APREP-DM: a Framework for Automating the Pre-Processing of a Sensor Data Analysis based on CRISP-DM

1st Hiroko Nagashima
Graduate School of Science
Tokyo Woman's Christian University
Tokyo, Japan
h18m001@cis.twcu.ac.jp

2nd Yuka Kato
Graduate School of Science
Tokyo Woman's Christian University
Tokyo, Japan
yuka@lab.twcu.ac.jp

***Abstract*—The need for analyzing data is increasing at an unprecedented rate. Well-known examples include customer behavioral patterns in shops, the autonomous motion of robots, and fault prediction. Pre-processing of data is essential for achieving accurate results. This includes detecting outliers, handling missing data, and data formatting, integration, and normalization. Pre-processing is necessary for eliminating ambiguities and inconsistencies. We here propose a framework called APREP-DM (for the Automated PRE-Processing for Data Mining) applicable to data analysis, including using sensor data. We evaluate two types of perspectives: (1) considering pre-processing in a test-case scenario involving pedestrian trajectory tracking, and (2) comparing APREP-DM with the outcomes of other existing frameworks from four different perspectives. We conclude that APREP-DM is suitable for analyzing sensor data.**

***Index Terms*—IoT, Sensor data, Data mining, Data Cleaning, Pre-Processing**

I. Introduction

It has recently become possible to analyze integrated data obtained from sensors or wearable devices in addition to using data on existing systems. Large amounts of various types of data can now be analyzed from multiple perspectives. Examples include the analysis of customer behavioral patterns in retail environments, the autonomous motion of robots, and the detection of fraud when using credit cards. However, the accuracy of such analyses is limited when importing raw data directly into the calculations performed by business intelligence tools, owing to the occurrence of data outliers, missing data, inconsistencies in units and in device specifications, or ambiguities within the data. To achieve the desired purpose, it is therefore necessary to duly consider the appropriate data conditions and formats of the generated data, within the specific context of a given analysis method.

Earlier studies have proposed frameworks for conducting data-mining tasks, covering the full work cycle, starting with the beginning of a project, to model maintenance [1]. A general overview is presented in Fig.1. Data analysts initially consider the aim of the project and transform the data accordingly before then evaluating the analysis model. Finally, they provide the necessary documentation. The transformation step, also called pre-processing, is the most time-consuming because of (1) the large quantity and variety of data; (2) the diversification of methods for data analysis; and (3) the large number of pre-processing tasks required. We therefore sought to reduce the pre-processing time by automating parts of the data-mining process.

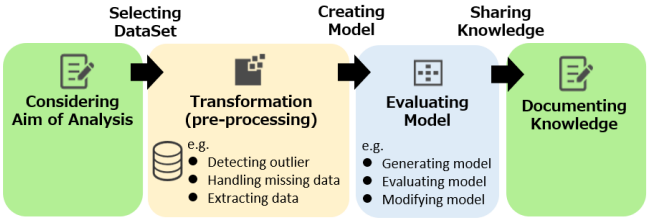

Fig. 1. Data-mining workflow. Green denotes an interaction with the client, yellow a data transformation, and blue the analysis model.

We here propose a new data mining framework called APREP-DM (Automated PRE-Processing for Data Mining) for controlling the automated tasks that feature in the pre-processing steps in sensor-data analyses. These processes are necessary owing to the noise and missing data that almost inevitably occur in sensor-derived data. Data from sensors through network sometimes delay or not to receive. We might receive data after long later if a network delay occurs. That is we have to check the outliers and missing data by ourselves, and modify them by any methodology. These processes take a long time without APREP-DM, that ought to use 80% resource in the analysis system [2]. As obtaining relevant information is a pre-requirement to pre-processing, this framework design is based on CRISP-DM [3], a well-known framework for data mining.

The contributions of this paper include:

- Determining whether or not pre-processing and data-cleaning tasks can be automated.
- Verifying the effectiveness of APREP-DM in a sensor-data analysis using a test-case scenario.

