Towards Construction of an Explanation Framework for Whole Processes of Data Analysis Applications: Concepts and Use Cases

Hiroshi Ishikawa
Graduate School of Systems Design
Faculty of System Design
Tokyo Metropolitan University
Hino, Tokyo
E-mail: ishikawa-hiroshi@tmu.ac.jp

Yukio Yamamoto
Japan Aerospace Exploration Agency
Sagamihara, Kanagawa
E-mail: yamamoto.yukio@jaxa.jp

Masaharu Hirota
Department of Information Science
Faculty of Informatics
Okayama University of Science
Okayama, Okayama
E-mail: hirota@mis.ous.ac.jp

Masaki Endo
Division of Core Manufacturing
Polytechnic University
Kodaira, Tokyo
E-mail: endou@uitec.ac.jp

***Abstract*- The main contribution of the paper is to address the necessity of both macro and micro explanations for Social Big Data (SBD) applications and to propose an explanation framework integrating both of these, allowing SBD applications to be more widely accepted and used. The framework provides both a macro explanation of the whole procedure and a micro explanation of the constructed model, as well as an explanation of the decisions made by the model. For a macro explanation of the application, we introduce a data model for abstractly describing all processes from data acquisition to data analysis. We explain the processes based on the data model. For the micro explanation, we illustrate the basis of the interpretation of the analytical model and the decisions made when applying it. We describe some of the specific features of the explanation framework proposed through multiple use cases.**

Keywords*- *social big data; explanayion; data model; data management; data mining.

I. INTRODUCTION

We are surrounded by big data, which are waiting to be analyzed and used. Big data are real data, such as automobile driving data and space observation data, generated from real world measurement and observation, social data derived from social media, e.g., Twitter and Instagram, and open data published by highly public groups, e.g., weather data and evacuation location data. These are generally called social big data (SBD) [9] [11]. Furthermore, SBD are inherently represented by multimedia (MM). By integrating and analyzing social big data, new knowledge can be obtained, which is expected to bring new value to society.

Further, as the horizon of applications whose main task is data analysis has spread, the following problems have emerged:

- Application to science, e.g., lunar and planetary science

Analytical applications in this field require strictness as science. That is, explanation of the protocol (procedure) of analysis and explanation of the reason for decisions are required. In addition, as to the interpretation of the analytical model, it is necessary to explain the input data (for learning and test) and the data manipulation on the data, and the procedure (algorithm and program) for model construction. In order to interpret the individual results, it is necessary to explain the input data (actual data) and the reasons for the decisions.

- Application to Social Infrastructure, e.g., Mobility as a Service (MaaS)

Analytical applications in this field require consent of practitioners. That is, the analysis result must be consistent with the practitioners' own experiences, and especially in the case of applications such as ones related to human life, it is necessary to fulfill the accountability to the concerned parties. Interpretation of both a model and individual results is necessary as with science. In addition, especially if the data about the generic users are utilized in applications, interpretation of the model is also important in order to get rid of the general users' concerns.

In order for social big data to widely be used, it is necessary to explain to the user the application system. Both microscopic description, that is, interpretation of the analytical model and explanation of individual decisions and macroscopic description, that is, description of the

 ISBN: 978-1-61208-697-2

whole process including the data manipulation and the model construction are required.

First of all, the reason why a macro explanation is necessary is described below. In order for social big data applications to be accepted by users, it is necessary to ensure at least their reliability. Since information science is one area of science, we should guarantee reproducibility as science. In other words, it is necessary to ensure that third parties can prepare and analyze data according to given explanation and can get the same results.

In addition, in order for the service to be operatable, it is necessary for the final user of the service to be convinced of how the service processes and uses the personal information. In addition, if the users can be convinced of the description of way of using the personal information, the progress of data portability can be advanced based on the EU's GDPR law on personal information protection [5] and Japan-based information bank to promote the use of personal information [18].

Next, a micro explanation is necessary for the following reasons. In order for analysts of social big data and field experts using the data to accept decisions made by the constructed model, it is assumed that they must understand the structure, actions and grounds of the model and are satisfied with them as well.

Up to now, the authors have been involved in the development of a wide range of social big data use cases ranging from tourism, disaster prevention to lunar and planetary science [12] [26]. In the course of these processes, from the users of the use cases, we have often received questions as to what kind of data are processed, what kind of model are created as the core of analysis, and furthermore, what are the grounds for the decisions. In other words, from the development experiences of multiple use cases, we have come to think that both the macro explanation proposed in this paper and the micro explanation emerging in AI are urgently needed.

To date, the authors created multiple seismic source classifiers of the lunar earthquakes (moonquakes) in the field of lunar and planetary science using the Balanced Random Forest [3], and the features, e.g., the distance between the moon and the earth, were calculated and studied for extracting features strongly related to cause of moonquakes as a micro explanation [12]. With regard to a macro explanation, the authors also showed that by observing many use cases, social big data applications should include different digital ecosystems such as data management (database operation) and data analysis (data mining, machine learning, artificial intelligence), we have noticed that it is necessary to have a method to generally describe the whole process of application consisting of such a hybrid digital ecosystem. Therefore, as a framework to describe processes in an abstraction level independent of a specific programming language, we have come to think of adopting a data model [8] developed in the field of database and proposed a framework for description using mathematical concept of set family [10]. As described in the subsequent section of the related works, the micro explanation research is being actively carried out, whereas as far as research on the framework for the macroscopic description is not known except our work.

