Uncertainty and grey data analytics

Yingjie Yang
School of Computer Science and Informatics, De Montfort University – City Campus, Leicester, UK, and
Sifeng Liu and Naiming Xie
Nanjing University of Aeronautics and Astronautics, Nanjing, China

Received 13 August 2019
Accepted 16 September 2019

Abstract

Purpose – The purpose of this paper is to propose a framework for data analytics where everything is grey in nature and the associated uncertainty is considered as an essential part in data collection, profiling, imputation, analysis and decision making.
Design/methodology/approach – A comparative study is conducted between the available uncertainty models and the feasibility of grey systems is highlighted. Furthermore, a general framework for the integration of grey systems and grey sets into data analytics is proposed.
Findings – Grey systems and grey sets are useful not only for small data, but also big data as well. It is complementary to other models and can play a significant role in data analytics.
Research limitations/implications – The proposed framework brings a radical change in data analytics. It may bring a fundamental change in our way to deal with uncertainties.
Practical implications – The proposed model has the potential to avoid the mistake from a misleading data imputation.
Social implications – The proposed model takes the philosophy of grey systems in recognising the limitation of our knowledge which has significant implications in our way to deal with our social life and relations.
Originality/value – This is the first time that the whole data analytics is considered from the point of view of grey systems.

Keywords Uncertainty, Data incompleteness, Grey data analysis, Grey data analytics, Grey data collection
Paper type Research paper

1. Introduction

With the rapid development of big data technology and artificial intelligence (AI), we are increasingly living in a data-driven world where machine intelligence is assisting us in many parts of our life, such as journey planning, automatic driving, security control and health care service. As a result, our society is increasingly depending on the data we collected, and the quality of the data is determining the quality of our life. Although veracity has been recognised as another feature of big data in addition to its volume, velocity and variety (Wang and He, 2016), the significance of uncertainties has not received sufficient attention in the current big data oriented data analytics research and applications.

Big data sets usually assemble data from different sources with different accuracies and reliability, and the associated uncertainties can be even worse. The increased volume helps to deal with uncertainties like noise in defining probability distribution in many cases, but it does nothing with uncertainties like incompleteness and inconsistency. The variety and its associated diversity of data sources and interpretations can easily lead to wider fluctuations

This research was funded by National Natural Science Foundation of China and Royal Society, UK.

Marine Economics and Management
Vol. 2 No. 2, 2019
pp. 73-86
Emerald Publishing Limited
2516-158X
DOI 10.1108/MAEM-08-2019-0006

and more significant uncertainties. As a result, the increased data volume might introduce even more uncertainties. Hariri *et al.* (2019) have pointed out in their recent paper that "little work has been done in the field of uncertainty when applied to big data analytics as well as in the artificial intelligence techniques applied to the datasets".

Although the current research applications are heavily focussing on big data, small data does not disappear at all, and it is still a major challenge in many real-world data analytics. In some areas, data collection is expensive, and it is not realistic to carry out large number of costly experiments to collect data, such as the geological sampling in oil industry. In some cases, the nature of the data itself limits its possible amount, such as yearly GDP for a country, which is itself limited and small. Even if big data do exist, due to the completely change of environment and other determinants, only a small portion of data is still relevant, such as the social economic situation in China and the USA after the trade war. For small data, the impact of uncertainties is even bigger. To this end, the research in dealing with uncertainties in small data is more common than the case in big data. In addition to the mainstream methodologies based on probability distribution (Mostofian and Zuckerman, 2019), the theory of grey systems has achieved significant progress in dealing with small and incomplete data (Liu *et al.*, 2016). Although the uncertainty representation and quantification in grey systems are developed for small data originally, its concepts and methodologies are applicable to big data as well (Yang and Liu, 2018). Hariri *et al.* (2019) listed most available uncertainty models with potential for big data, but grey model is missing in their list. Here, based on a comparative analysis of the existing uncertainty models, we discuss the feasibility of grey models for uncertainties in big data and the complementary feature between big data and small data models in data analysis, and propose the concept of grey data analytics (GDA).

Section 2 gives a review on the major uncertainty models available in computational intelligence, especially the related concepts in grey systems. Section 3 compares the available models in Section 2 and discusses the feasibility of grey models in uncertainty representation of big data. Then, Section 4 defines the concept of GDA and outlines its uncertainty representation framework. Following the proposed GDA, Section 5 discusses the complementary feature between big data and small data in GDA is also discussed. In the end, Section 5 concludes the paper.

2. Review of related uncertainty models

According to *Cambridge Dictionary* (Cambridge University Express, 2019), uncertainty refers to "a situation in which something is not known, or something that is not known or certain". Considering the source of uncertainties, it can be classified as two different groups: subjective and objective. The subjective uncertainties are usually caused by our interpretation, such as our language descriptions. In this case, the object itself is certain, and the uncertainty is only caused by our subjective description. For example, the temperature for a specific time at a specific location is certain, but our feeling of "hot" or "cold" is a subjective description, which may change from one person to another. The objective uncertainty, on the other hand, is caused by the object itself rather than our description. For example, the weather tomorrow is unknown yet, and this "unknown" is not caused by our description, so it is an objective uncertainty. The two different types can certainly be combined together to form a more complicated situation where both subjective and objective uncertainties are at present. For example, "if tomorrow's weather is hot" involves both subjective and objective uncertainties together. To deal with different types of uncertainties, a number of different models have been developed, such as probability (Feller, 1968), Bayesian (Bernardo and Smith, 2009) and belief function models (Cuzzolin, 2014) for randomness, fuzzy sets (Zadeh, 1965) for fuzziness, rough sets (Pawlak, 1982) for roughness and grey systems (Deng, 1982) for greyness.

