Maximum Likelihood Evidential Reasoning-Based Hierarchical Inference with Incomplete Data

Xi Liu, Swati Sachan, Jian-Bo Yang, Dong-Ling Xu
Decision and Cognitive Sciences Research Centre
The University of Manchester
Manchester, UK
xi.liu-2; swati.sachan; jian-bo.yang; Ling.Xu@manchester.ac.uk

Abstract—Data mining requires a pre-processing task where data are prepared, cleaned, integrated, transformed, reduced and discretized to ensure data quality. Incomplete data are commonly encountered during data cleaning, which can have major impact on the conclusions that will be drawn from the data. In order to effectively carry out inferential modelling or decision making from incomplete independent variables or explanatory variables and consider different types of uncertainties, this paper adopts a data-driven inferential modelling approach, Maximum Likelihood Evidential Reasoning (MAKER) framework, which takes advantage of incomplete datasets without any imputation that may be required by other conventional machine learning methods. The MAKER framework reflects the plausibility of different values of missing data and expresses data-driven support for different values of missing data.

Index Terms—Data-Driven Inference, Evidence Reasoning, Incomplete Data, Rule Extraction, Decision Making

I. INTRODUCTION

Data in the real world is rarely clean and uniform. Often, they tend to be incomplete, noisy, and inconsistent, and pre-storing data by populating missing values is an important task. Processing data is very important as there is the risk that it can lead to mispredictions or classifications of any given model used.

It is stated in [1] that it is appropriate to remove the data with missing values if the missing data are ‘missing (complete) at random’ and the observed data are ‘observed at random’. However, when missing values depend on hypothetical values or some other variables’ values, single imputation methods, such as mean or median substitution, and last observation carried forward (LOCF), have been frequently used but not recommended, due to the problem of biased distribution and increased noise. In order to deal with such issues, a more commonly used practice of multiple imputation was initially proposed by [2], which has been widely promoted and is now the basic solution method for incomplete dataset problems, see [3], [4] and among others. A variety of multiple imputation methods solve this problem by randomly drawing multiple imputations from the imputation distribution and also by introducing an additional error variance to each imputation [5]. Depending on the type of imputed variables (i.e. continuous, ordinal, nominal) and missing data patterns (i.e. monotone, non-monotone), logistic regression, discriminant regression, discriminant regression and Markov chain Monte Carlo (MCMC) are commonly used models of imputation. The validity of these methods depends on the cause of missing data (missing data mechanism); the pattern and extent of missing data; whether outputs or explanatory variables are missing.

Bayesian approaches have also been introduced to deal with missing data by treating the missing data as latent variables and thus integrate them out from the joint density of the observed values and the missing values [6]. The model of missingness is fitted simultaneously as a joint model, which ensures the consistency of estimation. It is not an appropriate method when a set of multiple pieces of evidence is not combined based on their arrival. Meanwhile, it is also impractical for arbitrarily large dimensional problems. In certain settings, the imputation method is simply an approximation of the Bayesian approach.

In order to effectively carry out referential modelling from incomplete independent variables or explanatory variables and take different types of uncertainties into account, this paper adopts a data-driven inference modelling approach, Maximum Likelihood Evidential Reasoning (MAKER) framework, which can take advantage of incomplete datasets without any imputation that may be required by other conventional machine learning methods. A general optimal learning model is constructed, where the reliability of evidence and probabilistic prediction can be learnt from historical data by maximising the likelihood of true state. Unlike Bayes rules, MAKER framework does not depend on prior probability. Prior, however, can be treated as independent evidence and used to incorporate additional information into the evidence set.

The rest of paper is structured as follows. Section II introduces the dataset built into the MAKER framework. In Section III, we briefly review the MAKER framework and elaborate the inference process in the case of incomplete data. An example of insurance fraud detection in Section IV is proposed to illustrate the performance of the proposed framework. Section V presents the conclusions and discussions.

