This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Likelihood Analysis of Imperfect Data

Jian-Bo Yang, Dong-Ling Xu, Xiaobin Xu, and Chao Fu

***Abstract*—This article investigates how to make use of imperfect data gathered from different sources for inference and decision making. Based on Bayesian inference and the principle of likelihood, a likelihood analysis method is proposed for acquisition of evidence from imperfect data to enable likelihood inference within the framework of the evidential reasoning (ER). The nature of this inference process is underpinned by the new necessary and sufficient conditions that when a piece of evidence is acquired from a data source it should be represented as a normalized likelihood distribution to capture the essential evidential meanings of data. While the explanation of sufficiency of the conditions is straightforward based on the principle of likelihood, their necessity needs to be established by following the principle of Bayesian inference. It is also revealed that the inference process enabled by the ER rule under the new conditions constitutes a likelihood inference process, which becomes equivalent to Bayesian inference when there is no ambiguity in data and a prior distribution can be obtained as a piece of independent evidence. Two examples in decision analysis under uncertainty and a case study about fault diagnosis for railway track maintenance management are examined to demonstrate the steps of implementation and potential applications of the likelihood inference process.**

***Index Terms*—Bayesian inference, decision making under uncertainty, evidential reasoning (ER), likelihood analysis of data, likelihood principle.**

I. Introduction

DATA and judgements play an essential role in inference, modeling, and decision making, for example, in evidence-based multiple attribute decision analysis under uncertainty [22], [37], [40], [41], [42], [43] and probabilistic rule-based system modeling and learning [4], [23], [28], [38], [39], [44], [45]. Data analysis through probabilistic inference is extensively investigated [3], [18]. There is also significant research in using ambiguous judgements for decision making under uncertainty [1], [42], [43]. However, it remains as a challenge to make use of imperfect data for probabilistic inference and evidence-based decision making [46], [47]. This is because imperfect data is often generated from routine processes, such as collecting operational data and recording daily activities, and therefore can be highly imbalanced and associated with various types of uncertainty, such as randomness and ambiguity. Data imperfection can include not only the imbalance and ambiguity but also other features, such as incompleteness and inaccuracy [46]. This article will focus only on addressing how to analyze random data that is also imbalanced and ambiguous. Table I describes the variables that will be used to establish a likelihood method to analyze such imperfect data. Note that these variables are denoted in consistence with the notations used in our previous research, such as [46] which is in turn based on Dempster's original work [10], [11] and his system view [13] on probabilistic inference with randomness and ambiguity.

Imbalanced data means that there is a disproportionate ratio of observations in some system states or classes. It is one of the potential problems in data mining, inference, machine learning (ML), and decision making. Data level solutions for handling imbalanced data include different forms of resampling, in particular undersampling and oversampling [25]. For instance, random undersampling is a nonheuristic method that aims to balance class distribution through random elimination of majority class examples; however, it can discard potentially useful data which could be important for inference and decision making [24]. Random oversampling is another nonheuristic method for balancing class distribution through random replication of minority class examples, but can increase the chance of causing overfitting [7]. In probabilistic inference and evidence-based decision making, the probability distribution of target population is estimated from sample distribution. If sample distribution is drawn randomly, it can be used to estimate population distribution. The problem that is common to resampling methods is that once resampling is performed, the sample can no longer be regarded as random [25] and it is thus not appropriate to use the sample to estimate population distribution. Since the purpose of data analysis in this article is to estimate population distribution for probabilistic inference and decision making, nonresampling methods will be investigated.

Among different types of uncertainty, randomness is used to signify well-defined statistical properties, and ambiguity can be used to represent properties incurred due to missing data.

Manuscript received 20 December 2022; accepted 9 March 2023. This work was supported in part by the Alliance Strategic Research Fund of The University of Manchester under Grant AA15032; in part by the European Union Horizon 2020 Research and Innovation Programme RISE (REMESH) under Grant 823759; in part by the U.S. Air Force Office of Scientific Research under Grant FA2386-15-1-5004; in part by the Natural Science Foundation of China under Grant U1709215, Grant 61433001, and Grant 71622003; and in part by the InnoHK initiative, the Government of the HKSAR, and Hong Kong Laboratory for AI-Powered Financial Technologies. This article was recommended by Associate Editor Y. Dong. *(Corresponding author: Jian-Bo Yang.)*

Jian-Bo Yang and Dong-Ling Xu are with the Alliance Manchester Business School, The University of Manchester, M15 6PB Manchester, U.K. (e-mail: jian-bo.yang@manchester.ac.uk).