Among these models, probability models are the mainstream tools applied in both big data and small data analysis to deal with objective uncertainty. They are very effective for objective uncertainties caused by noise, such as randomness. Due to the convenience of the data-driven derivation of probability distribution, they are applied widely in many other related uncertainty models as well, such as Bayesian and belief function models. For uncertainties related with randomness, these models have been proved to be effective both for small data and big data. However, the overall probability can easily hide some specific local issues which may lead to ignorance of the locality of some uncertainty changes in the case of big data. Furthermore, probability models require probability distribution and assume randomness in uncertainties; this is not always applicable for uncertainty modelling when uncertainties other than randomness are involved. In addition to probability-related models, rough sets and grey systems are two models for objective uncertainties different from randomness, and fuzzy sets are defined for subjective uncertainty. The probability-based models are well known, and here we review only the basic concepts in fuzzy sets, rough sets and grey systems.

In human language, we have many terms to describe something between two extremes. For example, "very hot" is not the hottest, but it is more hot than most other candidates. If we consider the hottest and coldest as the two extremes, then "very hot" located between the middle point (neither hot nor cold) and the hottest. Obviously, we have a situation where something cannot be simply classified into one extreme entirely. To represent such a situation, Zadeh (1965) proposed the concept of fuzzy sets:

Definition 1. Fuzzy sets (Zadeh, 1965): let U denote a universe of discourse. Then, a fuzzy set A in U is defined as a set of ordered pairs:

$$\mathrm{A} = \left\{ \left\langle x, \mu_{\mathrm{A}}(x) \right\rangle : x \in U \right\}, \tag{1}$$

where $\mu_{\mathrm{A}}: U \rightarrow [0, 1]$ is the membership function of A and $\mu_{\mathrm{A}}(x)$ is the grade of belongingness of x with regard to the fuzzy set A.

For each element x in a set A, there is always a membership $\mu_{\mathrm{A}}(x)$ to reveal the degree for the element x belonging to A. This membership takes a value between 0 and 1, and it can be any value between the two extremes. In this way, the relationship between an element and a set does not necessarily be categorically belonging (1) or not belonging (0), and it could be partly belonging ($x < 1$) and partly not belonging ($x > 0$) at the same time. Such a facility frees the set definitions from the categorically extremes before, and makes it possible to represent concepts like "very much" and "slightly". This is a revolutionary change for set theory. It should be noted that the object x itself is determined, and the only uncertainty here is the fuzziness represented by the partial membership. It is our artificial ambiguous classification of x to A which causes this fuzziness, and it is completely a subjective justification. For example, if two different persons are asked to give their membership value for a government's performance, more likely they come up with two different membership values for the same government under the same context. It shows that fuzzy sets are an ideal tool to describe subjective uncertainty.

There are many other extensions of fuzzy sets. The fuzzy membership values can be replaced with an interval, then an interval-valued fuzzy set (Sambuc, 1975) is obtained. It can also be replaced with two membership functions values, one for the membership of belonging to the set, and the other one for the membership not belonging to the set, which produces an intuitionistics fuzzy set (Atanassov, 1999). If the fuzzy membership itself is considered as a fuzzy set, it turns into a type-2 fuzzy set (Mendel and John, 2002). If the fuzzy membership is considered as a rough set, then it becomes an R-Fuzzy set (Yang and Hinde, 2010). If the interval membership is extended to include discrete set, it appears as a hesitant fuzzy set (Xu, 2014); if the membership values near to the two extremes and the rest are classified into three different groups, it leads to a shadowed set (Pedrycz, 1998), etc.

Rough sets provide a different facility to describe a set. It describes an undetermined set using approximations. There are many different interpretations of rough sets, here is a definition based on the set-oriented interpretation of rough sets (Yao, 1996):

Definition 2. Rough sets (Yao, 1996): let the pair apr = (*U*, *B*) be an approximation space on *U* and let the *U*/*B* denote the set of all equivalence classes of *B*. *B* is an equivalence relation on *U*. A set which is a union of the empty set ∅ and the objects of *U*/*B* is called a definable set. The family of all definable sets in approximation space apr are denoted by Def (apr). Given by two subsets $\underline{A}$, $\overline{A} \in$ Def (apr) with $\underline{A} \subseteq \overline{A}$, the pair ($\underline{A}$, $\overline{A}$) is called a rough set.

Different from fuzzy sets, a rough set approximates a set A using two definable sets $\underline{A}$ and $\overline{A}$. A is known to be included by $\overline{A}$, and it contains $\underline{A}$. The boundary region between $\underline{A}$ and $\overline{A}$ represents the uncertain region where it is not known if it is part of A or not. Therefore, the roughness can be measured using the cardinality of the uncertain region against the cardinality of the whole possible set $\overline{A}$:

Definition 3. Roughness of approximation (Sambuc, 1975): the roughness $R^{\circ}(A)$ for a set A approximated by ($\underline{A}$, $\overline{A}$) is defined as the significance of the uncertain objects to the set:

$$R^{\circ}(A) = \frac{|\overline{A} - \underline{A}|}{|\overline{A}|}. \tag{2}$$

The larger the boundary region is, the bigger the roughness is. The roughness will be 0 if the boundary region disappears when the two definable sets are identical. Considering the fact that a rough set is defined through partitions of the concerned domain (universe), a finer partition means a smaller boundary region and hence a lower roughness. In this sense, a rough set is defined through information granularity, and a finer granularity brings a more accurate approximation.