II. DATASET

Suppose, \( H = \{h_1, \ldots, h_n, \ldots, h_N\} \) is a set of mutually exclusive and collectively exhaustive outputs, with \( h_i \cap h_j \neq \emptyset \) for any \( i \neq j \), with \( \emptyset \) being the empty set. \( H \) is referred as
TABLE I: Example of a complete and unambiguous dataset

<table>
<thead>
<tr>
<th>Attributes</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1,1</td>
<td>V1,2</td>
<td>V2,1</td>
<td>V3,1</td>
<td>h1</td>
</tr>
<tr>
<td>V1,1</td>
<td>V1,2</td>
<td>V2,2</td>
<td>V3,2</td>
<td>h2</td>
</tr>
<tr>
<td>V1,1</td>
<td>V1,2</td>
<td>V2,3</td>
<td>V3,3</td>
<td>h3</td>
</tr>
<tr>
<td>V1,1</td>
<td>V1,3</td>
<td>V2,1</td>
<td>V3,1</td>
<td>h1</td>
</tr>
<tr>
<td>V1,1</td>
<td>V1,3</td>
<td>V2,2</td>
<td>V3,2</td>
<td>h2</td>
</tr>
<tr>
<td>V1,1</td>
<td>V1,3</td>
<td>V2,3</td>
<td>V3,3</td>
<td>h3</td>
</tr>
</tbody>
</table>

A frame of discernment. The set of independent variables can be represented as:

\[ \mathcal{V} = \{V_1, \ldots, V_k, \ldots, V_K \}, \ k \in \{1, \ldots, K\} \quad (1) \]

where the \( k \)th attribute \( V_k \) contains \( C_k \) (\( k \in \{1, \ldots, K\} \)) categories, which is represented by \( V_{f,k} \) (\( f \in \{1, \ldots, C_k\} \)).

The MAKER framework can be implemented on complete/incomplete and unambiguous/ambiguous datasets. A dataset is called complete if it has full records of input attributes, while an unambiguous dataset indicates that the output is complete. Table I shows an example of a complete and unambiguous dataset, where all input variables (attributes) \( V_1, V_2 \) and \( V_3 \) are complete, so as the output. \( V_1, V_2 \) and \( V_3 \) are also set to contain 4, 3, and 4 categories, respectively. Table II shows an example of an incomplete and ambiguous dataset. Missing outputs are labelled as ‘NA’, where ‘NA’ \( \subseteq \mathbb{H} \).

For further utilizing the MAKER framework, this dataset is partitioned into five sub-tables (Tables IIa - Iic) to include all data records for inference without imputation or deletion. Tables IIa - Ila contain complete records of attributes \( V_1, V_2 \) and \( V_3 \), separately, which can be used to calculate the basic probability in (3). Tables IIb and Iic contain complete records for the combination of \( V_1 \) and \( V_2 \) and the combination of all three input variables, which can be further utilized to obtain the joint basic probability in (4). Note that, there is no requirement for a complete dataset, e.g. Table I, to be split into sub-tables, as data records will remain the same. Once (joint) basic probabilities are obtained, the interrelationship between the pair of input variables can be further derived by adopting (8).

III. MAKER FRAMEWORK FOR INCOMPLETE DATA

A. MAKER framework

Let \( e_{f,k} \) denote the \( f \)th piece of evidence from the \( k \)th input variable \( V_k \) at \( V_k = V_{f,k} \) and \( e_{f,k}(h) \) be an element of evidence \( e_{f,k} \) pointing exactly to proposition. In the MAKER framework, pieces of evidence obtained from an input variable set \( \mathcal{V} \) are combined together to generate the combined support for a proposition. Multiple pieces of evidence are joined together by considering four elements: basic probability [7], the interrelationship between each pair of evidence, the reliability and weight of each piece of evidence.

