assume x_i is an object relevant to A and $x_i \in U$. Let i = 1, 2, 3, …, n and n is the cardinality of U. The degree of greyness of set A is defined as:

$$g_A^{\circ} = \frac{\sum_{i=1}^{n} g_A^{\circ}(x_i)}{n} \tag{4}$$

According to the given definition, the uncertainty caused by incomplete information for the evaluation of students under different attributes can be measured using the degree of greyness for objects and sets.

III. GREY MODELS

A grey number is represented as a set of possible values coming up as the images of its instances. In system evolution, however, the image is usually materialised as a single value coming from the set of possible values. Such an image can be considered as a combination of the underlined number and its uncertain noise. The aim of a grey model is to capture the relationship between a new image value of a grey number and its previous (historical) records in the form of a sequence of image values. Due to the involvement of the uncertain noise and the limited number of data samples, it is usually difficult to establish a meaningful model using other algorithms, e.g. statistics. In grey systems, however, a number of grey operators are applied to reveal the ground truth values of the involved grey numbers so as to establish a meaningful model to give a more reliable prediction. Here, we will illustrate this using the basic GM(1,1) model, which is the foundation for most other extended grey models [12], [13].

Consider a sequence

$$X^{(0)} = (x^{(0)}(1), x^{(0)}(2), \ldots, x^{(0)}(n))$$

The curve ab demonstrated in Figure 1 does not show any obviously identifiable pattern in the data. Therefore, it is difficult to setup a feasible model for the sequence. As aforementioned, each value in the sequence can be considered as a combination of the ground truth value together with a noise component, hence

$$x^{(0)}(i) = x^{(0)}(i) + \Delta(i)$$

In grey systems, there are a series of weakening operators to weaken the impact of these noise components. The 1-AGO (once accumulating generation operator) of $X^{(0)}$ is one of the simplest weakening operators:

$$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i); k = 1, 2, \ldots, n$$

The resulting sequence

$$X^{(1)} = (x^{(1)}(1), x^{(1)}(2), \ldots, x^{(1)}(n))$$

displays a clear growing tendency as shown by the curve ac in Figure 1.

Considering the random feature of the noise components, we have

$$\lim_{k\to\infty}\sum_{i=1}^{k} x^{(0)}(i) = \lim_{k\to\infty}\sum_{i=1}^{k} x_i^{(0)}(i)$$

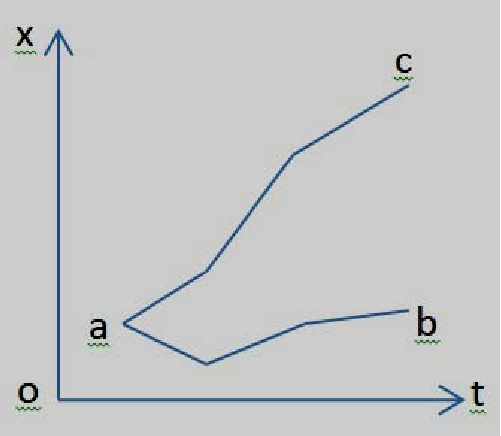

Fig. 1. Impact of grey operators

Obviously, when k is big enough or the 1-AGO opertor is applied many times, the resulting sequence approximates to the real sum values of the sequences with less impact from the noise. It shows that the 1-AGO operator is effective in weakening the impact of noise.

Now, we derive another sequence from $X^{(0)}$:

$$Z^{(1)} = (-, z^{(1)}(2), \ldots, z^{(1)}(n))$$

where

$$z^{(1)}(k) = \frac{1}{2}\left(x^{(1)}(k) + x^{(1)}(k-1)\right); k = 2, 3, \ldots, n$$

Then, we have the first order and single variable grey forecasting model, abbreviated as GM(1,1):

$$x^{(0)}(k) + az^{(1)}(k) = b$$

The parameter *a* and *b* can be calculated using the least square method

$$[a \quad b]^T = (B^T B)^{-1} BY$$

where

$$B = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \cdots\cdots & \cdots \\ -z^{(1)}(n) & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} -x^{(1)}(2) & 1 \\ -x^{(1)}(3) & 1 \\ \cdots\cdots & \cdots \\ -x^{(1)}(n) & 1 \end{bmatrix}$$

Then, a time response equation of GM(1,1) model can be established as

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{b}{a}\right)e^{-ak} + \frac{b}{a}, \quad k=1, 2, \ldots, n$$

With this equation, a reverse operation will give the prediction values for the original sequence.