The main contribution of the paper is to address the necessity of both macro and micro explanations for SBD applications and to propose an explanation framework integrating both of them. This will allow SBD applications to be more widely accepted and used. Although this paper describes our research-in-progress, we propose an integrated framework for explanation and introduce a part of its functions through case studies. In Section II, we introduce our explanation framework. Through use case examples of macroscopic description and microscopic description, we describe the features of the proposed approach in Sections III and IV, respectively.

II. OUR APPROACH

A. Explanation Framework

For a macro explanation of applications, the goal is to facilitate a data model for abstractly describing the entire processes from data acquisition to data analysis and to explain the processes based on the description. For the micro explanation, we aim to show the basis of the interpretation of the constructed model and the individual decisions made when applying it.

1) Construction of a theoretical foundation for integrated explanation

For that purpose, we build a theoretical framework of the technical foundation that integrates the following micro and macro explanatory methods.

a) Macro explanation function: The application system is a hybrid ecosystem consisting of data management and data mining (including machine learning and Artificial Intelligence, or AI), and the function must be able to describe the application seamlessly. Moreover, it must be able to describe the application in a high level not depending on individual environments or programming languages. Therefore, we first create a framework to unify the hybrid ecosystem based on the data model approach. In other words, we develop a method to provide macro explanations with the constituent elements (data structure and data manipulation) of the model based on the mathematical family of sets as a basic unit. The explanation mechanism provided by the proposed framework presents as a macro explanation a sequence of operations on databases to the user based on the model of SBD applications consisting of data management and data mining, as in a use case depicted in Section III.

b) Micro explanatory function: We develop an explanatory method independent of analytical model by extending explanatory functions based on attributes or constituent elements, which is an emergent approach in AI, discussed in the related work subsection. In other words, in model categories for structured data consisting of attributes, such as Support Vector Machine (SVM) and decision trees, we develop a method for systematically discovering subsets of attributes with strong influence on analysis results based on multiple weak classifiers. Especially this function is used to interpret the model itself. In model categories like Deep Neural Network (DNN) suitable for non-structured data such as images, we develop a method of explaining the analysis result based on the constituent elements or decomposition of the image with the use of annotation or attention. Especially this function is used to show the basis of individual decisions. For the micro explanation of the reasons for decisions, if the analysis target is image data, a part of the image which leads to the conclusion is indicated by concepts or words as its annotations based on a heat map. If the object is structural data, that is, it consists of attributes, the micro explanation is presented in terms of the contribution ratios of the attributes as in a use case depicted in Section IV.

2) Collection of use cases and verification of basic technology

First, we collect several different kinds of use cases (tourism, mobility service, lunar exploration). We generate concrete explanations as targets for typical ones, using the integrated explanatory platform developed in items *a* and *b* and verify its feasibility

3) Implementation of Explanation generation and presentation method

Based on the theoretical framework of the integrated infrastructure, an automatic generation method of explanation and a presentation function of explanations are implemented. We evaluate their effectiveness by performing the experiments. We also incorporate InfoGraphics [23] as a method of presenting explanations to users since the users are not always analysis experts.

Basically, for micro explanation, we create explanations of individual decisions by solving partial problems that restrict information existing in original problems.

In this research, we aim to develop both the emerging microscopic-explanatory functions and macroscopic-explanatory functions and to build a framework for integrating two kinds of explanations.

B. Related Research

As a trend other than the authors' research, researches corresponding to micro explanatory functions have become active in AI, what is so called eXplainable AI (XAI) at present.

First, there is an attempt [14] to try to give a basic definition to the possibility of interpretation of a model in machine learning and a research [4] on the evaluation method of interpretability.

Next, individual studies on XAI are roughly classified into (1) description based on features, (2) interpretable model, and (3) derivation of explanation model. A research is done to create a classification rule for explanation by creating a subset of features in SVM as a category of (1) [15]. In addition, in the image classification using Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM), there is a research to generate explanations based on both image features and class features [6]. Further there is a research introducing the explanation vector to make explicit the most important attributes [1]. In the category of (2), there is a research using a AND/OR tree to discover the components of the model [22] and a research to make models that can be interpreted by considering the generation process of features [13]. A research deriving description with reference of any classifier of the local approximation model falls into the category (3) [20].

While developing along the approaches of (1) and (3) as a micro explanatory technique, we aim to build a comprehensive explanation basis by conducting research on macroscopic explanation technology.

In addition, although there is an application of infographics to a tourism use case [25], our research aims at basic research that can be widely used for visualization of explanation of general analysis.

III. CASE STUDY: MACRO EXPLANATION OF TOURISM APPLICATION

We will describe a case that explains how our data is used in analysis application. For that purpose, an integrated data model is introduced as a macroscopic description of an analytical application which is a hybrid ecosystem. Then the application is described using the integrated model as a basis for macro explanation.

A. Integrated Model

We propose our SBD data model consisting of data structures and operations in the following subsections.

1)Data model for SBD

Our SBD model uses a mathematical concept of a *family* [24], a collection of sets, as a basis for data structures. Family can be used as an apparatus for bridging the gaps between data management operations and data analysis operations.

Basically, our database is a *Family*. A Family is divided into *Indexed family* and *Non-Indexed family.* A Non-Indexed family is a collection of sets.