Fuzzy sets describe a set by means of fuzzy membership for the relationship between each element and the set. It indicates the strength of the belongings of an element to the set. For a rough set, however, it defines a set through approximation with two definable sets. Fuzzy sets focus on the subjective fuzziness, but rough sets highlight the objective roughness, and they are two different models for different uncertainties.

As a model for small and incomplete data, the theory of grey systems was first proposed by Professor Deng (1982). It divides systems into three different categories: white where everything is known, black where nothing is known and grey systems where part is known and another part is unknown. More specifically, grey systems take grey numbers and its associated degree of greyness as its fundamental concepts:

Definition 4. Grey numbers (Yang and John, 2012): let $\Omega \subset R$ be the universe, $g^{\pm} \in \Omega$ be an unknown real number within a union set of closed or open intervals:

$$g^{\pm} \in \cup_{i=1}^{n} \left[a_i^-, a_i^+\right] \subseteq \Omega, \tag{3}$$

$i = 1, 2, \ldots, n$, n is an integer and $0 < n < \infty$, $a_i^-, a_i^+ \in \Omega$ and $a_{i-1}^+ < a_i^- \leqslant a_i^+ < a_{i+1}^-$. For any interval $[a_i^-, a_{ik}^+] \subseteq \cup_{i=1}^{n} [a_i^-, a_i^+] \subseteq \Omega$, p_i is the probability for $g^{\pm} \in [a_i^-, a_i^+]$.

Definition 5. Degree of greyness of a grey number (Yang and John, 2012): let $\Omega \subset R$ be the universe and $g^{\pm} \in \cup_{i=1}^{n} [a_i^-, a_i^+] \subseteq \Omega$, μ is a measurement defined on Ω. The degree of greyness of $g^{\pm}$ is defined as follows:

$$g^{\circ}(g^{\pm}) = \frac{\mu(g^{\pm})}{\mu(\Omega)}. \tag{4}$$

A grey number can be represented using an interval, a discrete set or a combination of both. Its associated uncertainty can be measured by its degree of greyness. In this way, a number with a known scope but unknown location can be represented by a grey number, and its associated degree of greyness.

In grey systems, a set with everything is known is called a white set, and a set with nothing known is called a black set. A set with only partial information known is referred as a grey set:

Definition 6. Grey sets (Yang and John, 2012): for a set $A \subseteq U$, if the characteristic function value of x with respect to A can be expressed with a grey number $g_A^{\pm}(x) \in \cup_{i=1}^{n} [a_i^-, a_i^+] \in D[0,1]^{\pm}$:

$$\chi_A : U \to D[0,1]^{\pm}, \tag{5}$$

then A is a grey set.

The characteristic function takes value as a grey number between 0 and 1. It could be represented with an interval, a discrete set or a combination of intervals and discrete values. When the two extremes of the grey numbers meet together, it becomes a single number. If the characteristic function is a fuzzy membership function, a fuzzy set defined in Definition 1 is obtained. Therefore, a fuzzy set is a special case of a grey set. Each element in a grey set is associated with a grey number, so a degree of greyness can be defined for each element:

Definition 7. Degree of greyness for an element (Yang and John, 2012): let U be the finite universe of discourse, $x \in U$. For a grey set $A \subseteq U$ the characteristic function value of x with respect to A is $g_A^{\pm}(x) \in D[0,1]^{\pm}$. The degree of greyness $g_A^{\circ}(x)$ of element x for set A is expressed as follows:

$$g_A^{\circ}(x) = \left| g^+ - g^- \right|. \tag{6}$$

With the degree of greyness of each element, the degree of greyness of a set is a natural description of the uncertainty of a grey set:

Definition 8. Degree of greyness for a set (Yang and John, 2012): let U be the finite universe of discourse, and let A be a grey set and $A \subseteq U$. Assume x_i is an element relevant to A and $x_i \in U$. Let $i = 1, 2, 3, \ldots, n$ and n is the cardinality of U. The degree of greyness of set A is defined as follows:

$$g_A^{\circ} = \frac{\sum_{i=1}^{n} g_A^{\circ}(x_i)}{n}. \tag{7}$$

The degree of greyness is a convenient indicator for information incompleteness, but it still needs other measurements to different some special situations, such as the relative uncertainty (Yang *et al.*, 2014).

3. Comparative analysis of the existing models and their feasibility for data analytics

In our previous work, it has been proved that these existing uncertainty models are closely related with each other although they have difference as well (Yang and John, 2012). In Definition 6, the characteristic function is equivalent to a fuzzy membership function when its resulted grey number becomes a white number (the two extremes meet together). In this special case, it turns out to be equivalent to a fuzzy set. In this sense, a white set without greyness can still be fuzzy. It shows that the greyness is different from fuzziness, and they represent different uncertainties. Greyness is caused by incomplete information of the

object, and it is objective. A grey set can be turned into a white set when more information is added. However, this is not the case with fuzzy sets, no matter how much information is added, a fuzzy set is still fuzzy. Fuzziness is caused by our ambiguous classification of objects; this subjective uncertainty has nothing to do with incomplete information. Therefore, a grey set combines objective uncertainty together with subjective uncertainty. In fact, those extended fuzzy sets have also combined objective uncertainties with subjective uncertainty in some way. A fuzzy membership value itself is subjective, but its incompleteness is actually objective. Therefore, those extended fuzzy sets overlap with grey sets in many cases. For example, both interval-valued fuzzy sets and hesitant fuzzy sets overlap with grey sets significantly under some conditions. However, there is a crucial difference between them: a grey set is a single valued set even if its characteristic function is represented as a grey number, but interval-valued fuzzy sets or hesitant fuzzy sets are still multi-valued sets.