978-1-5386-9151-9/19/$31.00 ©2019 IEEE

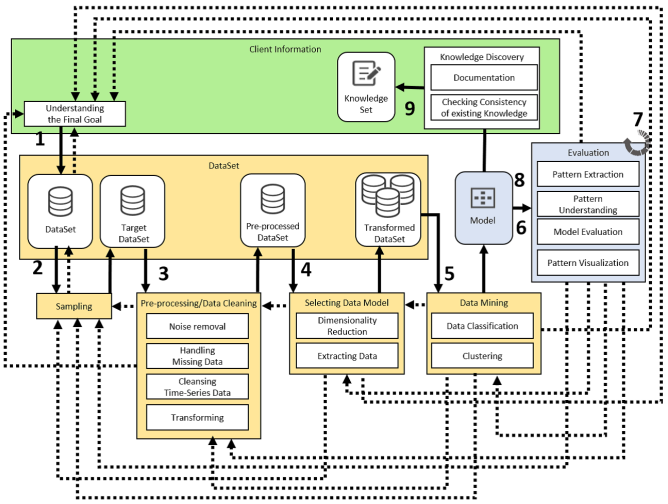

Fig. 2. Overview of KDD framework (drawing based on [4], Figure 1).

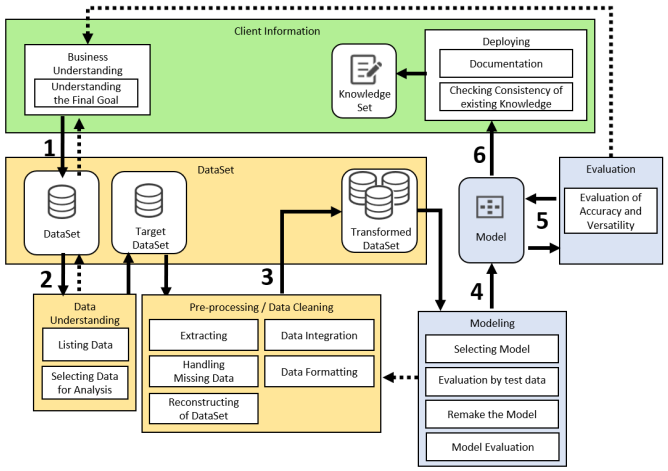

Fig. 3. Overview of CRISP-DM framework (drawing based on [3])

Section II provides a brief overview of the relevant literature. Section III outlines the principle and provides some details on APREP-DM. Section IV evaluates APREP-DM from two perspectives, firstly by performing the analysis on the test scenario using sensor data and secondly by comparing APREP-DM with previously established frameworks. Section V closes with some concluding remarks.

II. Related Work

We focus on four familiar frameworks used for data mining: Knowledge Discovery in Database (KDD)[4][5], CRoss Industry Standard Process for Data Mining (CRISP-DM) [3], and Sample, Explore, Modify, Model and Assess (SEMMA) [6]

The features of these frameworks and the specific steps they involve are described as follows.

A. KDD

KDD (Fayyad et al., 1996), the oldest data-mining framework, involves nine steps in one cycle. A notable feature is that a step in the process may be repeated if necessary. The process steps, depicted in Fig.2, are the following:

1) Learning the application domain: Understanding the application domain and the business aim of the analysis.
2) Creating a target dataset: Extracting data or sampling to generate a target dataset for analysis.
3) Data cleaning and pre-processing: Removing noise, mapping missing data, or transforming time-sequence information.
4) Data reduction and projection: Identifying data trends by dimensionality reduction or using transformation values, i.e., data reduction or data projection.
5) Choosing the function of data mining: Selecting the model to achieve the final goal of the analysis among data integration, classification, or clustering.
6) Choosing the data mining algorithm: Evaluating the model and considering the analysis model.
7) Data mining: Running the model, e.g. using regression or clustering.
8) Interpreting: Understanding the results and visualizing patterns and the model.
9) Using the acquired knowledge: Documenting the outcome and checking for conflicts with earlier results.

Within the KDD framework, steps may be repeated and improved through successive iterations. However, this can increase the complexity significantly owing to the need to consider the return points at each step. The KDD framework can therefore be very time-consuming.