An Indexed family is defined as follows:

- $\{Set\}$ is a Non-Indexed family with *Set* as its element.
- $\{Set_i\}$ is an Indexed family with Set_i as its *i-th* element. Here, *i: Index* is called *indexing set* and *i* is

an element of Index.

- Set is {<time space object>}.
- *Set*$_i$ is {<*time space object*>}$_i$. Here, *object* is an identifier to arbitrary identifiable user-provided data, e.g., record, object, and multimedia data appearing in social big data. *Time* and *space* are universal keys across multiple sources of social big data.
- {*Indexed family*$_i$} is also an Indexed family with *Indexed family*$_i$ as its *i-th* element. In other words, Indexed family can constitute a hierarchy of sets.

Please note that the following concepts are interchangeably used in this paper.

- Singleton family ⇔ set
- Singleton set ⇔ element

As described later in this section, we can often observe that SBD applications contain families as well as sets and they involve both data mining and data management. Please note that a family is also suitable for representing hierarchical structures inherent in time and locations associated with social big data.

If operations constructing a family out of a collection of sets and those deconstructing a family into a collection of sets are provided in addition to both family-dedicated and set-dedicated operations, SBD applications will be described in an integrated fashion by our proposed model.

2) SBD Operations

SBD model constitutes an algebra with respect to Family, as follows.

SBD consists of Family data management operations and Family data mining operations. Further, Family data management operations are divided into Intra Family operations and Inter Family operations.

First, Intra Family Data Management Operations are described as follows:

a) Intra Indexed Intersect (*i*:*Index Db p*(*i*)) returns a singleton family (i.e., set) intersecting sets which satisfy the predicate *p*(*i*). Database *Db* is a Family, which will not be mentioned hereafter.

b) Intra Indexed Union (*i:Index Db p*(*i*)) returns a singleton family union-ing sets which satisfy *p*(*i*).

c) Intra Indexed Difference (*i:Index Db p*(*i*)) returns a singleton family, that is, the first set satisfying *p*(*i*) minus all the rest of sets satisfying *p*(*i*)

d) Indexed Select (*i:Index Db p1*(*i*) *p2*(*i*)) returns an Indexed family with respect to *i* (preserved) where the element sets satisfy the predicate *p1*(*i*) and the elements of the sets satisfy the predicate *p2*(*i*). As a special case of true as *p1*(*i*), this operation returns the whole indexed family. In a special case of a singleton family, Indexed Select is reduced to Select (a relational operation).

e) Indexed Project (*i:Index Db p*(*i*) *a*(*i*)) returns an Indexed family where the element sets satisfy *p*(*i*) and the elements of the sets are projected according to *a*(*i*), attribute specification. This also extends also relational Project.

f) Intra Indexed cross product (*i:Index Db p*(*i*)) returns a singleton family obtained by product-ing sets which satisfy *p*(*i*). This is extension of Cartesian product, one of relational operators.

g) Intra Indexed Join (*i:Index Db p1*(*i*) *p2*(*i*)) returns a singleton family obtained by joining sets which satisfy *p1*(*i*) based on the join predicate *p2*(*i*). This is extension of join, one of relational operators.

h) Select-Index (*i:Index Db p*(*i*)) returns *i:Index* of *set*$_i$ which satisfy *p*(*i*). As a special case of true as *p*(*i*), it returns all index.

i) Make-indexed family (*Index Non-Indexed Family*) returns an indexed Family. This operator requires *order-compatibility*, that is, that *i* corresponds to *i-th set of Non-Indexed Family*.

j) Partition (*i:Index Db p*(*i*)) returns an Indexed family. Partition makes an Indexed family out of a given set (i.e. singleton family either w/ or w/o index) by grouping elements with respect to *p* (*i:Index*). This is extension of "groupby" as a relational operator.

k) ApplyFunction (*i:Index Db f*(*i*)) applies *f*(*i*) to *i-th* set of DB, where *f*(*i*) takes a set as a whole and gives another set including a singleton set (i.e., Aggregate function). This returns an indexed family. *f*(*i*) can be defined by users.

Second, Inter Family Data Management Operations are described as follows:

All are assumed to be Index-Compatible

a) Indexed Intersect (*i:Index Db1 Db2 p*(*i*)) union-compatible

b) Indexed Union (*i:Index Db1 Db2 p*(*i*)) union-compatible

c) Indexed Difference (*i:Index Db1 Db2 p*(*i*)) union-compatible

d) Indexed Join (i:Index Db1 Db2 p1(i) p2(i))

e) Indexed cross product (*i:Index Db1 Db2 p*(*i*))

Finally, Family Data Mining Operations are described as follows:

a) Cluster (*Family method similarity* {*par*}) returns a Family as default, where Index is automatically produced. This is an unsupervised learner.

b) Make-classifier (*i:Index set:Family learnMethod* {*par*}) returns a classifier (Classify) with its accuracy. This is a supervised learner.

c) Classify (*Index/class set*) returns an indexed family with class as its index.

d) Make-frequent itemset (*Db supportMin*) returns an Indexed Family as frequent itemsets, which satisfy *supportMin*.

e) Make-association-rule (*Db confidenceMin*) creates association rules based on frequent itemsets *Db*, which satisfy *confidenceMin*. This is out of range of our algebra, too.

Please note that the predicates and functions used in the above operations can be defined by the users in addition to the system-defined ones such as Count.

B. Tourist Applications

We describe a case study, finding candidate access spots for accessible Free Wi-Fi in Japan [16]. This case is classified as integrated analysis based on two kinds of social data.