Similar to grey sets, rough sets focus mainly on objective incomplete information as well. It is the incomplete information which led to coarse partitions over a given information system. Different from grey sets and fuzzy sets that take individual characteristics function values of each element to define the set, a rough set is approximated through two definable sets constructed from the available information. If more information added, the partition can be finer which leads to a better approximation. In this way, the set could be accurately defined when the two definable sets become identical. Therefore, similar to grey sets, a rough set can be turned into a definable set when all information required is available. Although both grey sets and rough sets are dealing with objective uncertainty, they are quite different from probability models which are still uncertain even if more information is available. Randomness is actually independent from incompleteness. However, data incompleteness due to small data size may have impact on the measurement of probability. In this sense, an increased data size may lead to a more accurate probability, but it will not remove randomness from the data.

Let U be the universe, and A is a set defined on U. For any element x in U, $\mu_A(x)$ is the fuzzy membership value for x belonging to $\underline{A}$; $g_A(x)$ is the degree of greyness of x in A, and g_A is the degree of greyness of A; $(\underline{A}, \overline{A})$ is the rough approximation of A; $p_A(x)$ is the probability distribution of x. i is the available information or data. Then, we have the following properties:

- $\lim_{i\to\infty} \mu_A(x) = \mu_A(x)$;
- $\lim_{i\to\infty} g_A(x) = 0$ and $\lim_{i\to\infty} g_A = 0$;
- $\lim_{i\to\infty} |\overline{A} - \underline{A}| = 0$; and
- $\lim_{i\to\infty} p_A(x) = p_A(x)$.

The first one and the last one show that the subjective fuzziness and the objective randomness do not change when more data are available. However, the increased data may turn a grey number or a grey set into a white number or a white set as shown in the second one. The third one indicates that a rough approximation can become accurate when more data are available. The 0 cardinality of their boundary set means that every element in the universe has a known relationship with the set A: either in or out.

The impact of available new data or information on different models is shown in Figure 1. The addition of new data or information can remove the information incompleteness, so it leads to a white set from a grey set, or a definable set from a rough set. For probability models and fuzzy sets, the additional information does not change their randomness or fuzziness at all. In real-world applications, however, there are many situations where a probability or fuzzy

membership adopted to represent the incompleteness as well, and it is certainly true that such "fuzziness" and "randomness" would change when more data or information become available.

From this property, it is clear that both probability models and fuzzy sets target on uncertainties that do not change with the amount of data available. Compared with probability model and fuzzy sets, grey sets and rough sets focus on information incompleteness, and provide feasible models to consider incompleteness in data sets. In data analytics, randomness has been well defined by probability models and its derived models (Hariri *et al.*, 2019). Fuzziness has also been implemented in fuzzy database (Petry and Bosc, 1996). However, there is no systematics way in dealing with information incompleteness so far. In relational databases, incomplete information is mainly represented by "null" marks (Date, 2000) in relational databases. It refers to either applicable but unknown, or not applicable. This is a very restrictive representation, and it cannot express partial information where some incomplete information might be available. For example, a person's salary might be known between £30,000 and £40,000, but the exact number is not known. With the "null" mark, this available information is not possible to be represented, hence we lose available information. In big data, incomplete data are mainly treated by data imputation (Shobha and Nickolas, 2018) in data pre-processing. However, no matter what imputation methods taken, the derived values are not a factual value and it is not as reliable as other values. These can cause significant problems later on. For example, an image with pedestrian in the road might be a rare case; an imputation from other images might end up with suggestions for the car to cross over it. If the pre-processed data are prepared by imputation, the intelligent system established from such data has no way to detect this, and it might cause serious problem in the end. A much preferred way to deal with it is to preserve the original information as accurate and precise as it is, and catch up the existence of uncertainty in its application stage. In this way, the established intelligent system has a chance to take care when such uncertainty is involved. Following our aforementioned analysis, it is clear that grey sets and rough sets have potential to facilitate such a representation.

Similar to our previous discussion on the difference between grey sets and extended fuzzy sets, a grey set is a single valued set whereas a rough set is a multi-set. A grey set can be considered as a singleton rough set with an empty lower approximation. In terms of its partition, there is only one participation included in the rough set and it is known that there is only one element in the partition belonging to the set although other elements in the partition cannot be separated due to the limited known information. Therefore, rough sets could be applied if multi-set representation is necessary. Here, only singleton values are considered, so we focus mainly on the application of grey sets and grey systems in data analytics.

Although grey systems are mainly focussing on small and incomplete data sets, the definition of grey sets does not exclude data sets with large size. In fact, as aforementioned, a big data set consists of many small data sets, and the summarisation of a big data set still leads to small data sets. There are always occasions where small data sets are necessary even if in big data. In this sense, a combination of big data with small data through grey sets is an ideal option to deal with uncertainties in data sets, being it big or small.