Xiaobin Xu is with the Department of Automation, Hangzhou Dianzi University, Hangzhou 310005, China.

Chao Fu is with the School of Management, Hefei University of Technology, Hefei 230009, China.

This article has supplementary downloadable material available at https://doi.org/10.1109/TSMC.2023.3262585, provided by the authors.

Digital Object Identifier 10.1109/TSMC.2023.3262585

2168-2216 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Manchester. Downloaded on April 11,2023 at 09:43:10 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
DESCRIPTION OF VARIABLES

Variable	Description
θ	a set of system states
Θ	system space or frame of discernment
2^{Θ}	the power set of Θ
$\varnothing$	empty set
$c_{i,j}$	the likelihood associated with sample data I_0, to which test result t_j is observed given that state h_i is true
e_0	sample prior distribution
e_j	the j^{th} piece of evidence
$e_{j,k}$	the evidence of observing $x_k = j$
$f_k(t)$	train acceleration data from sensor k at time t
h_i	the i^{th} singular system state
I_0	sample data
$Ir(t)$	the irregularity or absolute vertical displacement of rail track at time t
L	the total number of test results
m_j	basic probability mass distribution
$m_{\theta,j}$	the joint probability that evidence e_j points to state θ and state θ is true, referred to as the basic probability mass
$m_{2^{\Theta},j}$	The probability mass that is earmarked to power set 2^{Θ} for evidence e_j
$\widetilde{m}_{\theta,e_0\wedge e_1}$	the probability mass that both evidence e_0 and evidence e_1 jointly support state θ
N	the total number of singular system states
$p_{i,0}$	the prior probability of state h_i
$p_{i,j}$	the basic probability to which test result t_j points to state h_i
$p_{\theta,j}$	the basic probability that evidence e_j points precisely to a set of states θ
$p_{i,jk}$	the basic probability that evidence $e_{j,k}$ points to state h_i,
$p_{\theta,e_0\wedge e_1}$	the combined probability that e_0 and e_1 jointly support state θ
w_j	the weight of evidence e_j,
$s_{\theta,j}$	the number of observations where both state θ is true and test result t_j is observed
S_{θ}	the number of all observations where state θ is true
S	the total number of all observations in sample
t_j	the j^{th} test result
$t_{i,j}$	the i^{th} interval of the readings of sensor j
T_j	the number of all observations where test result t_j is observed
x_k	the result from the k^{th} random test

Missing data is defined as the data value that is not stored for a variable in observation of interest, and can be classified into three types [21]: 1) missing completely at random (MCAR); 2) missing at random (MAR); and 3) missing not at random. A number of methods for handling missing data have been developed. The most common approach for dealing with missing data is to simply delete those cases with missing data and analyze the remaining data, known as complete case analysis or listwise deletion. However, when data do not meet the assumption of MCAR, listwise deletion may cause bias in the estimates of parameters [14]. Opposite to data deletion is to impute missing data, including single imputation and multiple imputation. The former includes a number of approaches, such as the mean substitution approach to replace missing data by the mean value of a variable [27], the last observation approach to replace missing data by the last observed value before the missing value [17], and maximum likelihood approaches, e.g., expectation–maximization to create a new data set, in which all missing values are imputed with values estimated by maximum likelihood methods [12]. The latter is a strategy for handling missing data by replacing missing values with a set of plausible values which contain the natural variability and uncertainty of the true values [33]. Multiple imputation is shown to produce valid statistical inference that reflects uncertainty associated with the estimation of missing data, although the statistical principles of multiple imputation may be difficult to understand [33]. However, deleting missing data or imputing missing data can change the features of data, causing concerns on interpretability and trustworthiness when data is used for probabilistic inference and evidence-based decision making.