B. CRISP-DM

CRISP-DM was proposed by CRISP-DM consortium of companies (such as NCR, SPSS, and DaimlerChrysler) that perform data mining. There are six steps in one cycle. Its features are (1) an initial clarification of the priority and end-goal criteria, and (2) the inclusion of iterations between "business understanding" and "data understanding" and between "data preparation" and "modeling". The framework is illustrated in Fig. 3 and comprises the following steps:

1) Business understanding: Clarifying the client's aim and defining the priority and success criteria.
2) Data understanding: Understanding the data used within the project and evaluating the data if required.
3) Data Preparation: Performing the necessary data transformations, e.g., extracting target data, handling missing data, and reconstruction of the dataset.
4) Modeling: Selecting the model, e.g., decision tree, or neural network.
5) Evaluation: Using an application to evaluate the model accuracy and versatility.
6) Deploying: Summarizing the process and sharing knowledges.

CRISP-DM involves a step where the client's priority and the success criteria of the project based on "business understanding" are decided. On the other hand, this framework

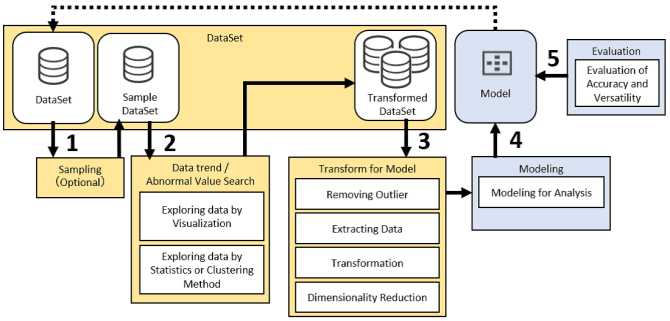

Fig. 4. Overview of SEMMA framework (drawing based on [6])

does not treat outliers in the data-preparation step. Thus, the potential impact of outliers on the project outcome makes CRISP-DM unsuitable for sensor-data analyses. Although the CRISP-DM framework does not treat outliers, CRISP-DM consortium consider the outliers and missing data. CRISP-DM is sometimes used for other data mining framework base like APREP-DM because it does not depend on specific products. e.g., Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM) by IBM[7][8].

C. SEMMA

SEMMA, a framework proposed by SAS Institute, involves five steps in one cycle. This framework was designed for a data-mining product of SAS Institute called "SAS Enterprise Miner". The steps are therefore classified according to the product's functions. The features of this framework are (1) data extraction for random sampling and (2) the exploration of data trends. SEMMA is often used by enterprises. The steps forming SEMMA are outlined in Fig. 4 and involve:

1) Sampling: The extracting of a portion of a large data set is done by random sampling. This is an optional step.
2) Exploration: Understanding data trends by visualization, statistics, clustering, etc.
3) Modification: Adding new items, extracting or transforming data; if necessary, removing outliers or reducing dimensionality.
4) Modeling: Making a model for analysis methods, e.g., decision trees or neural networks.
5) Assessment: Evaluating the model in terms of usability, reliability, or accuracy.

SEMMA includes a data-sampling step when dealing with a large dataset and is therefore amenable to a trial-and-error approach to data mining. On the other hand, it does not include the "business understanding" and "deploying" steps. Client information and knowledge must therefore be managed outside the framework.

III. APREP-DM

APREP-DM involves automating parts of the pre-processing of sensor data, including common data-cleaning tasks such as detecting outliers and handling missing data.