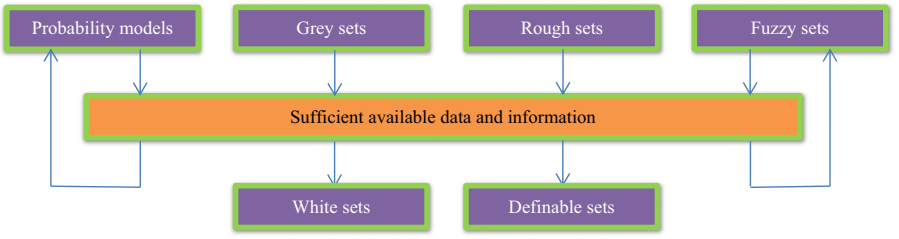

Figure 1.
The impact of new information to uncertainty models

4. Grey data analytics (GDA)

With the development of Industry 4.0 (Popkova *et al.*, 2019) and Society 5.0 (Salgues, 2018), our society is developing towards a direction where everything is connected to provide data to enable intelligent machines to help people in nearly every aspect of our life. The human society is speeding up into an unknown situation where machines and human are mixed together in a much more interactive way. The ability to deal with unknown situations is a nature of human which machines have failed to grasp so far. However, the competition of AI applications between countries means we have to deal with this challenge nowadays. Data analytics is the foundation of AI applications, and a machine can only be as clever as it can grasp from its data. One of the most common uncertainties in data analytics is data incompleteness, such as missing values, incomplete values and inaccurate values. Due to the imperfect devices for data collection, dynamic environment and human errors, the incompleteness is inevitable in a data set, and it is more serious for a large data set assembling data from different sources. The current data pre-processing like data imputation effectively hide these uncertainties from users, which helps the data processing later on but may bring in false inputs to data analysis and lead to serious problems. For example, a traffic accident is definitely rare in a traffic data set recording traffic situation for years. If some data for the accident are missing, a data imputation will more likely to derive data from normal data sets, which effectively remove accident from the database. We can imagine what will happen if a driverless car acts according to such kind of intelligence. Nearly all data analysis tools prefer exact values in data analysis, and assume that all data are perfect reflection of the real-world situations. Although this method is applicable in most cases without significant problems, it has a fundamental problem in that it cannot deal with unknown situations. To avoid such problems, it is essential to take data imperfection as a nature, and make it an essential part for every operation in data analytics.

Applying the concept of grey sets, we consider everything in data analytics as grey in nature. It includes data samples, data storage and data analysis tools. For example, a data sample may not be completely known and there might be some unknown components, a big data set can be considered as a grey set consisting of data samples as elements, as shown in Figure 2. Then their uncertainty measurement can be derived using Equation (7).

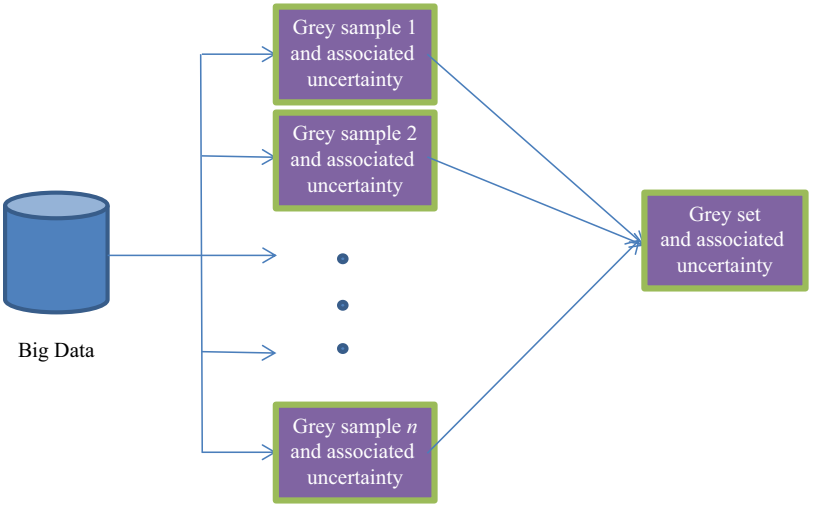

Figure 2.
Data samples and grey sets

Similarly, an analytical tool may also be partly known and partly unknown, and its results are also grey (partial known). In this way, the process of data analytics can be turned into a process taking all data and tools as grey in nature, and carrying out grey data management, grey analysis and grey decision making. Such a data analytics has a crucial difference from the current one: it highlights the uncertainty rather than hiding it, and we call it GDA, as shown in Figure 3.

In Figure 3, the grey data management refers to all facilities and operations to collect and manage data with uncertainties. It involves data collection, data pre-processing and imputation, data representation and storage and data management. The grey tools indicate various data analysis tools whose operations introduce further uncertainties, such as its accuracy, reliability, etc. There grey tools have to take account the data incompleteness and its own associated incompleteness. The grey decision support is the final stage of the grey data analysis. The results from the last stage need to be processed and interpreted with incompleteness involved so as to provide support to business decision making.

4.1 Grey data management

The grey data management refers to the first stage in GDA. It includes grey data collection, grey data imputation, grey data storage and management as shown in Figure 4.