From the above discussions, two questions emerge that need to be addressed. First, data collected from different sources can be imbalanced, leading to difficulty in using the data for probabilistic inference and evidence-based decision making. The *first question* is therefore how to analyze imbalanced data without resampling or changing its random nature. Second, data generated from routine activities can be both imbalanced and ambiguous, e.g., incurred due to missing data. The *second question* is how to deal with both imbalance and ambiguity common to imperfect data in the processes of analyzing data in a principled manner. This article aims to investigate these questions.

In an imbalanced dataset, there is a disproportionate ratio of observations in some states. In other words, the prior probabilities of states acquired from the imbalanced dataset are not

Authorized licensed use limited to: University of Manchester. Downloaded on April 11,2023 at 09:43:10 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

equal. Using prior probabilities and new observations, represented as likelihoods, to generate posterior probabilities constitutes the fundamental principle of Bayesian inference [18], [19]. Extensive research on Bayesian inference has been conducted, such as the estimation of prior probability [16], [20], prior-free statistical inference [10], [11], [15], [31], [32], and inference with default priors [16]. The investigation of the first question will be based on exploiting the strengths of Bayesian inference, and a likelihood method will be explored to analyze imbalanced data.

To address the second question, randomness and ambiguity in data will be investigated on the basis of Dempster's original work on probabilistic inference with imperfect data [10], [11], summarized as Dempster's rule [29], and his original thinking on system view [13], where both types of uncertainty can be represented by allowing the assignment of basic probabilities to not only singular system states but also their subsets as a whole. Considerable research has been conducted on the basis of Dempster's original work. For example, Shafer [29], [30] and Smets [35] constructed belief functions to show that the application of Dempster's rule to their belief functions can approximate Bayesian inference in general when sample size is very large but only lead to the same result as that of Bayes' rule for a rather special case with a single frequency distribution [30]. Aickin [2] proposed to construct credibility functions and modify Dempster's rule for likelihood inference, which leads to the computations that are quite different from those of Smets. In Aickin's approach, a credibility function is generated by dividing all likelihoods by the maximum likelihood, which satisfies the principle of likelihood [5].

Under Dempster's system view for probabilistic inference, the authors proposed to acquire evidence as likelihood distribution generated from sample data [46], [47]. This concept has been successfully applied to several areas, such as probabilistic rule-based data classification [38], fault diagnosis of marine systems [39], an evidential reasoning (ER)-based prediction of ICU admission and in-hospital death of trauma patients [23], and financial services [28]. However, it remains as an open question whether it is necessary to acquire evidence as normalized likelihood from imperfect data or this is just an alternative approach suggested as a rule of thumb.

One of the main purposes of addressing the above two questions is to facilitate probabilistic inference and evidence-based decision making in the framework of the ER rule [46], which is established to enhance and augment Dempster's rule. The ER rule requires that evidence be acquired as basic probability distribution, so that inference driven by the ER-rule is deemed to be probabilistic. However, what principles should be followed to acquire evidence from imperfect data remains open for investigation. This article will first investigate how to acquire evidence on the common scale of relative frequency from unambiguous data by following the two principles: 1) the principle of likelihood [5] and 2) the principle of Bayesian inference [19], referred to as Bayesian principle in the rest of this article. It will then explore how to acquire evidence from ambiguous data by following the same principles.

In the above context, this article will focus on establishing new necessary and sufficient conditions required as the basis for constructing a likelihood inference process. Conceptually, the conditions require that evidence can be acquired from sample data as normalized likelihood distribution, independent of the prior distribution of the sample data, referred to as sample prior. Such evidence is referred to as prior-independent evidence in this article. We will show that the combination of multiple pieces of prior-independent evidence using the ER rule forms a symmetrical likelihood inference process, leading to the posterior probability that treats prior distribution as total ignorance or unknown by default. If a prior distribution becomes known, it will be taken into account as a piece of independent evidence and combined with other evidence in the ER process, which ought to result in the same posterior probability as Bayesian posterior when there is no ambiguity in data.

The focus of Section II will be on the establishment of the new necessary and sufficient condition to acquire evidence from imbalanced data without ambiguity for conventional probabilistic inference, where evidence is profiled as classical probability distribution with probability assignable to singular states only. In Section III, the condition will be extended to acquire evidence from imbalanced and ambiguous data to support augmented probabilistic inference and evidence-based decision making, where evidence can be profiled as basic probability distribution with probability assignable to any singular states or any subset of singular states. These conditions are intended to provide a theoretical foundation and a principled and practical method to acquire evidence from imperfect data for likelihood inference, probabilistic modeling, and evidence-based decision making, such as multiple criteria decision analysis under uncertainty [22], [43] and belief rule-based modeling and learning [1], [4], [39].