The GM(1,1) is the simplest grey model, and it has been extended into many different more complicated models. Limited by the space, we would not cover other grey models here.

IV. THE ROLE OF GREY MODELS IN DATA ANALYSIS

The last section has demonstrated the basic grey model GM(1,1), and it has been shown that a grey model applies weakening operators to reduce the impact of noise in the original sequence and then establish data models with less impact from the noise components. In this way, it will establish a feasible model with only a few data samples. It greatly reduces the required quantity of data samples and makes it possible to establish feasible models using small data. This is a distinctive advantage in comparison with most other data analytic models. Although statistical models

have a long history in solving small data problems, their requirement on data sample size is still much higher than grey models. The popular Big Data models, such as neural networks, require much more data to make sense. Therefore, these models will not work where grey models work. In this sense, grey models provide a beneficial complementary tool for data analytics.

However, the applications of grey models are mostly isolated from the popular Big Data analysis. Many people consider it as an exclusive opposite to Big Data technology. The fact of their complimentary roles with each other received little attention. Here we will examine the possibility to bring the two models together and make use of their merits for different data.

A. Uncertainty measurement

In the current Big Data research, most investment has gone to the first three Vs: Volume, Velocity and Variety. Significant progress has been made in these areas as well. However, the fourth V: Veracity has not received the attention it deserves. The increase in data volume does not necessarily reduce uncertainties in Big Data; on the contrary, it more likely to increase uncertainty as a result of its variety. Big Data is usually collected from multiple sources where each source has its own interpretation and data accuracy. Due to the variety of the data sources, required attributes in one source may not be recorded in another source, which will certainly lead to more missing values and incomplete information. The mixed accuracies will also certainly bring in wider fluctuations in data accuracy and reliability. All these will raise incomplete data which may have significant impact on the reliability of the data analysis result.

As demonstrated in section II, grey numbers, grey sets and their associated degree of greyness can be applied to quantify the information incompleteness in data sets. Furthermore, extended incompleteness measurements have also been defined for grey numbers and grey sets [14]. All these provide a foundation to measure the information incompleteness in Big Data. We can consider the whole data set as a grey set and each sample as a grey element of the set. A degree of greyness will then be derived for each sample and a corresponding degree of greyness can thus be calculated for the whole data set following the equations in Definition 6 and 7. In this way, the greyness of Big Data can be quantified through grey systems as shown in Figure 2.

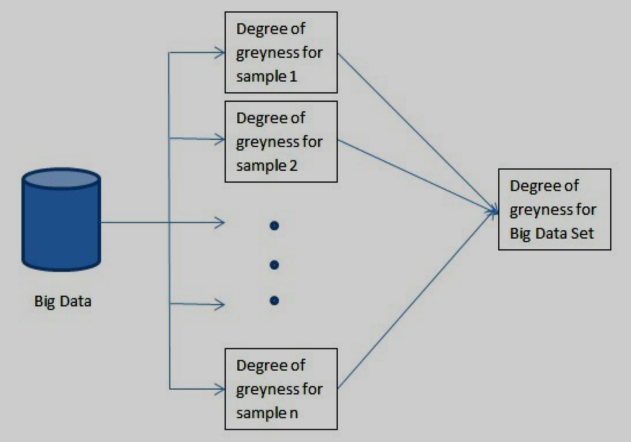
Degree of greyness for sample 1
Degree of greyness for sample 2
Big Data
Degree of greyness for sample n
Degree of greyness for Big Data Set

Fig. 2. Greyness of Big Data

Obviously, if the degree of greyness is too large, the data analysis derived from such Big Data may give misleading results, which will damage the reputation of Big Data and reduce people's confidence on Big Data technology. Therefore, grey systems provide a useful tool in quantifying the incompleteness in Big Data.