TABLE II: Example of an incomplete and ambiguous dataset

(a) Sub-table 1

<table>
<thead>
<tr>
<th>Attributes</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1,1</td>
<td>h1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V1,1</td>
<td>h2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V1,1</td>
<td>h2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V1,1</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V1,1</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V1,1</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) Sub-table 2

<table>
<thead>
<tr>
<th>V2</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>V2,1</td>
<td>h1</td>
</tr>
<tr>
<td>V2,2</td>
<td>h2</td>
</tr>
<tr>
<td>V2,2</td>
<td>h2</td>
</tr>
<tr>
<td>V2,1</td>
<td>NA</td>
</tr>
<tr>
<td>V2,1</td>
<td>NA</td>
</tr>
<tr>
<td>V2,1</td>
<td>NA</td>
</tr>
</tbody>
</table>

(c) Sub-table 3

<table>
<thead>
<tr>
<th>V3</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>V3,1</td>
<td>h1</td>
</tr>
<tr>
<td>V3,2</td>
<td>h3</td>
</tr>
<tr>
<td>V3,2</td>
<td>h3</td>
</tr>
<tr>
<td>V3,1</td>
<td>NA</td>
</tr>
<tr>
<td>V3,1</td>
<td>NA</td>
</tr>
<tr>
<td>V3,1</td>
<td>NA</td>
</tr>
</tbody>
</table>

A piece of evidence \( e_{f,k} \) is profiled by a belief distribution or basic probability distribution as follows:

\[ e_{f,k} = \left\{ (e_{f,k}(h), p_{h,f,k}), \forall h \subseteq \mathbb{H} \right\} \quad \text{and} \quad \sum_{h \subseteq \mathbb{H}} p_{h,f,k} = 1 \quad (2) \]

where \( p_{h,f,k} \) denotes the basic probability that evidence \( e_{f,k} \) points to proposition \( h \), which is obtained from an input variable \( V_k \).

The evidence from single input variable \( V_k \) can be transformed to a basic probability distribution for all \( h \subseteq \mathbb{H} \). Let \( l_{h,f,k} \) be the likelihood of observing \( f \)th piece of evidence from the \( k \)th input variable \( V_k \) given proposition \( h \), the basic probability, obtained from a normalized likelihood, is given by:

\[ p_{h,f,k} = p(e_{f,k}(h)) = \frac{l_{h,f,k}}{\sum_{X \subseteq \mathbb{H}} l_{X,f,k}} \quad \forall h \subseteq \mathbb{H}, \quad (3) \]

which is called one-dimensional evidence acquisition process, where the likelihood is obtained by generating a contingency Table.

Similarly, joint basic probability can be acquired from a joint likelihood function following the Bayesian principle. Let
$e_{i,t}$ and $e_{j,m}$ be two pieces of evidence acquired from two input variables $V_t$ and $V_m$ at $V_t = V_{i,t}$ and $V_m = V_{j,m}$. Then, the joint basic probability that both evidence $e_{i,t}$ and $e_{j,m}$ point to proposition $h$ is given by:

$$p_{h,i,l,jm} = \frac{l_{h,i,l,jm}}{\sum_{X \subseteq H} l_{X,i,l,jm}} \quad \forall h \subseteq H,$$

(4)

where $l_{h,i,l,jm}$ is the joint likelihood that both $V_{i,t}$ and $V_{j,m}$ are observed given proposition $h$.

Note that joint basic probability given by (4) combines independent evidence under the assumption that all evidential elements are fully reliable. In reality, evidence is seldom fully reliable and evidence reliability hence needs to be taken into account to measure the degree of its support for proposition $h$. Let $r_{h,i,l} = p(h|e_{i,l}(h))$ be the reliability of evidence $e_{i,l}$ pointing to proposition $h$, for it is defined in [7] as the conditional probability to which a proposition $h$ is true given that $e_{i,l}$ points to $h$, and can be estimated by examining the process with how data is generated and how $e_{i,l}$ is acquired from the data. Furthermore, the reliability of evidence $e_{i,l}$ is:

$$r_{i,l} = \sum_{h \subseteq H} r_{h,i,l} p(h,i,l)$$

(5)

If $e_{i,l}$ and other evidence are acquired from the same data source, then the probability mass that proposition $h$ is supported by evidence $e_{i,l}$ is:

$$m_{i,l}(h) = p(h|e_{i,l}(h))p(e_{i,l}(h)) = r_{i,l} p(e_{i,l}(h)).$$

(6)