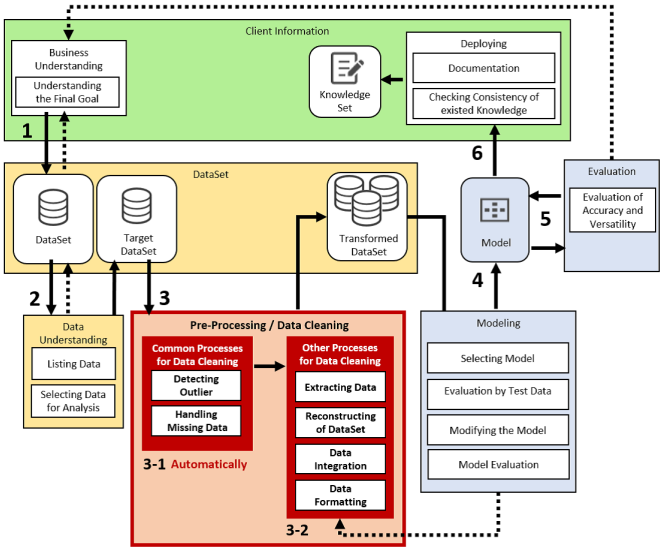

Fig. 5. Overview of APREP-DM framework

A. Overview

We focused on the pre-processing step of the analysis model, which is divided into two sub-steps. One is common data-cleaning sub-step (e.g., detecting outliers or handling missing data); the other is transformation sub-step (e.g., extracting or integrating data, reconstruction, or adding new items) generate to import data into the analysis model. The common data-cleaning sub-step can run based on statistics or by clustering or classifying items. Therefore we define the common data-cleaning sub-step is composed automatically tasks. The other sub-step needs find a suitable model by trial-and-error for the aim of the analysis result. Therefore we define the other sub-step cannot be automatically. The business aims and the criteria for priority and success (in the "business understanding" step) can be used to determine the criteria for defining outliers and the way by which missing data can be filled. Thus, APREP-DM includes "business understanding" step before "detecting outliers" and "handling missing data" steps.

As mentioned in Section I, APREP-DM follows CRISP-DM in that it does not depend on specific products. Three iterations are used in APREP-DM: "business understanding" and "data understanding", "pre-processing" and "modeling", and "business understanding" and "evaluation". Details are provided in Section III-B.

B. Details

APREP-DM involves firstly a "business understanding" step, followed by "data understanding", to select the target data to be analyzed in a sensor-data project. The steps of the process are outlined in Fig.5, and include:

1) Business understanding: Clarifying the client's aim and the criteria defining priority and success.
2) Data understanding: Understanding the data used within the project and evaluating the data if required.

3) Data Preparation
   (3-1) Performing common data-cleaning tasks automatically, e.g., detecting outliers and handling missing data. This sub-step is automatically.
   (3-2) Performing the relevant transformations, e.g., extracting data, reconstructing of the dataset, to create a model suited to the final goal.
4) Modeling: Selecting a model, e.g., decision tree, or neural network.
5) Evaluation: Using applications to evaluating the model accuracy and versatility.
6) Deployment: Summarizing the process and sharing knowledges.

The two pre-processing steps, (3-1) and (3-2), are not performed simultaneously. Common data-cleaning tasks (e.g., detecting outliers or handling missing data) are performed first. Other data-cleaning processes (e.g., reconstructing of dataset, data integration, or data formatting) are performed subsequently. Being able to decide conditions for outliers in business understanding are important for running automatic step (3-1). We cannot use APREP-DM unless we can define the conditions of outliers.

As mentioned in Section II-B, APREP-DM is based on CRISP-DM, which does not involve detecting and removing outliers during pre-processing. Nonetheless, outliers impact the statistics significantly and should therefore be detected early. This is especially important when dealing with sensor-derived data.

One feature of APREP-DM is that it does not remove outliers during pre-processing. This is because outliers can be important for detecting anomalies or unexpected behavior. Thus, APREP-DM detects outliers, and removes them if necessary.

IV. Evaluation of APREP-DM

We evaluated the proposed framework from two perspectives:

1) A scenario evaluation, involving analyzing a system using sensor data.
2) A qualitative evaluation, comparing APREP-DM with other existing frameworks.

A. Scenario evaluation

We clarify pre-processing processes that can automatically as uses a scenario evaluation.

1) Evaluation approach: We considered a scenario while analyzing the data derived from sensors, with the aim of defining the tasks to include in pre-processing. The scenario involved analyzing customer behavioral using multiple three-dimensional (3D) range-imaging sensors:

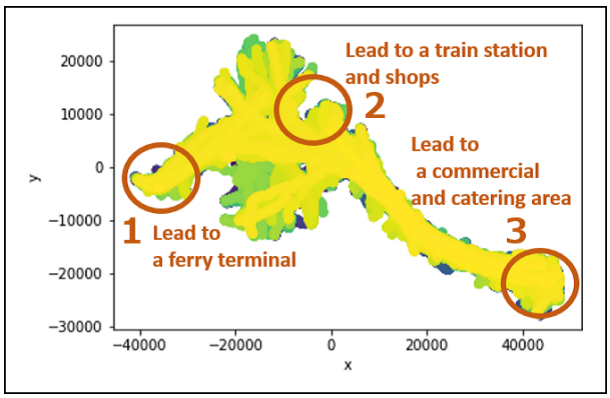

Fig. 6. Map of a shopping mall using 3D range-imaging sensor data

A system aimed at predicting a customer behavioral in a shopping mall with three exits, using data describing where an individual customer exits the mall depending on their point of entry. Based on this result, we deliver suitable coupons of the shopping mall by a push function on a mobile application.