For data collection, data requirement is the first step for the whole process. For a given business requirement at a specific environment and time, there are always incomplete information involved for the data requirement. For example, it is usually not clear in the beginning what kind of data analysis will be involved later on, and what are the influential factors in the data analysis. It leads to an incomplete data requirement which needs to be refined. However, it is difficult to know when we do not have a complete picture of data requirements, so it is usually the case that data collection starts with an incomplete data requirement. Having established the data requirements, then a strategy to collect data has to be designed, such as the devices to be used, location to be selected, time to collect the data, the people who collect the data, etc. Together with data collection strategy, a specific data representation has to be determined for the data to be collected, such as its format,

Figure 3.
Grey data analytics

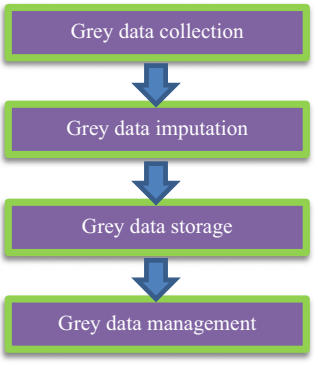

Figure 4.
Operations in grey data

components, accuracy and errors representation, etc. Obviously, all these may introduce data incompleteness, such as neglected important components, wrong formats, unfitted accuracy or error representation. Furthermore, any environment change during the data collection, such as temperature, humility, social events, etc., can introduce further errors and unexpected fluctuations. To make it even worse, different persons may have different skills and knowledge on data collection, which may introduce human errors and differences. All these can potentially introduce data incompleteness in data collection, as shown in Figure 5. In GDA, the incompleteness will be captured right from the data collection stage. Each data sample will be represented as a grey data set, with grey number as its underlined representation of its incompleteness.

Having collected data, the next step is pre-processing, such as data cleaning and imputation to improve the data quality. The current data imputation simply removes uncertainty from data and makes it no difference from other fact data. In GDA, instead of convert unknown into an unreliable known value, the imputation will target to reduce the degree of greyness and other uncertainty measurement rather than completely remove uncertainty. If no other information available, a grey data or even black data will be kept instead of an imputed "white" data. The difference between the imputations of GDA and the current data analytics is shown in Figure 6.

For the data after grey imputation, a data representation and storage strategy has to be drawn to keep not only the data values itself but also all the information on its incompleteness/greyness. To this end, a single value represented by multi-valued sets has to

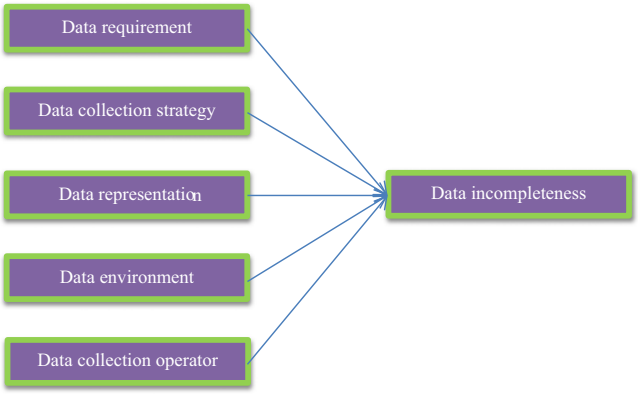

Figure 5.
Uncertainty in data collection

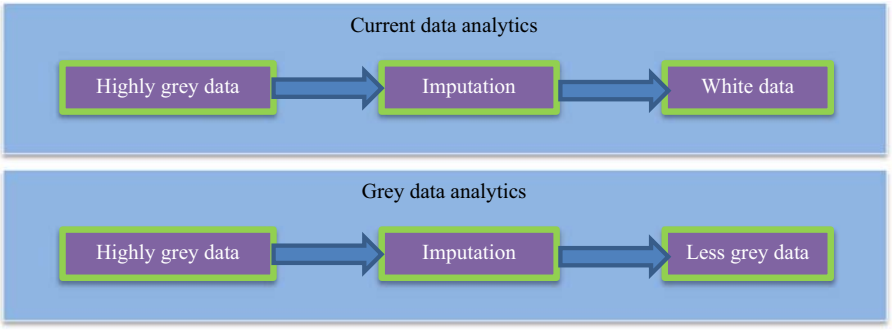

Figure 6.
Comparison between data imputation in GDA and the current data analytics

be facilitated in data storage facility. For relational databases, this requirement does not conflict with its requirement of a single value cell in tables; however, a storage of a set representing this single value has to be enabled. For unstructured databases, this is not an issue. With such a data storage, corresponding data retrieval and data manipulation will have to facilitate such data as well. In relational model, the involvement of grey data makes operations like left join, right join and full join more common than they are in the present relational queries. For unstructured data, similar mechanism for data with grey numbers will have to be facilitated as well.

4.2 Grey data analysis

There have been many data analysis tools available in the market, and most of these tools assume perfect data as inputs and give perfect data as outputs as well. However, it is not only the input data could be grey, the tools themselves could also add further uncertainties into the process. No matter what tools we are using, they involve many parameters for their computation. In most cases, these parameters are some kind of approximation and will introduce errors and bias. In addition to this, the way to use these tools can also lead to new incompleteness to the system, such as the specific structure and algorithm chosen when many different alternatives are available, as shown in Figure 7.

In GDA, all processes implemented by data analysis tools are considered as grey, and hence associated with degree of greyness as well. By combining the uncertainty from tools and the uncertainty from data, it is possible to derive its propagation from inputs to outputs with respect to the specific operations of each individual tool. Such an operation will certainly reveal more information for intelligent systems to take right actions with full consideration of various possibilities.