This section describes our proposed method of detecting attractive tourist areas where users cannot connect to accessible Free Wi-Fi by using posts by foreign travelers on social media.

Our method uses differences in the characteristics of two types of social media:

Real-time: Immediate posts, e.g., Twitter

Batch-time: Data stored to devices for later posts, e.g., Flickr

Twitter users can only post tweets when they can connect devices to Wi-Fi or wired networks. Therefore, travelers can post tweets in areas with Free Wi-Fi for inbound tourism or when they have mobile communications. In other words, we can obtain only tweets with geo-tags posted by foreign travelers from such places. Therefore, areas where we can obtain huge numbers of tweets posted by foreign travelers are identified as places where they can connect to accessible Free Wi-Fi and /or that are attractive for them to sightsee.

Flickr users, on the other hand, take many photographs by using digital devices regardless of networks, but whether they can upload photographs on-site depends on the conditions of the network. As a result, almost all users can upload photographs after returning to their hotels or home countries. However, geo-tags annotated to photographs can indicate when they were taken. Therefore, although it is difficult to obtain detailed information (activities, destinations, or routes) on foreign travelers from Twitter, Flickr can be used to observe such information. In this study, we are based on our hypothesis of “A place that has a lot of Flickr posts, but few Twitter posts must have a critical lack of accessible Free Wi-Fi.” We extracted areas that were tourist attractions for foreign travelers, but from which they could not connect to accessible Free Wi-Fi by using these characteristics of social media. What our method aims to find is places currently without accessible Free Wi-Fi.

Our method envisaged places that met the following two conditions as candidate access spots for accessible free Wi-Fi:

- Spots where there was no accessible Free Wi-Fi
- Spots that many foreign visitors visited

We use the number of photographs taken at locations to extract tourist spots. Many people might take photographs of subjects, such as landscapes based on their own interests. They might then upload those photographs to Flickr. As these were locations at which many photographs had been taken, these places might also be interesting places for many other people to sightsee or visit. We have defined such places as tourist spots. We specifically examined the number of photographic locations to identify tourist spots to find locations where photographs had been taken by a lot of people. We mapped photographs that had a photographic location onto a two-dimensional grid based on the location at which a photograph had been taken to achieve this. Here, we created individual cells in a grid that was 30 square meters. Consequently, all cells in the grid that was obtained included photographs taken in a range. We then counted the number of users in each cell. We regarded cells with greater numbers of users than the threshold as tourist spots.

[Integrated Hypothesis] Based on different data generated form Twitter and Flickr by using our generalized difference method, the fragment collects attractive tourist spots for foreign visitors but without accessible free Wi-Fi currently (See Figure 1):

$DB_{t/visitor}$ ← Tweet DB of foreign visitors obtained by mining based on durations of their stays in Japan;

$DB_{f/visitor}$ ← Flickr photo DB of foreign visitors obtained by mining based on their habitations;

T ← Partition (i:*Index grid* $DB_{t/visitor}$ $p(i)$); This partitions foreign visitors tweets into grids based on geo-tags; This operation returns a indexed family.

F ← Partition (j:*Index grid* $DB_{f/visitor}$ $p(j)$); This partitions foreign visitors photos into grids based on geo-tags; This operation returns a indexed family.

Index1 ← Select-Index (i:*Index T Density*(i) >= *th1*); *th1* is a threshold. This operation returns a singleton family.

Index2 ← Select-Index (j:*Index F Density*(i) >= *th2*); *th2* is a threshold. This operation returns a singleton family.

Index3 ← Difference (*Index2 Index1*); This operation returns a singleton family.

Plaese note that Partition and Select-Index are family data management operations while Difference is a relational (set) data management operation.

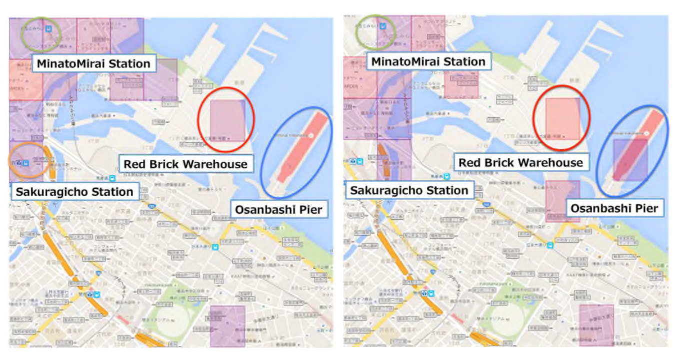

Figure 1. Differences of high-density areas of Tweets (left) and of Flickr photos (right).

We collected more than 4.7 million data items with geo-tags from July 1, 2014 to February 28, 2015 in Japan. We detected tweets tweeted by foreign visitors by using the method proposed by Saeki et al. [7]. The number of tweets that was tweeted by foreign visitors was more than 1.9 million. The number of tweets that was tweeted by foreign visitors in the Yokohama area was more than 7,500. We collected more than 5,600 photos with geo-tags from July 1, 2014 to February 28, 2015 in Japan. We detected photos that had been posted by foreign visitors to Yokohama by using our proposed method. Foreign visitors posted 2,132 photos. For example, grids indexed by *Index3* contain "Osanbashi Pier." Please note that the above description doesn't take unique users into consideration.

IV. CASE STUDY: MICRO EXPLANATION FOR SCIENCE APPLICATION

In this section, we present the case of determining features important for interpreting the constructed model by reducing features with small contribution ratios.