II. Likelihood Analysis of Imbalanced Data

The following example is used to explain what issues are to be addressed in this article and what challenges there exist in addressing these issues, in particular how to acquire evidence from imperfect data for likelihood inference and evidence-based decision making.

Example 1: A manufacturer claims that its test method, named as Method I, is highly effective for detecting steroid use. An experiment is conducted to investigate this claim, in which the method is used to test 2220 individuals, as shown in Table II. Most of the tests lead to either positive (t_1) or negative (t_2) results, and the majority of the individuals are classified as either steroid user (h_1) or steroid free (h_2); however, there are 220 ambiguous cases in which a test result is inconclusive or it is unknown whether an individual is a steroid user or steroid free, or both. If the method is used to test an athlete and the test result is positive, what is the probability that the athlete is a steroid user, so that appropriate actions can be taken with confidence to deter steroid abuse?

The data presented in Table II is imperfect in the sense that there are 220 ambiguous cases and the data is imbalanced as there are far more steroid free individuals than steroid users in this sample. In general, the imperfection of data may also include incompleteness and inaccuracy [46], although these

Authorized licensed use limited to: University of Manchester. Downloaded on April 11,2023 at 09:43:10 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
SAMPLE DATA GENERATED BY USING METHOD I (I_0)

Frequency		Test results: Positive (t_1)	Negative (t_2)	Inconclusive
System states	*Steroid user* (h_1)	190	10	3
	Steroid free (h_2)	270	1530	30
	Unknown ($\Theta = \{h_1, h_2\}$)	60	120	7

TABLE III
CLEANED DATA (I_0)

Frequency	t_1	t_2	Total
h_1	190	10	200
h_2	270	1530	1800

TABLE IV
BAYESIAN DATA ANALYSIS

Likelihood	t_1	t_2	Prior
h_1	0.95	0.05	0.1
h_2	0.15	0.85	0.9

features are not the focus of investigation in this article. While the likelihood analysis method proposed in this article does not require undersampling or oversampling to generate balanced data for inference and decision making, it is important to note that sampling is a technique widely used in ML, such as k-fold cross validation. If the likelihood analysis method is applied in such circumstances, the resampling of data into k folds needs to follow the rule that the likelihood function of the split data in each fold is as consistent as possible with that of the original data for every state.

The decision outcome of the above example has considerable impact on the sport life of the athlete personally and also wider implication in deterrence of steroid abuse. As such, it is necessary to use a principled approach to analyze the data for robust inference and decision making. Bayesian inference is regarded as a principled approach for statistical data analysis and probabilistic inference [18]. In this section, Bayes' rule will be first introduced in Section II-A as the core of Bayesian inference and illustrated using the above example. The ER rule will then be introduced in Section II-B to analyze ambiguity in data. To establish the ER rule as a principled approach for analysis of ambiguous data and probabilistic inference, Section II-C will focus on identifying a new sufficient and necessary condition to ensure that the ER rule generates the same posterior probability as Bayes' rule does when prior probability becomes known and there is no ambiguity in data.

A. Bayes' Rule for Inference With Unambiguous Data

Bayesian view of probability as degree of belief is based on the algebra of probable inference [8], [9], [36], the Kolmogorov axioms of probability, the Jeffreys's theory of probability [6], [20], among others. Under this view, probability can be constructed from partial information and updated when new information becomes available. In this article, Bayes' rule is presented for combining prior probability with new information to generate posterior probability.

Let e_0 stand for sample prior distribution, acquired from sample data I_0 as follows:

$$e_0 = \left\{ \left(h_i,\ p_{i,0}\right), i = 1, \ldots, N, \sum_{i=1}^{N} p_{i,0} = 1 \right\} \quad (1)$$

where h_i is the ith system state, $p_{i,0}$ is the prior probability of h_i calculated from sample data I_0, or $p_{i,0} = p(h_i|I_0)$, and N is the number of states that are mutually exclusive and collectively exhaustive.