If $e_{i,l}$ is acquired from a data source that is different from other evidence, such as Table II, characterized by a probability function $p_{i,l}$ constructed from data $V_l$ only, the probability mass that proposition $h$ is supported by evidence $e_{i,l}$ is updated by [7] as follows:

$$m_{i,l}(h) = w_{h,i,l} p(e_{i,l}(h)).$$

(7)

where $w_{h,i,l} = a_{i,l} r_{i,l} p_{i,l}(h|e_{i,l}(h))$ is referred to as the weight of $e_{i,l}(h)$, which is proportional to the conditional probability to which $h$ is true given that $e_{i,l}$ points to $h$ measured by $p_{i,l}$, with $a_{i,l}$ a positive constant for all $h \subseteq H$. From (6) and (7),

$$w_{h,i,l} = r_{h,i,l} \quad \text{when} \quad p_{i,l} = p.$$

[7] defined an interdependence index to characterize the interrelationship between each pair of input variables when there are multiple inputs. Following the above (joint) basic probability, the degree of interdependence $I_{A,B,i,jm}$ between a pair of evidence $e_{i,l}(A)$ and $e_{j,m}(B)$, $A \cap B = h$, $h \subseteq H$ acquired from two different input variables $V_{i,t}$ and $V_{j,m}$ is formulated as follows:

$$I_{A,B,i,jm} = \begin{cases} 0 & \text{if } p_{A,i,l} = 0 \text{ or } p_{B,j,m} = 0, \\ \frac{p_{A,B,i,jm}}{p_{A,i,l} p_{B,j,m}} & \text{otherwise.} \end{cases}$$

(8)

where the single basic probabilities $p_{A,i,l}$, $p_{B,j,m}$ and the joint probability $p_{A,B,i,jm}$ are derived from the same joint-table, such as Table IIId.

The interdependence index has the following property:

$$I_{A,B,i,jm} = \begin{cases} 0 & \text{if } e_{i,t}(A) \text{ and } e_{j,m}(B) \text{ are disjoint,} \\ 1 & \text{if } e_{i,t}(A) \text{ and } e_{j,m}(B) \text{ are independent.} \end{cases}$$

(9)

### B. Conjunctive MAKER Rule

Considering to generate the combined probability $p_{h,i,l,jm}(h)$ that proposition $h$ is jointly supported by two pieces of evidence $e_{i,l}$ and $e_{j,m}$, a conjunctive MAKER rule is as follows:

$$p_{h,i,l,jm}(h) = \begin{cases} 0 & h = \emptyset \\ \frac{\tilde{m}_{i,l,jm}(h)}{\sum_{r \subseteq H} m_{i,l,jm}(r)+\tilde{m}_{i,l,jm}(P(H))} & h \subseteq H, h \neq \emptyset \end{cases}$$

(10)

where

$$\tilde{m}_{i,l,jm}(h) = [(1 - r_{j,m})m_{i,l}(h) + (1 - r_{i,l})m_{j,m}(h)]$$

(11)

is the combined probability mass of evidence $e_{i,l}$ and $e_{j,m}$ for proposition $h$. The residual support, which is earmarked to the power set $P(H)$, is given by

$$\tilde{m}_{il,jm}(P(H)) = m_{i,l}(H)m_{j,m}(H).$$

The parameter $\gamma_{A,B,i,l,jm} = r_{i,l,jm} r_{j,m} p_{i,l} p_{j,m}$ is the ratio of the joint reliability of the two pieces of evidence and the product of their individual reliabilities, referred to as reliability ratio for short. $r_{i,l}$, $r_{j,m}$, $w_{i,l}$, $w_{j,m}$ and $\gamma_{A,B,i,l,jm}$ in (11) are the parameters that need to be trained from input-output datasets for inference.