We apply Balanced Random Forest [3] which extends Random Forest [2], a popular supervised learning method in machine learning, to lunar and planetary science to verify the key features in analysis. Our verification method tries to confirm whether the known seismic source labels can be reproduced by Balanced Random Forest using the features described below based on the features constructed from the moonquakes with the seismic source label of the known moonquake as the correct label.

A. Features for Analysis

TABLE I shows the parameters in the coordinate systems used in this section. We use as seismic source of moonquakes the position on the planets of the moon, the sun, the earth, and Jupiter (*X , y , z*), velocity (*v x , v y , v z*), and distance (*lt*) . Based on the time of moonquake occurrence, we calculate and use features using SPICE [17]. Here, sun perturbation is the solar perturbation. The IAU MOON coordinate system is a fixed coordinate system centered on the moon. The z axis is the north pole direction of the moon, the x axis is the meridian direction of the moon, the y axis is the right direction with respect to the plane xz. The IAU EARTH coordinate system is a fixed coordinate system centered on the earth. Here, the z axis is the direction of the conventional international origin, the x axis is the direction of the prime meridian, and the y axis is the right direction with respect to the xz plane.

We also calculate the period of the perigee at the distance of earth from moon, the period based on the period of the perigee, the periods of the *x* coordinate and the *y* coordinate of the solar perturbation. *sin* and *cos* values are calculated from these periodic features and the phase angle based on them. In addition, at the positions moon from earth and sun from earth, we calculate the *cos* similarity as the features of the sidereal moon. As all possible combinations of these features, a total of 55 features are used in experiments described in this paper.

B. Balanced Random Forest

Random Forest is an ensemble learning that combines a large number of decision trees and is widely used in fields such as data mining and has a characteristic that the contribution ratio of features can be calculated. However, Random Forest has a problem such that when there is a large difference in the number of data to be learned depending on class labels, the classifier is learned biased towards classes with a large number of data. Generally, we address the problem of imbalanced data by weighting classes with a small number of data. However, if there is any large skew between the numbers of data, the weight of data belonging to classes with a small number will become large, which is considered to cause over fitting to classes with a small number of data. Since the deep moonquakes have a large difference in the number of events for each seismic source, it is necessary to apply a method considering imbalanced data.

As analysis considering imbalanced data, we apply Balanced Random Forest [3], which makes the number of samples even for each class when constructing each decision tree. Balanced Random Forest divides each decision tree based on the Gini coefficient. Gini coefficient is an index representing impurity degree, which takes a value between 0 and 1. The closer it is to 0, the higher the purity is, that is, the less variance the data have. The contribution ratio of the feature is calculated for each feature by calculating the reduction ratio by the Gini coefficient at the branch of the tree. The final contribution ratio is the average value of contribution ratios of each decision tree.

C. Experiment Setting

Here, we describe experiments for evaluating features effective for seismic source classification, together with the results and considerations. Based on the classification performance and the contribution ratio of the features by Balanced Random Forest, we analyze the relationship between the seismic sources in the features used in this paper.

The outline of feature analysis is shown below.

- Features are calculated based on the time of occurrence of moonquake.
- Balanced Random Forest is applied to each pair of all seismic sources.
- Classification performance and the contribution ratio of the features by Balanced Random Forest are calculated and analyzed.

In this paper, as one-vs-one method, by constructing the classifier for every pair of two seismic sources in the dataset, we perform analysis paying attention to

 ISBN: 978-1-61208-697-2

characteristics of each seismic source and the relationship between seismic sources. 100 Random Forests are constructed for each classifier. The number of samples used to construct each decision tree are taken 50 by bootstrap method. Also, scikit-learn [19] was used to construct each decision tree in Random Forest.

In this paper, we perform the following analysis as feature selection.

• We create a classifier that learns all of the extracted 55 features.

TABLE I. PARAMETERS IN THE COORDINATE SYSTEMS COMPUTED USING SPICE.

Target	Observer	Coordinate system	Parameter
EARTH BARYCENTER	**MOON**	IAU MOON	**earth_from_moon**
SOLAR SYSTEM BARYCENTER	**MOON**	IAU MOON	**sun_from_moon**
JUPITER BARYCENTER	**MOON**	IAU MOON	**jupiter_from_moon**
SOLAR SYSTEM BARYCENTER	**EARTH BARYCENTER**	IAU EARTH	**sun_from_earth**
JUPITER BARYCENTER	**EARTHBARYCENTER**	IAU EARTH	**jupiter_from_earth**
SUN	**SOLAR SYSTEM BARYCENTER**	IAU EARTH	**sun_perturbation**

TABLE II. NUMBER OF DATA FOR EACH SEISMIC SOURCE.