Let $c_{i,j}$ stand for the likelihood associated with sample data I_0, to which test result t_j is observed given that state h_i is true, that is $c_{i,j} = p(t_j|h_i, I_0)$. If test result t_j is observed as new information, Bayes' rule can be described as follows, to generate the posterior probability that state h_i is true given both I_0 and the observation of t_j

$$p\left(h_i|t_j, I_0\right) = \frac{p\left(t_j|h_i, I_0\right)p(h_i|I_0)}{\sum_{n=1}^{N} p\left(t_j|h_n, I_0\right)p(h_n|I_0)} = \frac{c_{i,j}p_{i,0}}{\sum_{n=1}^{N} c_{n,j}p_{n,0}}. \quad (2)$$

Bayesian inference based on the above Bayes' rule asserts that the combination of prior probability with new information ought to result in posterior probability. This assertion should be followed for probabilistic inference with data. In this article, this assertion is referred to as Bayesian principle and will be used as a condition for establishing a new evidence acquisition method from sample data.

A question is whether the above Bayesian inference can be directly applied to analyzing Example 1 to find the required probability. The short answer is no as Bayesian analysis does not directly handle ambiguous cases. To conduct Bayesian data analysis and inference, the ambiguous cases of the sample data need to be processed or "cleaned." The most common approach is perhaps to delete the ambiguous cases. If all of the 220 ambiguous cases are removed by deleting the last row and the last column of Table II, we will have a set of "cleaned" data of 2000 individuals, with 200 steroid users and 1800 individuals being steroid free, as shown in columns 2 and 3 of Table III.

Based on the data of Table III, Bayesian data analysis is conducted to find the prior probabilities that an individual implied by the sample data is a steroid user or steroid free, denoted by $p_{1,0} = p(h_1|I_0)$ and $p_{2,0} = p(h_2|I_0)$, respectively. The likelihoods that an individual has a positive test result given that this person is a steroid user or steroid free, denoted by $c_{1,1} = p(t_1|h_1, I_0)$ and $c_{2,1} = p(t_1|h_2, I_0)$, respectively, can be calculated as follows:

$$p_{1,0} = p(h_1|I_0) = \frac{200}{200 + 1800} = 0.1$$
$$p_{2,0} = p(h_2|I_0) = \frac{1800}{200 + 1800} = 0.9 \quad (3)$$

Authorized licensed use limited to: University of Manchester. Downloaded on April 11,2023 at 09:43:10 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

$$c_{1,1} = p(t_1|h_1, I_0) = \frac{190}{200} = 0.95$$
$$c_{2,1} = p(t_1|h_2, I_0) = \frac{270}{1800} = 0.15 \tag{4}$$

as summarized in the second and last columns of Table IV. The sample prior probability distribution generated from (3) is represented by $e_0 = \{(h_1, 0.1), (h_2, 0.9)\}$.

Bayesian inference can be conducted to generate posterior probability as follows:

$$\begin{aligned} p(h_1|t_1, I_0) &= \frac{p(t_1|h_1, I_0)p(h_1|I_0)}{p(t_1|h_1, I_0)p(h_1|I_0) + p(t_1|h_2, I_0)p(h_2|I_0)} \\ &= \frac{0.95 \times 0.1}{0.95 \times 0.1 + 0.15 \times 0.9} = 0.413. \end{aligned} \tag{5}$$

The above result suggests that there is a probability of 0.413 that the athlete is a steroid user.

The above-generated posterior probability is dependent on the sample prior, which may or may not express one's prior belief that the athlete is a steroid user or steroid free before any evidence is acquired. If there is no prior knowledge about the athlete's state of steroid use, one may assume that the prior is uniformly distributed, or it is equally likely that the athlete is steroid user or steroid free. Under this assumption of uniform prior, Bayesian inference leads to the following result:

$$p(h_1|t_1, I_0) = \frac{0.95 \times 0.5}{0.95 \times 0.5 + 0.15 \times 0.5} = 0.8636. \tag{6}$$

B. Basic Probability Distribution and the ER Rule

Dempster's rule and system view [10], [11], [13] provides a unified framework to model and infer with both randomness and ambiguity. In the framework, probability distribution is extended to basic probability distribution to model evidence, in which ambiguity is explicitly measured.