### C. An Example of Conjunctive MAKER Rule for Incomplete Data

This framework can recursively combine evidence from multiple input variables. Take for example with three input variables $V_1$, $V_2$, and $V_3$, with $C_k$ ($k \in \{1, 2, 3\}$) categories. The calculation processes are elaborated as follows:

- Using (3) to calculate the single basic probability from the Tables IIa - IIc for input variables $V_1$, $V_2$, and $V_3$: $p_1(e_{f_1,1}(h))$, $p_2(e_{f_2,2}(h))$, $p_3(e_{f_3,3}(h))$, $f_k \in \{1, \ldots, C_k\}$, $\forall h \subseteq H$.

- Calculate the single probability mass $m_{f_1,1}(h)$, $m_{f_2,2}(h)$, $m_{f_3,3}(h)$ by importing the above single probabilities into (7), $\forall h \subseteq H$.

- Derive the interrelationship $I_{h,f_1,f_2}$ in (8), by using the single basic probabilities $p(e_{f_1,1}(h))$, $p(e_{f_2,2}(h))$ and the joint basic probability $p_{h,f_1,f_2}$ calculated from Table IIId.

- Derive the interrelationship $I_{h,f_1,f_2,f_3}$ in (8), by using the single basic probabilities $p(e_{f_1,1}(h))$, $p(e_{f_2,2}(h))$, $p(e_{f_3,3}(h))$ and the joint basic probability $p_{h,f_1,f_2,f_3}$ calculated from Table IIe.
• The combined probability mass \( \tilde{m}_{f_1, f_2}(h) \) of the first two pieces of evidence \( e_{f_1, 1} \) and \( e_{f_2, 2} \) pointing to preposition \( h \) can be obtained by adopting (11):

\[
\tilde{m}_{f_1, f_2}(h) = [(1 - r_{f_2, 2}) m_{f_1, 1}(h) + (1 - r_{f_1, 1}) m_{f_2, 2}(h)] \\
+ \sum_{A \cap B = h} \gamma_{A, B, f_1, f_2} I_{A, B, f_1, f_2} m_{f_1, 1}(A)m_{f_2, 2}(B)
\]

• Similarly, the combined probability mass of evidence \( e_{f_1, 1} f_2, 2, f_3 \) pointing to preposition \( h \) can be derived by adopting (11):

\[
\tilde{m}_{f_1, f_2, f_3}(h) = \left[ (1 - r_{f_3, 3}) \tilde{m}_{f_1, f_2}(p(H)) + (1 - r_{f_1, 1})m_{f_3, 3}(h) \right] \\
+ \sum_{A \cap B = h} \gamma_{A, B, f_1, f_2, f_3} I_{A, B, f_1, f_2, f_3} \tilde{m}_{f_1, f_2}(A)m_{f_3, 3}(B)
\]

• Normalise \( \tilde{m}_{f_1, f_2, f_3}(h) \) by using (10) and obtain the optimal combined joint basic probability \( p_{f_1, f_2, f_3}(h) \).

D. Generation of Subtables

In decision making, there are many attributes involved, and there are often interdependent relationships among them. Considering that the number of sub-tables will increase exponentially as the number of attributes increases, we propose an efficient method for grouping attributes to form sub-tables. Using R clustering analysis, input variables are divided into several independent classes, and sub-tables can be established accordingly for MAKER iterative calculation.

R clustering analysis is a kind of cluster analysis that is generally used to classify attributes. Here we utilise spectral clustering [8] to achieve the clustering of attributes, which generally starts from the correlation matrix (or Adjacency matrix) \( S \) of data and changes to a Laplacian matrix \( L \) as follows:

\[
S = \begin{pmatrix}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
\cdots & \cdots & \cdots \\
1 & 1 & 1
\end{pmatrix}
\]

\[
L = D - S
\]

\[
L = \begin{pmatrix}
l_{11} & l_{12} & l_{13} \\
l_{21} & l_{22} & l_{23} \\
l_{31} & l_{32} & l_{33} \\
\cdots & \cdots & \cdots \\
l_{k1} & l_{k2} & l_{k3} \\
\cdots & \cdots & \cdots \\
l_{kk}
\end{pmatrix}
\]

where \( s_{ij} \) represents a measure of the pairwise similarity between data points with indices \( i \) and \( j \). \( D \) is a degree matrix which is diagonal with

\[
d_{ij} = \begin{cases}
\sum_j s_{ij} & \text{if } i = j \\
0 & \text{otherwise}
\end{cases}
\]