Seismic source	A1	A5	A6	A7	A8	A9	A10	A14	A18	A20	A23	A25	A35	A44	A204	A218
Number of data	441	76	178	85	327	145	230	165	214	153	79	72	70	86	85	74

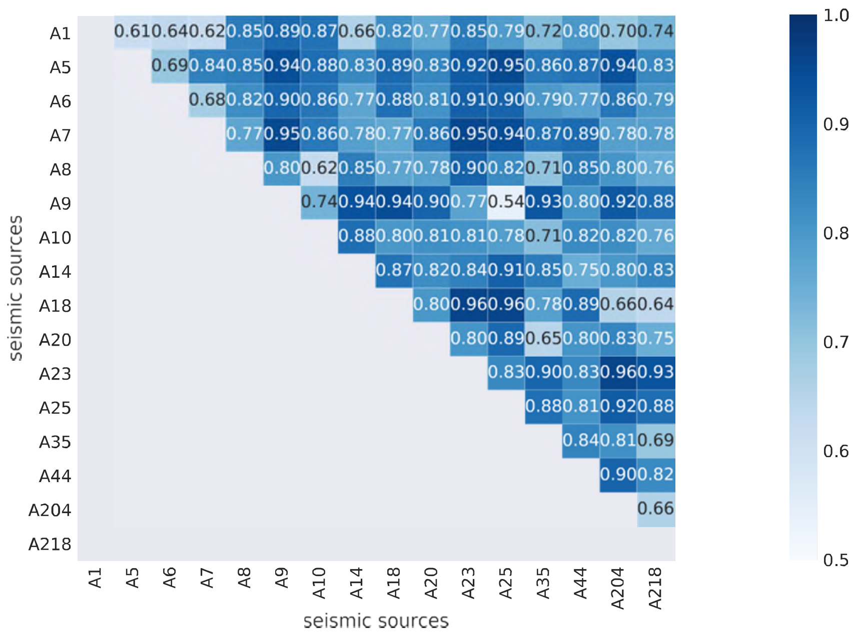

Figure 2. Averages of F-values for pairs of seismic sources.

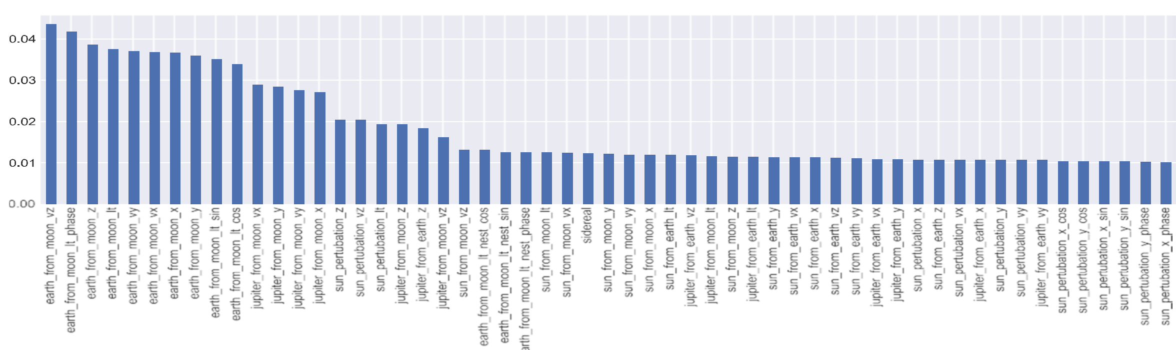

Figure 3. Averages of contribution ratios for each feature.

• Using the Variance Inflation Factor (VIF), we construct a classifier after reducing features.

Here, VIF is one of the indicators used to evaluate *multicollinearity*. In this paper, in order to make VIF of each feature 6 or less, experiments were conducted on a subset with reduced features. Based on the experimental results using all features, we calculate VIF and delete features with 6 or more VIF. To calculate VIF, statsmodel [21] was used.

TABLE II shows the dataset in this paper. We select events of 16 seismic sources whose observed number of moonquake events is 70 or more.

In this paper, the precision ratio, recall ratio, and F-value are used as indexes for evaluating the performance of classification of seismic sources.

The precision ratio is an index for measuring the accuracy of the classification, and the recall ratio is an index for measuring the coverage of the classification. F-value is the harmonic mean of recall and precision ratios and is an index in consideration of the balance of precision and recall. The score of the classifier in this paper is the average value of the F-values of the two classes targeted by the classifier.

D. Experiment Results

1) Experimental results using all features

a) Classification performance

Figure 2 is the average of the F-values of classifiers for each seismic source. The vertical axis and the horizontal axis show seismic sources, each value is a score of the average of F-value of classifier. In Figure 2, the highest classification performance is 0.96 and it is observed in multiple pairs of seismic sources. Also, the lowest classification performance is 0.54 as of classifier between A9 and A25. Figure 2 shows that some classification is difficult depending on combinations of seismic sources. Also, the number of classifiers with 0.9 or higher as classification performance is 20, about 17% of the total number of the classifiers. The number of classifiers with 0.8 or more and less than 0.9 is 60, 50% of the total. The number of classifiers with performance below 0.6 is only one. Most of the classifiers show high classification performance and show that the positional relationships of the planets are effective for the seismic source classification of the deep moonquakes.

b) Contribution ratio of features

Figure 3 shows the average value of contribution ratios for each feature. All features with the higher contribution ratios are those of the earth when they are calculated as the moon as the origin. In addition, it shows that the contribution ratios of Jupiter 's features are high when the moon is the origin while those of earth features is high when the moon is the origin. By comparing features when the moon is the origin and when the earth is the origin, the features with the moon as the origin has a higher contribution ratio than the features with the earth as the origin. Figure 3 indicates that relationships between the moon and the Earth affect the classification most strongly. However, there is a possibility that correlation between features, then it is necessary to further analyze each feature from view point of mutual independence. Therefore, in the following subsection, considering the correlations between features, we will describe the experimental results after feature reduction using VIF.