Suppose h_i is the ith system state and $\Theta = \{h_1, \ldots, h_N\}$ is the whole set of mutually exclusive and collectively exhaustive system states, referred to as system space. The power set of Θ consisting of a total number of 2^N subsets of Θ, denoted by 2^Θ, is given as follows:

$$\begin{aligned} 2^\Theta = \{&\varnothing, \{h_1\}, \ldots, \{h_N\}, \{h_1, h_2\}, \ldots, \{h_1, h_N\}, \\ &\ldots, \{h_1, \ldots, h_{N-1}\}, \Theta\}. \end{aligned} \tag{7}$$

The jth piece of evidence e_j is modeled as a basic probability distribution (BPD) as follows:

$$e_j = \left\{ (\theta,\ p_{\theta,j}) \quad \forall \theta \subseteq \Theta,\ \sum_{\theta \subseteq \Theta} p_{\theta,j} = 1 \right\} \tag{8}$$

where $(\theta,\ p_{\theta,j})$ is a basic element of evidence e_j, and $p_{\theta,j}$ is the probability that evidence e_j points precisely to a set of states θ as a whole. $p_{\theta,j}$ cannot be further assigned to any subset of θ and is referred to as basic probability for θ. Basic element $(\theta,\ p_{\theta,j})$ is called a focal element of e_j if basic probability $p_{\theta,j} > 0$. In the rest of this article, h_i is reserved to mean the ith singular state or state i, whilst θ is an element of the power set, which can mean either a singular state or a set of states in general and is simply called state θ for short. In (8), ambiguity is explicitly measured by basic probabilities assigned to subsets of states other than singular states. For instance, basic probability assigned to the system space Θ measures the degree of unknown or global ignorance.

The ER rule [46] is established on the basis of the framework of (7) and (8), and is briefly introduced in this section in the context of probabilistic inference. The establishment of the ER rule originates from the observation that evidence generated from sample data is in general not fully reliable due to data quality and is always associated with a degree of reliability. The weight of evidence e_j, denoted by w_j, is defined as the conditional probability that a system state is true given that evidence e_j points to the state. In applications, such as information fusion, w_j can be regarded as a measure for the ability of information source, from which e_j is acquired, to provide correct identification of system states [34].

In the above context, parameter $m_{\theta,j} = w_j p_{\theta,j}$ is defined as the joint probability that evidence e_j points to state θ, and state θ is true. In other words, $m_{\theta,j}$ is referred to as the basic probability mass that evidence e_j supports state θ. In this context of evidence supporting states, e_j can be represented as the following basic probability mass distribution by m_j

$$m_j = \left\{ (\theta,\ m_{\theta,j}) \quad \forall \theta \subseteq \Theta;\ \left(2^\Theta,\ m_{2^\Theta,j}\right) \right\} \tag{9}$$

where $(\theta,\ m_{\theta,j})$ is the basic element of evidence e_j supporting state θ, and $(2^\Theta,\ m_{2^\Theta,j})$ is the residual support of evidence e_j incurred due to its reliability, with $m_{2^\Theta,j} = 1 - w_j$. The residual support needs to be earmarked to power set 2^Θ, rather than assigned to any singular state or any subset of states by evidence e_j alone. This is because it is not reliable to do so without taking into account other evidence conjunctively, while the conjunction of the power set with any state is the same as the latter, so that the residual support can be assigned to the focal elements of the other evidence conjunctively. Note that $\sum_{\theta \subseteq \Theta} m_{\theta,j} + m_{2^\Theta,j} = w_j \sum_{\theta \subseteq \Theta} p_{\theta,j} + (1 - w_j) = 1$ always holds since there is $\sum_{\theta \subseteq \Theta} p_{\theta,j} = 1$ in (8), so (9) is a probability distribution augmented from (8).