Then spectral clustering can be used to achieve the clustering of attributes, by partitioning \( K \) input variables into several independent subsets, \( \mathcal{V}_i, l = 1, 2, \ldots, n \), with \( \mathcal{V}_i \cap \mathcal{V}_j = \emptyset \) and \( \bigcup_{l=1}^{n} \mathcal{V}_l = \mathcal{V} \). The MAKER framework can be further established based on these \( \mathcal{V}_l \) classes.

Fig 1 displays a hierarchical MAKER framework, composed of four sub-MAKER frameworks. The outputs of sub-MAKER frameworks 1 and 2 are taken as the inputs to sub-MAKER framework 3, which is taken as input to sub-MAKER framework 4 together with the independent input state subclass 5. Each of the four sub-MAKER frameworks can be investigated following the processes elaborated in Sec III-C. Take sub-MAKER framework 1 as an example. Two independent subsets of input variables \( V_1 \) and \( V_2 \) are generated, with \( V_1 = \{V_1, V_2, \ldots, V_K\} \) and \( V_2 = \{V_{k+1}, V_{k+2}, \ldots, V_K\} \). Following the procedures in Sec. III-C and optimising the two sub-classes separately, the combined probability mass of evidence \( e_{f_1, \ldots, f_k, h} \) and \( e_{f_{k+1}, (k+1), \ldots, f_K} \) pointing to preposition \( h \) can be obtained: via (13):

\[
\tilde{m}_{f_1, \ldots, f_k, h}(h) = \left[ (1 - r_{f_{k+1}, (k+1), \ldots, f_K}) \tilde{m}_{f_1, \ldots, f_k}(h) \\
+ (1 - r_{f_1, \ldots, f_k}) \tilde{m}_{f_{k+1}, (k+1), \ldots, f_K}(h) \right] \\
+ \sum_{A \cap B = h} \gamma_{A, B, f_1, \ldots, f_K} I_{A, B, f_1, \ldots, f_K} \tilde{m}_{f_1, \ldots, f_k}(A)\tilde{m}_{f_{k+1}, (k+1), \ldots, f_K}(B)
\]

E. Training of MAKER model

Here we discuss how to determine three significant parameters in the MAKER framework, namely the reliability and weight of each piece of evidence and the reliability ratio, that is, \( r_{i, l} \), \( r_{j, m} \), \( w_{i, l} \), \( w_{j, m} \) and \( \gamma_{A, B, i, l, j, m} \) in (11), which are the parameters that need to be assigned for inference.

The parameters are optimized by minimising the mean square error (or \( L_2 \) loss) function to reduce the difference between \( \hat{p}(m)(h) \) and \( p(h) \):

\[
\min \delta = \frac{1}{2M} \sum_{m=1}^{M} \sum_{k \in \Omega} \left( p(m)(h) - \hat{p}(m)(h) \right)^2,
\]

s.t. \( r_{h,i,l}, w_{h,i,l}, \gamma_{A,B,i,l,j,m} \in \Omega \)

where, \( \hat{p}(m)(h) \) and \( p(h) \) are the real probability and the estimated probability to which a proposition \( h \) is true given in the \( m \)th observation, respectively. \( \Omega \) is the feasible space of parameters, with constraints such as \( 0 < w_{h,i,l} \leq 1 \).
In this section, the MAKER framework is illustrated in insurance fraud detection. Majority of insurance companies adopt a set of rules to detect fraud. These rules can be interpretable and very effective in mitigating fraud risk. However, the rules are limited by fixed thresholds or being absolute and fail to capture interactive effects. MAKER, however, preserves the interpretability but also shows the interdependence between all possible combined features explicitly with high relative predictive performance.