2) Experimental results of feature reduction using VIF.

a) Classification performance

Figure 4 shows the average of the F-values of the classifier when the features are reduced. Similarly, as in Figure 2, the vertical axis and the horizontal axis are seismic sources, respectively, and each value is the score of the F-value of the classifier in Figure 4. In addition, the number of classifiers whose classification performance is 0.9 or higher is 26, about 22% of the total. 54 classifiers with 0.8 or higher but less than 0.9 are 45% of the total. There is one classifier whose classification performance is less than 0.6. Compared with Figure 2, these show that the classification performance does not change significantly.

b) Contribution ratio of features

Figure 5 shows the average value of the contribution ratios of each seismic source after feature reduction. After reducing features, earth features when the origin is the moon are reduced to 4 features of the top 10 features which existed before feature reduction. The four features between top 11 and 14 positions of the features of Jupiter when the origin is the moon, as shown in Figure 3, are reduced to one feature. Other parameters of Jupiter are thought to have been affected by other features. The subset of the features after feature reduction is considered to have small influence of multicollinearity. Therefore, there is a possibility that the features of the Earth and some of the features of Jupiter are effective for classification when the moon is the origin,

E. Discussion of methods and features

By using Balanced Random Forest, contribution ratios of features can be easily calculated in addition to classification performance, so it is useful for feature analysis like the scientific research described in this section. However, in this method, there is room for consideration of parameters of classification techniques depending on the seismic sources as the classification targets. Moreover, in order to obtain higher classification performance, it is necessary to consider many classification methods. Furthermore, it is necessary to apply a method considering waveform information. In addition, since the findings obtained in this paper are only correlations, it is difficult to directly estimate the causal mechanism of the deep moonquakes. However, the results of this paper are shown to be useful for new analysis and knowledge creation of experts. If the knowledge of experts is available, the elucidation of the causal relationships between the seismic sources and the planetary bodies and ultimately that of the causal mechanism of the moonquakes can be expected.

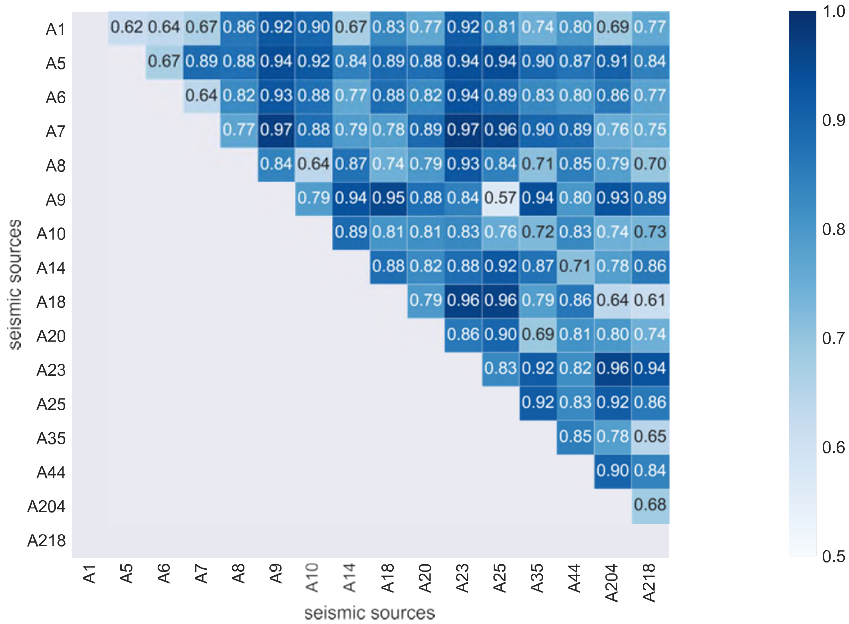

Figure 4. Averages of F-values for pairs of seismic sources after feature reduction.

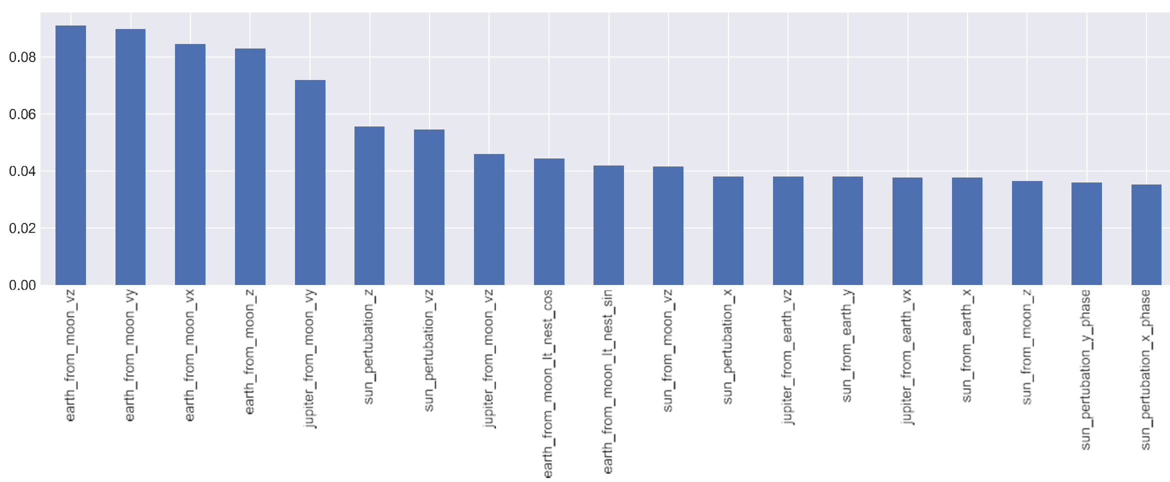

Figure 5. Averages of contribution ratios for each feature after feature reduction.