In addition to the benefit from uncertainty tracking, GDA provides another possibility to combine the data analysis tools for big data and small data together as well. Although big data has significantly enhanced the application of AI, there are still real-world situations where big data technology cannot work and models for small data are still essential (Kennedy *et al.*, 2017; Martin-Diaz *et al.*, 2017; Thinyane, 2017). It is well known that grey prediction models work better for small data sets while models like neural networks work better for big data sets. The two different tools are usually used in different situations separately. However, they do have different merits and are complementary to each other. The big data models have strength in evaluating the long-term and large-scale analysis and prediction, and it is reliable when a general consideration on large scale over long-term period. However, it is not good to evaluate a very specific location at a specific time due to the impact of data from other areas and time. On the contrary, the grey models are good at local and short duration predictions, but they struggle with large-scale analysis. Based on

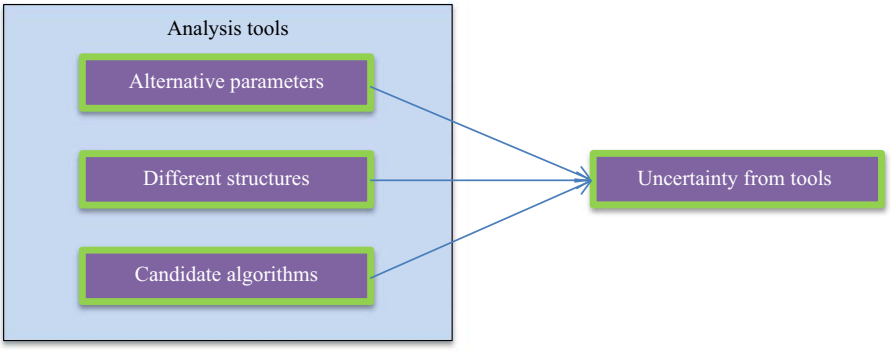

Figure 7.
Uncertainty from data analysis tools

the common facilities for grey data in GDA, the two different models could be combined together to cover both large-scale and small area, long duration and short-term analysis. The big data model will be called first to conduct general analysis and prediction for long term and large area. Then the identified interesting locations and time slots will be focussed and related data will be extracted from large data set into a small data set, and then a grey model will be called upon to carry out local and short-term prediction so as to give further result specific to an identified location and time period. In this way, we can make full use of their different merits, as shown in Figure 8.

4.3 Grey decision support

In data analytics, all the data management and data analysis facilitate the final step – decision-making support. For GDA, this step involves construction of candidate solutions from GDA, comparison, ranking and optimisation of these candidate solutions and then the final grey decision making, as shown in Figure 9. The results from grey data analysis are usually grey in nature, and a full consideration of these grey results will lead to more candidate solutions than the current data analytics. The task to compare, rank and optimise these solutions is even more challenging, and may involve an iteration back to grey data analysis as well. Various models can be involved, such as grey incidence analysis, grey clustering, fuzzy clustering, linear programming, evolutionary algorithms, etc. The result is then fed into grey decision making where a final recommendation can be drawn in terms of visualised or other user friendly forms.

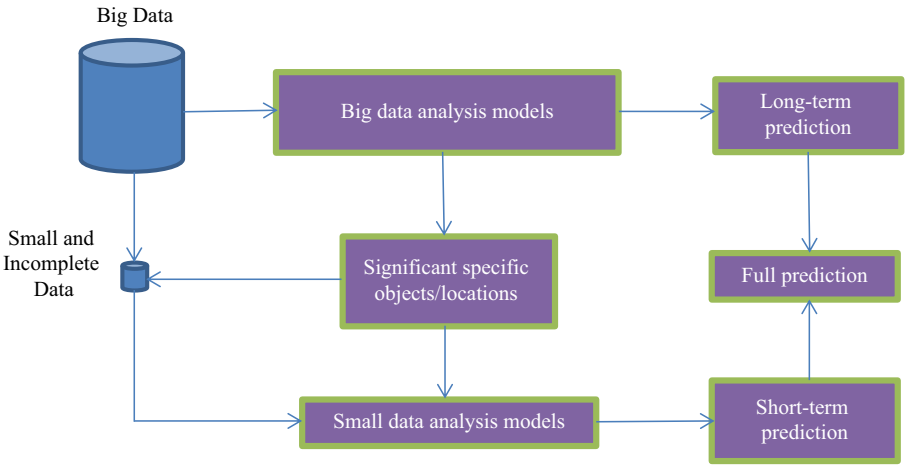

Figure 8.
The combination of big data models with small data models

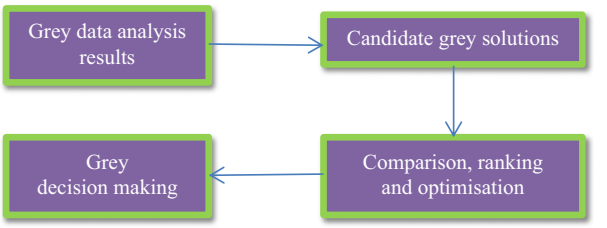

Figure 9.
Grey decision-making support

5. Conclusions

Based on a comparative analysis of the existing uncertainty models, the feasibility of grey sets and grey systems in representing information incompleteness is investigated. The analysis shows that grey sets and grey systems are an ideal option in capturing uncertainty like information incompleteness not only for small data nut also big data. On the basis of this comparative analysis, GDA is proposed as a novel concept for data analytics. The data collection, imputation, storage and management are then discussed, and the data analysis for grey data with full consideration of the possible imperfection of analysis tools is then analysed. The additive advantage to combine the big data analysis with small data model in data analysis stage is then highlighted. Based on the grey data management and grey data analysis, the process of their results for grey decision making is then discussed. The analysis in this paper shows that the proposed GDA opens a brand new field to be explored for the incoming AI enabled society. It will enable a better human machine coexistence and improve people's trust on AI. As the first step, this paper focusses mainly on the concept and there are much more work needed to get its full potential.