The data used in our study includes 3499 records, with output as whether it is ‘high-risk’ or ‘low-risk’. By conducting this research, four relevant criteria are selected, with the detailed sub-categories shown as follows:

- Missing Documents (M)
  - Yes (M1); No (M2)
- Financial difficulty (F)
  - Minor (F1); Moderate (F2); Severe (F3)
- On watch list (W)
  - Yes (W1); No (W2)
- Have frequent complaints (C)
  - Yes (C1); No (C2)

Fig. 2 shows the ratio of missing values in each input variable and certain combinations of variables. The proportion of missing values in each input variable is around 5%.

**A. Training and Validation Results of the MAKER rule**

The validation dataset is drawn from the same database as the training’s. 2800 samples are used as training data for model estimation and the remaining 699 samples are left as the validation dataset for the out-of-sample evaluation, where both training and validation datasets keep the same prior of the original dataset as shown in Table III.

### Table III: Comparison of the prior between the training/validation and original dataset

<table>
<thead>
<tr>
<th></th>
<th>training sample</th>
<th>test sample</th>
<th>all data</th>
</tr>
</thead>
<tbody>
<tr>
<td>High risk</td>
<td>0.2814</td>
<td>0.2818</td>
<td>0.2852</td>
</tr>
<tr>
<td>Low risk</td>
<td>0.7186</td>
<td>0.7182</td>
<td>0.7148</td>
</tr>
</tbody>
</table>

Fig. 3 displays the basic probability of single piece of evidence derived from the training dataset. The most dominant rules of high risk of fraud are extracted and simplified as follows:

- IF (Financial difficulty == Severe) THEN (High Risk)
- IF (Frequent Complaints == Yes) THEN (High Risk)

Table IV shows the combined joint basic probabilities, independent of the prior, with $\delta = 0.038$. Rules of interest could be generated accordingly which enables the claim handlers to analyse these data more intuitively and efficiently.

### Table IV: Optimal combined joint basic probability

<table>
<thead>
<tr>
<th>M</th>
<th>F</th>
<th>W</th>
<th>C</th>
<th>Fraud</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Minor</td>
<td>No</td>
<td>No</td>
<td>0.9983 0.0017</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>0.9983 0.0017</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>0.0296 0.9704</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>0.9154 0.0846</td>
</tr>
<tr>
<td>Yes</td>
<td>Moderate No</td>
<td>No</td>
<td>Yes</td>
<td>0.1288 0.8712</td>
</tr>
<tr>
<td>Yes</td>
<td>Moderate No</td>
<td>Yes</td>
<td>Yes</td>
<td>0.0018 0.9982</td>
</tr>
<tr>
<td>Yes</td>
<td>Moderate No</td>
<td>Yes</td>
<td>No</td>
<td>0.3654 0.6346</td>
</tr>
<tr>
<td>Yes</td>
<td>Moderate No</td>
<td>Yes</td>
<td>Yes</td>
<td>0.0026 0.9974</td>
</tr>
<tr>
<td>Yes</td>
<td>Severe No</td>
<td>No</td>
<td>Yes</td>
<td>0.9983 0.0017</td>
</tr>
<tr>
<td>Yes</td>
<td>Severe No</td>
<td>Yes</td>
<td>Yes</td>
<td>0.0962 0.0358</td>
</tr>
<tr>
<td>Yes</td>
<td>Severe No</td>
<td>Yes</td>
<td>No</td>
<td>0.9984 0.0016</td>
</tr>
<tr>
<td>Yes</td>
<td>Severe No</td>
<td>Yes</td>
<td>Yes</td>
<td>0.3052 0.6948</td>
</tr>
<tr>
<td>Yes</td>
<td>Severe No</td>
<td>Yes</td>
<td>No</td>
<td>0.2659 0.7341</td>
</tr>
</tbody>
</table>

Conventional machine learning approaches, such as Logistic Regression (LR), Neural Network (NN), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB)
and Gradient Boosted Machine (GBM) are compared with the MAKER framework. Note that, imputation methods are required before adopting these ML approaches; hence we pre-process the missing dataset by using single mode imputation and multiple imputation in [9].