We used a dataset comprising multiple range images obtained using 3D sensors [9] and also a meteorological dataset[10]. The sensor dataset monitored the walking trajectories of shopping-mall customers (40,851 per day on average, including duplicates). The data were gathered for 92 days over approximately one year (between 9:40 and 20:20 every Wednesday and Sunday from October 2012 to November 2013). The customers' locations were measured continuously at a rate of 10 - 40 Hz using multiple 3D range-image sensors. The floor plan of the shopping mall is depicted in the upper part of Fig. 6. The data, stored in CSV format, contained the following elements: UNIX time, *person_id*, position *pos_x*, *pos_y*, *hight* [mm], *velocity* [mm/s], *body_angle* of motion [rad], and *facing_angle* [rad]. The meteorological dataset comprised information for Osaka, where the shopping mall was located. Data were downloaded for the period covering October 2012 to November 2013, including the *date*, *temperature[°C]*, and *weather* parameters.

The shopping mall has three exits, leading (1) to a ferry terminal, (2) to a train station, shops, and offices via escalators and elevators, and (3) to a commercial and catering area on the eastern side. Machine learning was performed using a support vector machine (SVM), currently one of the most commonly-used pattern-recognition models. The features selected for the SVM were the weather, the entry location of the customers, a flag specifying whether or not the data were recorded on a weekday, and the averages of velocity. These data were then used to predict the exit location of an individual customer.

2) Evaluation Result: The sensor and meteorological data must be combined under cleaned condition before being input into the SVM (Fig. 7). We checked data cleaning tasks, and verified APREP-DM processes.

We first considered the pre-processing of the sensor data.