B. Full spectrum data analysis

When Big Data are collected from long term recording and large areas, the general trends will be easy to capture through Big Data analysis. For example, yearly sales and large groups with common shopping interests will be easily to be identified in Big Data. However, the general trends are too general to be extended to individual shops or customers, and the daily sales at a specific shop for specific group of people at a specific time slot would be difficult to predict through Big Data technology. The probability obtained from large populations may not be a true reflection of a few specific individuals. Under such a situation, a grey model has the potential to be much superior to the Big Data model. Grey models focus on limited recent data for specific local objects, it has fewer opportunities to be drowned in the Big

Data ocean where remote and irrelevant information can change the prediction.

However, in comparison with Big Data models, grey models focus on limited recent data for specific objects only, and thus it is not good at identifying general trends and specific areas to be further investigated from a large volume of data. Therefore, its role is complementary to Big Data models rather than competition. They have merits for different situations and hence can be connected to carry out more reliable data analysis. Models for Big Data can be applied as a general model to identify general trends and the specific location to be highlighted, while grey models could be called in, using the local short term data to derive local short term prediction. In this way, we can make use of their merits and connect the two models as illustrated in Figure 3.

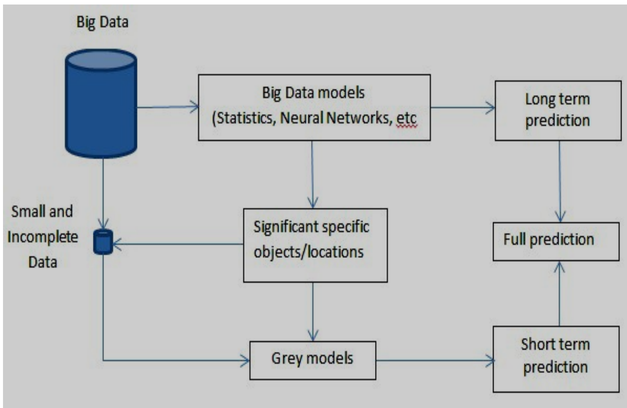
Big Data
Big Data models (Statistics, Neural Networks, etc
Long term prediction
Small and Incomplete Data
Significant specific objects/locations
Full prediction
Grey models
Short term prediction

Fig. 3. Grey models in data analysis for prediction

In Figure 3, Big Data models (e.g. Statistics, Neural Networks, etc.) will firstly be employed to extract long term and general trends/pattern in data. The results are the identified general pattern meaningful when large volumes of data are considered together. This general pattern does not necessarily match the fact data in a specific location for a specific time slot. However, such an analysis may reveal that these significant locations (objects) need to be further investigated. Based on the identified specific locations (time slots), small data sets corresponding to these identified specific location/time slots will be extracted from the Big Data. Then grey models will be applied to conduct short term prediction for specific locations (objects). The full prediction will be a combination of the general long terms trends and the short term forecasting for specific locations/objects. In this way, the two unrelated models can be connected together so as to establish a more informative full spectrum data analysis and prediction.

V. THE ROLE OF GREY SYSTEMS IN OTHER ASPECTS OF DATA ANALYTICS

Section IV discusses the role of grey models in data analysis. However, the role of grey systems is not only restricted within data analysis; it has a potential impact on other aspects of data analytics as well. Figure 4 demonstrates a simple data analytics profile from the point of view of grey systems. On the left side, the grey sources represent various data sources from other business sectors. Their common feature is that they are not perfect, and there is partially unknown or missing data in additional to their different accuracies and interpretations. These data sets are collected together and form the grey data which is a combination of data with different partial information, accuracies and interpretations. The data then undergoes a profiling/cleaning process where part of the partial information are whitenised, while other data is quantified with their greyness. The result is then stored into grey storage where greyness in data hierarchy is taken as an important factor in their distributed storage. The stored data is then retrieved to carry out data mining and grey analysis as illustrated in section IV. The result will be fed into the grey decision making process as business intelligence.