V. CONCLUSION

In this paper, we proposed a general framework of explanation necessary to widely promote implementation of analytical applications using social big data. The procedure of a tourism application based on integrated data model was described as an example of a macro explanatory function. In addition, we used Balanced Random Forest as a micro explanatory function to extract features effective for the seismic source classification of the deep moonquakes from the temporal and spatial features of the planets. We will develop a micro explanatory function showing the basis of individual decisions in analysis and complete the whole explanation framework and at the same time we will verify the versatility of the explanatory framework by applying it to a wider variety of use cases in the future.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 16K00157, 16K16158, and Tokyo Metropolitan University Grant-in-Aid for Research on Priority Areas Research on social big data.

REFERENCES

[1] D. Baehrens et al., “How to explain individual classification decisions,” The Journal of Machine Learning Research, vol.11, pp. 1803-1831, August 2010.

[2] L. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5-32, 2001.

[3] C. Chen, A. Liaw, and L. Breiman, “Using random forest to learn imbalanced data,” University of California, Berkeley, pp. 1-12, 2004.

[4] F. D. Velez and B. Kim, “A roadmap for a rigorous science of interpretability,” pp. 1-13, 2017 (arXiv: 1702.08608, 2017).

[5] EU GDPR, https://eugdpr.org/ [retrieved: March, 2019].

[6] L. Hendricks et al., “Generating visual explanations,” Proc. European Conference on Computer Vision, pp. 3-19, Springer, 2016.

[7] M. Hirota, K. Saeki, Y. Ehara, and H. Ishikawa, “Live or Stay ?: Classifying Twitter Users into Residents and Visitors,” Proc. International Conference on Knowledge Engineering and Semantic Web (KESW 2016), pp. 1-2, 2016.

[8] H. Ishikawa, Database, Mori Kita Publishing, 2008 (in Japanese).

[9] H. Ishikawa, Social Big Data Mining, CRC Press, 2015.

[10] H. Ishikawa, D. Kato, M. Endo, and M. Hirota, “Generalized Difference Method for Generating Integrated Hypotheses in Social Big Data,” Proc. ACM MEDES International Conference, pp. 13-22, 2018.

[11] H. Ishikawa and M. Hirota, S. Yokoyama, Social Big Data Practiced with Full Stack JavaScript and Python Machine Learning Library, Corona Publishing, 2019 (in Japanese).

[12] K. Kato, R. Yamada, Y. Yamamoto, M. Hirota, S. Yokoyama, and H. Ishikawa, “Investigation of Orbit Parameters to Classify the Deep Moonquake Sources,” Journal of Space Science Informatics Japan, vol. 7, pp. 43-52, 2018 (in Japanese).

[13] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, “Human-level concept learning through probabilistic program induction,” Science vol.350, issue 6266, pp. 1332-1338, 2015.

[14] Z. C. Lipton, “The Mythos of Model Interpretability,” Communications of the ACM, vol. 61, no. 10, pp. 36-43, October 2018.

[15] D. Martens, B. Baesens, T. V. Gestel, and J. Vanthienen, “Comprehensible credit scoring models using rule extraction from support vector machines,” Rule extraction from support vector machines, pp. 33-63, 2008.

[16] K. Mitomi, M. Endo, M. Hirota, S. Yokoyama, Y. Shoji, and H. Ishikawa, “How to Find Accessible Free Wi-Fi at Tourist Spots in Japan,” Volume 10046 of Lecture Notes in Computer Science, pp. 389-403, 2016.

[17] NAIF, https://naif.jpl.nasa.gov/naif/ [retrieved: March, 2019].

[18] NRI, “information bank” acceptable to consumers?,” NRI Journal [retrieved: March, 2019] (in Japanese).

[19] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” Proc. CHI 2016 Workshop on Human Centered Machine Learning, pp. 1135-1144, 2016 (arXiv: 1602.04938v1 [cs.LG] 16 Feb 2016).

[21] S. Seabold and J. Perktold. “Statsmodels: Econometric and statistical modeling with python,” Proc. 9th Python in Science Conference, pp. 57-61, 2010.

[22] Z. Si and S. C. Zhu., “Learning and-or templates for object recognition and detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 9 pp. 2189-2205, 2013.

[23] W. V. Siricharoen, “Infographics: The New Communication Tools in Digital Age,” Proc. International Conference on E-Technologies and Business on the Web, pp. 169-174, 2013.

[24] D. Smith, R. St. Andre, and M. Eggen, A Transition to Advanced Mathematics, Brooks/Cole Pub Co., 2014.

[25] K. W. Su, C. L. Liu, and Y. W. Wang, “A principle of designing infographic for visualization representation of tourism social big data.” J Ambient Intell Human Comput, pp. 1-21, 2018, doi:10.1007/s12652-018-1104-9.

[26] T. Tsuchida, D. Kato, M. Endo, M. Hirota, T. Araki, and H. Ishikawa, “Analyzing Relationship of Words Using Biased LexRank from Geotagged Tweets,” Proc. ACM MEDES International Conference, pp. 42-49, 2017.

 ISBN: 978-1-61208-697-2