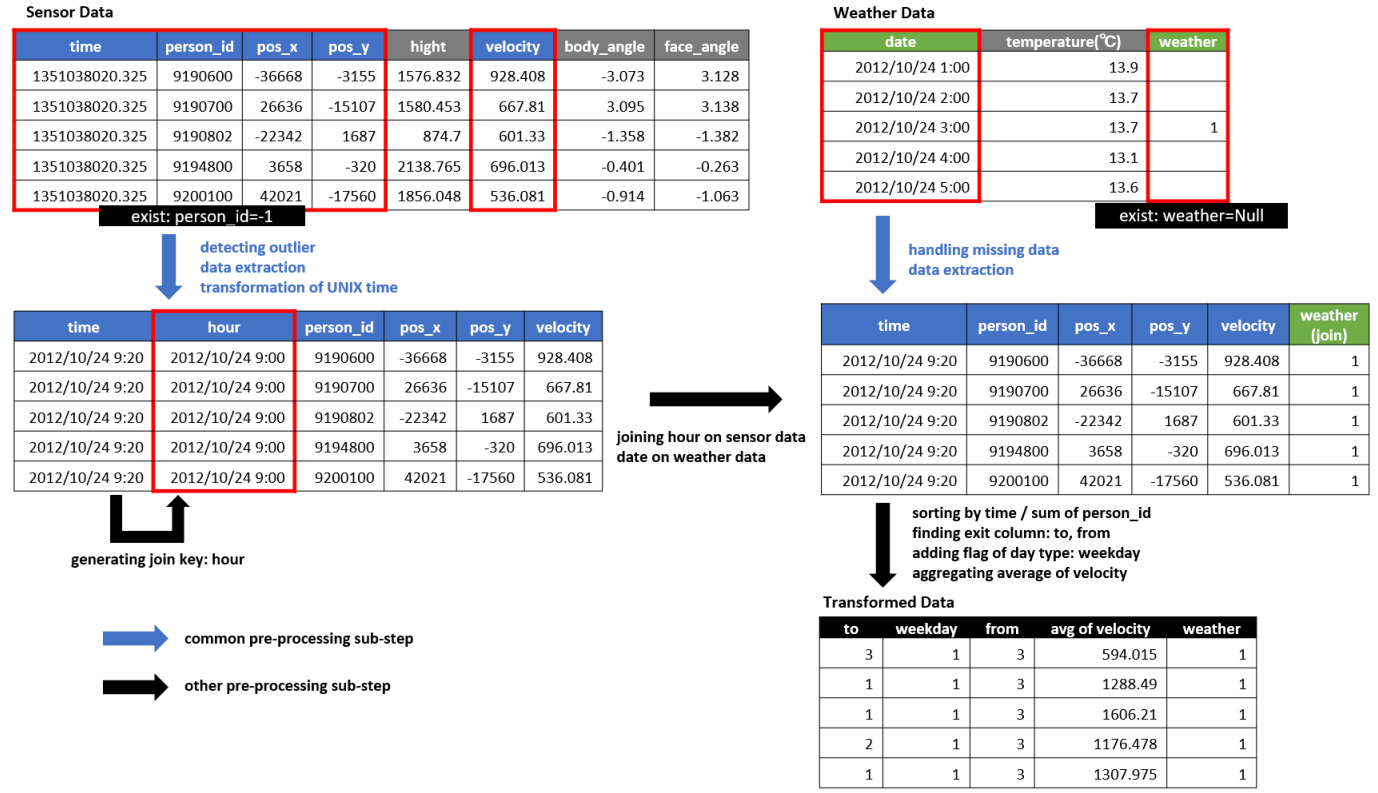

Fig. 7. Pre-processing example

Some outliers were identified, e.g., "*person_id = -1*" . As this project was not concerned with anomaly analysis, we removed the outliers rows from the dataset and extracted the required columns of data, i.e., *time*, *person_id*, *pos_x*, *pos_y*, and *velocity*. We then transformed the *time* from the UNIX time to date format [yyyy/mm/dd HH:mm:ss] in order to emulate the corresponding format in the meteorological dataset, namely [yyyy/mm/dd HH:mm]. In order to associate the sensor and meteorological data, we truncated all the dates into a unique format [yyyy/mm/dd HH:00]. We next considered the pre-processing of the meteorological data. This dataset contained some missing data in the *weather* column. We identified the rows with defined *weather* values and extracted the required columns, namely *date* and *weather*. We then associated the above sensor and meteorological data. Finally, we added a flag denoting whether or not a particular day was a *weekday*, and included the entry and exit locations of the customers' *exit*, as well as the aggregated the values of *velocity* per *person_id* to calculate the average.

In summary, the purpose of this scenario was to demonstrate the following pre-processing tasks:

1) Data understanding (format, contents of items).
2) Pre-processing.
   a) Removing irrelevant data.
      (i) Extracting the required data from the dataset.
      (ii) Transforming the dataset to take into account outliers and missing data entries.
      (iii) Unifying data units.
   b) Data integration, aggregation, and transforming data for analysis.
      (i) Creating association keys to join different datasets.
      (ii) Aggregating and calculating data.
      (iii) Generating necessary data. e.g., exit flag, and weather flag.

Following to apply this result to APREP-DM flow in III-B, 1) is the "business understanding" and "data understanding", 2a) is the "common data-cleaning sub-step", and 2b) is the "other data-cleaning sub-step".

Thanks to the pre-processing done for this quantitative, for the data obtained on Wednesday, November 14th, 2012, the number of rows in the dataset decreased from approximately 57 million to only 427.

B. *Qualitative evaluation*

1) Method of evaluation: The analysis model requires that the input data be suitably integrated by associating the appropriate columns. Furthermore, it is necessary to detect and remove outliers when using sensor data, and to include business understanding when deciding which parts of the pre-processing to automate. Finally, it is important that each step be easily iterated to improve the accuracy of the analysis model, so we evaluate about iteration.

The target project considered here involves an analysis of sensor-derived data. We evaluated the following aspects of APREP-DM and of earlier frameworks qualitatively:

- Adding data: Adding columns by the reconstruction, aggregate, or joining in the middle of the framework.
- Business understanding: Understanding the data specifically and clarifying the information needed to select the data required for the analysis.
- Small iteration: Iteration of each steps flexibility.
- Outlier detection: Detecting and (if necessary) removing outliers.

2) Evaluation Result: The results obtained for APREP-DM and for the earlier frameworks are compared in TABLE I.

Adding Data: KDD and SEMMA narrow down from first selection data. CRISP-DM, and APREP-DM involve data integration during pre-processing.