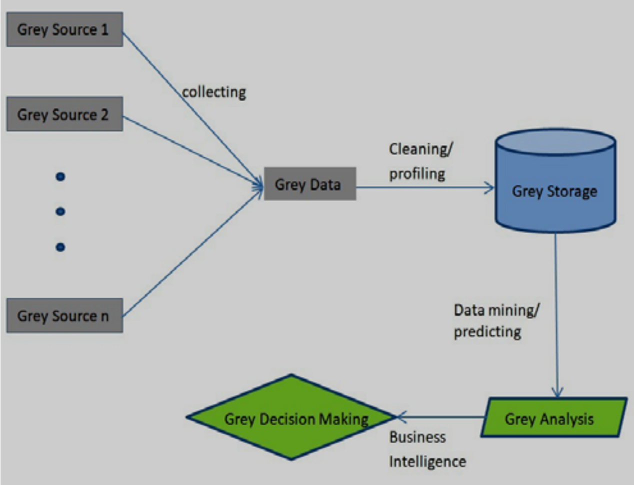

Fig. 4. Grey systems in data analytics

VI. CONCLUSIONS

Based on the review of the special features of grey systems and grey models, this paper identifies the role of grey systems in data analytics. A novel combination architecture is proposed to connect grey models with Big Data models. Grey models have some distinctive features which Big Data models do not have, while Big Data models cover areas which grey models are incapable to deal with. Therefore, a combination of grey models with Big data models can provide a more robust data analysis framework. In the same time, grey numbers and grey sets can be applied to quantify greyness (incompleteness) of Big Data so as to avoid useless data analysis operations. Furthermore, the potential grey analytics is also discussed. Our analysis shows that grey systems are actually beneficial to Big Data analysis in data science.

REFERENCES

[1] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, “Big data analytics: a survey”, Journal of Big Data, vol. 2, no. 1, pp. 1–32, 2015.

[2] O. Kennedy, D. R. Hipp, S. Idreos, A. Marian, A. Nandi, C. Troncoso, and E. Wu, “Small data” in Proceedings of the 2017 IEEE 33rd International Conference on Data Engineering, 2017, pp. 1475–1476.

[3] I. Martin-Diaz, D. Morinigo-Sotelo, O. Duque-Perez, and R. de J. Romero-Troncoso, “Early fault detection in induction motors using adaboost with imbalanced small data and optimized sampling”, IEEE Transactions on Industry Applications, vol. 53, no. 3, pp. 3066 – 3075, 2017.

[4] M. Thinyane, “Small data and sustainable development individuals at the center of data-driven societies” in Proceedings of the 2017 IEEE ITU Kaleidoscope: Challenges for a Data-Driven Society (ITU K), 2017.

[5] S. Liu, Y. Yang, , and J. Forest, Grey Data Analysis: Methods, Models and Applications. Springer-Verlag, 2016.

[6] Y. Lin, M. Chen, and S. Liu, “Theory of grey systems: Capturing uncertainties of grey information”, kybernetics: The International Journal of Systems and Cybernetics, vol. 33, pp. 196–218, 2004.

[7] S. Liu, T. Gao, and Y. Dang, Grey systems theory and its applications. Beijing: The Science Press of China, 2000.

[8] S. Liu and Y. Lin, Grey Information Theory and Practical Applications. Springer, 2006.

[9] J. Deng, “The control problems of grey systems”, Systems and Control Letters, 1982.

[10] Y. Yang and R. John, “Grey sets and greyness”, Information Sciences, vol. 185, no. 1, pp. 249–264, 2012.

[11] S. Liu, Z. Fang, Y. Yang, and J. Forrest, “General grey numbers and their operations”, Grey Systems: Theory and Application, vol. 2, no. 3, pp. 341–349, 2012.

[12] S. Liu and Y. Yang, “Explanation of terms of grey forecasting models”, Grey Systems: Theory and Application, vol. 7, no. 1, pp. 123–128, 2017.

[13] L. Wu, S. Liu, Y. Yang, L. Ma, and H. Liu, “Multi-variable weakening buffer operator and its application”, Information Sciences, vol. 339, pp. 98–107, 2016.

[14] Y. Yang, S. Liu, and R.John, “Uncertainty representation of grey numbers and grey sets”, IEEE Transaction on Cybernetics, vol. 44, no. 9, pp. 1508–1517, 2014.