Furthermore, in order to demonstrate that the MAKER framework outperforms or is comparable to the aforementioned ML approaches, two measurements are employed to evaluate the prediction accuracy for the numerical example, namely, the root mean square error (RMSE) and the mean absolute error (MAE). The measures of goodness of fit are given by:

$$\text{RMSE}(\text{params}) = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \sum_{h \subseteq H} \left( \hat{p}^{(m)}(h) - \hat{p}^{(m)}(h) \right)^2},$$

$$\text{MAE}(\text{params}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{h \subseteq H} \left| \left( \hat{p}^{(m)}(h) - \hat{p}^{(m)}(h) \right) \right|$$

For both these methods a lower score indicates a better forecast. Also, we show the results of AUC (Area Under The ROC Curve) value, which is invariant to prior probabilities or class prevalence in the data, to demonstrate the robustness of each approach. The validation results for all methods are given in Tab V.

TABLE V: The validation results of different approaches

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>MAE</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.1955</td>
<td>0.0784</td>
<td>0.9400</td>
</tr>
<tr>
<td>ANN</td>
<td>0.1953</td>
<td>0.0784</td>
<td>0.9400</td>
</tr>
<tr>
<td>SVM</td>
<td>0.1964</td>
<td>0.0802</td>
<td>0.9345</td>
</tr>
<tr>
<td>RF</td>
<td>0.1995</td>
<td>0.0415</td>
<td>0.9361</td>
</tr>
<tr>
<td>NB</td>
<td>0.1957</td>
<td>0.0737</td>
<td>0.9403</td>
</tr>
<tr>
<td>GBM</td>
<td>0.8460</td>
<td>0.7174</td>
<td>0.9123</td>
</tr>
<tr>
<td>MAKER</td>
<td>0.1879</td>
<td>0.0781</td>
<td>0.9516</td>
</tr>
</tbody>
</table>

As can be seen from the table, the ML methods with single mode imputation is generally inferior compared to its own performance with multiple imputation. The MAKER framework outperforms other ML approaches in terms of RMSE and AUC value, which implies MAKER can obtain robust and accurate predictions independent of the decision threshold. Note that, MAKER is inferior to RF in terms of MAE, however, RF is viewed as a black box model, as it generates a forest consisting of many trees, which is lack of interpretability when explicit rules need to be generated for reasoning. Moreover, although most of the machine learning models perform well on this data, except GBM, such methodologies depending on prior distributions will be affected in terms of accuracy to a large extent.

V. CONCLUSION AND DISCUSSION

Missing data is a common occurrence and can have drastic impact on the quality of predictive model. Widely used methods for incomplete dataset problems, such as simple deletion strategy, missing value imputation methods, change the statistical features of data, which may lead to biased parameter estimates.

In this paper, we investigated a data-driven inference modelling approach, Maximum Likelihood Evidential Reasoning (MAKER) framework, in which different types of uncertainty can be taken into account for inferential modelling, probabilistic prediction and decision making. Specifically, the reliability of evidence and interdependence between a pair of evidence are explicitly measured.

The MAKER framework is also straightforward for pattern detection and the results are highly intelligible and concise rules that can improve people’s understanding of a problem domain. In addition, MAKER is able to help gain insight from all possible combinations of input variables, which can be used as rules for decision making.

Nevertheless, the MAKER framework in this paper has its limitations. On one hand, the good performance of the MAKER framework would be affected if there is a lack of records in possible combinations. In addition to the spectral clustering of input variables, in the future research, the number of possible combinations within each sub-class should also be considered during the establishment of sub-tables. On the other hand, the MAKER framework can be extended to numerical analysis, by transforming continuous inputs into distributions on the optimised referential values, which can be a set of meaningful and distinctive evaluation standards for describing an attribute by subjective linguistic terms, see [10].

REFERENCES