Business Understanding: SEMMA does not involve a business understanding step. KDD, CRISP-DM, and APREP-DM do include a step in which the project aim is decided. Furthermore, CRISP-DM, and APREP-DM consider the priority of aim and criteria. CRISP-DM, and APREP-DM therefore have a more specific aim than KDD.

Small iteration: Although KDD can iterate any two steps, CRISP-DM, and APREP-DM regard the one cycle from the multiple steps of data mining flow. Therefore KDD is the smallest and the most flexible iteration among KDD, CRISP-DM, and APREP-DM. Although SAS Institute said that the iteration on data mining is natural, SEMMA does not step flow iteration clearly. So we decided to write the evaluation as "N/A".

Outlier Detection: CRISP-DM does not have any step of outliers in pre-processing. KDD, SEMMA, and APREP-DM have a process for outliers. Moreover, APREP-DM and SEMMA just detect outliers instead KDD removes outliers. Therefore, we can use APREP-DM or SEMMA abnormal analysis projects.

C. Discussion

We evaluated the scenario by confirming the pre-processing steps.

- Clarifying business understanding (aim and criteria) is essential when selecting data from the dataset.
- Data must be suitably cleaned for input into the analysis model. (When analyzing customers' trajectories using sensor data, we did not use the row-data directly.)
- Different datasets must be associated using appropriate joining keys, if necessary.

Following the result of the scenario evaluation, we performed a qualitative comparison of APREP-DM with earlier frameworks. We therefore conclude that APREP-DM is a well-balanced framework that is suited to analyzing sensor data.

TABLE I
COMPARISON WITH EARLIER FRAMEWORKS

Framework Name	Adding Data	Business Understanding	Small Iteration	Outlier Detection
KDD	+	++	+++	++
CRISP-DM	+++	+++	++	+
SEMMA	+	+	N/A	+++
APREP-DM	+++	+++	++	+++

The number of "+" symbols indicates the degree of adequacy.

V. CONCLUSION

We proposed using the APREP-DM framework to automate parts of data pre-processing. This is achieved by classifying commonly used tasks and specific processes for data cleaning. We evaluated APREP-DM using a data-analysis scenario and compared the results with the outcomes of earlier frameworks. We then verified that APREP-DM performed a superior sensor-data analysis.

This paper conclusions are the following:

- Defined automated data-transformation tasks (e.g., to deal with outliers and missing data within datasets) during pre-processing by extending the CRISP-DM framework.
- Clarified the importance of understanding the aim of the analysis and the success criteria before pre-processing as well as considering previous frameworks.
- By considering a sensor-data analysis scenario and by comparing the outputs obtained with earlier frameworks, we verified the efficacy of the proposed APREP-DM framework.

Further studies will be needed to verify by building system involved APREP-DM.

REFERENCES

[1] A. Azevedo and M. Filipe Santos, "KDD, semma and CRISP-DM: A parallel overview," 2008, pp. 182–185.
[2] S. Gulwani. Program by examples (and its applications in data wrangling). [Online]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/12/pbe16.pdf
[3] Cross Industry Standard Process for Data Mining Consortium. CRISP-DM by smart vision europe. [Online]. Available: http://crisp-dm.eu/reference-model/
[4] Fayyad, Usama and Piatetsky-Shapiro, Gregory and Smyth, Padhraic, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, no. 3, p. 37, 1996.
[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for Extracting Useful Knowledge from Volumes of Data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, 1996.
[6] SAS Institute Inc. SAS Enterprise Miner. [Online]. Available: https://web.archive.org/web/20120308165638/http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html/
[7] International Business Machines Corporation. Have you seen ASUM-DM? - SPSS Predictive Analytics. [Online]. Available: https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/
[8] International Business Machines Corporation, "Analytics Solutions Unified Method Implementations with Agile principles," *IBM Analytics sevice Datasheet*, 2016.
[9] D. Brscic, T. Kanda, T. Ikeda and T. Miyashita, "Person Tracking in Large Public Spaces Using 3-D Range Sensors," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 6, pp. 522–534.
[10] T. Ministry of Land, Infrastructure and Tourism. Japan meteorological agency. [Online]. Available: https://www.data.jma.go.jp/gmd/risk/obsdl/index.php