References

Atanassov, K.T. (1999), *Intuitionistic Fuzzy Sets*, Physica-Verlag, Heidelberg and New York, NY.

Bernardo, J.M. and Smith, A.F. (2009), *Bayesian Theory*, Vol. 405, Wiley, Hoboken, NJ.

Cambridge University Express (2019), *Cambridge Dictionary*, Cambridge University Express, available at: https://dictionary.cambridge.org/dictionary/english/uncertainty (accessed 12 July 2019).

Cuzzolin, F. (Ed.) (2014), *Belief Functions: Theory and Applications*, Springer International Publishing, Berlin.

Date, C.J. (2000), *An Introduction to Database Systems*, Addison-Wesley, p. 938.

Deng, J. (1982), "The control problems of grey systems", *Systems and Control Letters*, Vol. 1 No. 5, pp. 288-294.

Feller, W. (1968), *An Introduction to Probability Theory and its Applications*, ISBN 0-471-25708-7, Wiley.

Hariri, R.H., Fredericks, E.M. and Bowers, K.M. (2019), "Uncertainty in big data analytics: survey, opportunities, and challenges", *Journal of Big Data*, Vol. 6, p. 44, available at: https://doi.org/10.1186/s40537-019-0206-3

Kennedy, O., Hipp, D.R., Idreos, S., Marian, A., Nandi, A., Troncoso, C. and Wu, E. (2017), "Small data", *Proceedings of the 2017 IEEE 33rd International Conference on Data Engineering*, pp. 1475-1476.

Liu, S., Yang, Y. and Forest, J. (2016), *Grey Data Analysis: Methods, Models and Applications*, Springer-Verlag, p. 333.

Martin-Diaz, I., Morinigo-Sotelo, D., Duque-Perez, O. and Romero-Troncoso, R.J. (2017), "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.

Mendel, J.M. and John, R.I.B. (2002), "Type-2 fuzzy sets made simple", *IEEE Transactions on Fuzzy Systems*, Vol. 10 No. 2, pp. 117-127.

Mostofian, B. and Zuckerman, D.M. (2019), "Statistical uncertainty analysis for small-sample, high Log-variance data: cautions for bootstrapping and Bayesian bootstrapping", *Journal of Chemical Theory and Computation*, Vol. 15 No. 6, pp. 3499-3509.

Pawlak, Z. (1982), "Rough sets", *International Journal of Computer and Information Sciences*, Vol. 11 No. 5, pp. 341-356.

Pedrycz, W. (1998), "Shadowed sets: representing and processing fuzzy sets", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 28 No. 1, pp. 103-109.

Petry, F. and Bosc, P. (1996), *Fuzzy Databases: Principles and Applications*, Kluwer Academic Publishers, p. 226.

Popkova, E.G., Ragulina, Y.V. and Bogoviz, A.V. (2019), *Industry 4.0: Industrial Revolution of the 21st Century*, Springer, p. 253.

Salgues, B. (2018), *Society 5.0: Industry of the Future, Technologies, Methods and Tools*, Technological Prospects and Social Applications, Wiley-ISTE, p. 302.

Sambuc, R. (1975), "Fonctions ϕ-floues. application i'aide au diagnostic en patholologie thyroidienne", PhD thesis, Univ. Marseille.

Shobha, K. and Nickolas, S. (2018), "Imputation of multivariate attribute values in big data", *Smart Intelligent Computing and Applications*, Springer, Vijayawada, pp. 53-60.

Thinyane, M. (2017), "Small data and sustainable development individuals at the center of data-driven societies", *Proceedings of the 2017 IEEE ITU Kaleidoscope: Challenges for a Data-Driven Society, Nanjing, 27-29 November*.

Wang, X. and He, Y. (2016), "Learning from uncertainty for big data: future analytical challenges and strategies", *IEEE Systems, Man, and Cybernetics Magazine*, Vol. 2 No. 2, pp. 26-31.

Xu, Z. (2014), *Hesitant Fuzzy Sets Theory*, Springer, p. 466.

Yang, Y. and Hinde, C. (2010), "A new extension of fuzzy sets using rough sets: R-Fuzzy sets", *Information Sciences*, Vol. 180 No. 3, pp. 354-365.

Yang, Y. and John, R. (2012), "Grey sets and greyness", *Information Sciences*, Vol. 185 No. 1, pp. 249-264.

Yang, Y. and Liu, S. (2018), "Grey systems, grey models and their roles in data analytics", *International Journal of Simulation: Systems, Science and Technology*, Vol. 19 No. 3, pp. 8.1-8.6.

Yang, Y., Liu, S. and John, R. (2014), "Uncertainty representation of grey numbers and grey sets", *IEEE Transaction on Cybernetics*, Vol. 44 No. 9, pp. 1508-1517.

Yao, Y.Y. (1996), "Two views of the theory of rough sets in finite universe", *International Journal of Approximate Reasoning*, Vol. 15 No. 4, pp. 291-317.

Zadeh, L. (1965), "Fuzzy sets", *Information and Control*, Vol. 8 No. 3, pp. 338-353.

Further reading

Wang, X. and Huang, J.Z. (2015), "Uncertainty in learning from big data", *Fuzzy Sets and Systems*, Vol. 258 No. 1, pp. 1-4.

Corresponding author
Yingjie Yang can be contacted at: yyang@dmu.ac.uk