Suppose two pieces of evidence e_0 and e_1 with weights w_0 and w_1 are independent of each other in the sense that the information carried by e_1 does not depend on whether e_0 is known or not, and vice versa. They are profiled by m_0 and m_1 as in (9), with $j = 0$ and $j = 1$, respectively, where $m_{\theta,0} = w_0 p_{\theta,0}$ and $m_{\theta,1} = w_1 p_{\theta,1}$ are the basic probability masses that evidence e_0, and evidence e_1 support state θ, respectively, for any $\theta \subseteq \Theta$. The combined probability that state θ is true given that e_0 and e_1 are observed, denoted by $p_{\theta,e_0 \wedge e_1}$, is generated from m_0 and m_1 as follows [46]:

$$\begin{aligned} p_{\theta,e_0\wedge e_1} &= \begin{cases} 0, & \theta = \varnothing \\ \frac{\widetilde{m}_{\theta,e_0\wedge e_1}}{\sum_{D\subseteq\Theta} \widetilde{m}_{D,e_0\wedge e_1}}, & \theta \neq \varnothing \end{cases} \\ \widetilde{m}_{\theta,e_0 \bigwedge e_1} &= \Bigg[(1 - w_1)m_{\theta,0} + (1 - w_0)m_{\theta,1} \\ &\qquad + \sum_{B \bigcap C = \theta} m_{B,0} m_{C,1} \Bigg]. \end{aligned} \tag{10}$$

The recursive formula of the ER rule to combine multiple pieces of evidence in any order is also given [46], where it is

Authorized licensed use limited to: University of Manchester. Downloaded on April 11,2023 at 09:43:10 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

shown that Dempster's rule is a special case of the above ER rule when each piece of evidence e_j is assumed to be fully reliable, or $w_j = 1$ for any j.

C. Likelihood Method for Analyzing Data to Acquire Evidence

The main purpose of introducing the ER rule is to handle imbalanced and ambiguous data shown in Table II. Before the issue of ambiguity is addressed, it is necessary to investigate if the ER rule can be applied to inference with unambiguous data as shown in Table III in the same principled way as Bayesian inference. In this context, it is fundamental to investigate how evidence should be acquired from sample data in a principled manner so that it can be used to enable the ER rule for probabilistic inference. This is the focus of this article.

This section will be focused on how to acquire evidence from unambiguous data shown in Table III. If basic probabilities are assigned to singular states only, a BPD defined in (8) reduces to an ordinary probability distribution, as shown by (1). Such evidence is referred to as probabilistic evidence in this article. In this section, we establish a new necessary and sufficient condition for evidence acquisition, which ought to be followed so that the ER rule given in (10) can generate the same results as Bayes' rule does when used to combine two pieces of fully reliable probabilistic evidence.

Since prior probability distribution e_0 given by (1) is already a BPD, it can be directly used in the ER rule as the first piece of evidence, and what needs to be investigated is how to turn an observation into a new piece of evidence and represent it as a BPD as well, so that the combination of the new evidence with e_0 using the ER rule leads to posterior probability.

Let $p_{i,j}$ stand for the basic probability to which test result t_j points to state h_i, with $0 \leq p_{i,j} \leq 1$ and $\sum_{n=1}^{N} p_{n,j} = 1$ for any $i = 1, \ldots, N$ and $j = 1, \ldots, L$, so that basic probability is assigned to singular states only. Here, N and L are the total number of states in the system space and the total number of test results, respectively. If evidence e_j is acquired from test result t_j, it can be profiled as a BPD over the set of singular states as follows:

$$e_j = \left\{ \left(h_i,\ p_{i,j}\right), i = 1, \ldots, N, \sum_{i=1}^{N} p_{i,j} = 1 \right\} j = 1, \ldots, L. \quad (11)$$

It is straightforward to deduce that when applied to combine two pieces of independent probabilistic evidence e_0 and e_j, the ER rule given by (10) reduces to

$$p_{h_i, e_0 \wedge e_j} = \frac{\left[(1 - w_j) m_{i,0} + (1 - w_0) m_{i,j}\right] + m_{i,0} m_{i,j}}{\sum_{n=1}^{N} \left(\left[(1 - w_j) m_{n,0} + (1 - w_0) m_{n,j}\right] + m_{n,0} m_{n,j}\right)}. \quad (12)$$

If it is further assumed that e_0 and e_j each have the highest weight in that $w_0 = w_j = 1$, the ER rule will reduce to the following format:

$$p_{h_i, e_0 \wedge e_j} = \frac{p_{i,0} p_{i,j}}{\sum_{n=1}^{N} \left(p_{n,0} p_{n,j}\right)}. \quad (13)$$

By comparison of (13) with (2), one can observe that making basic probability $p_{i,j}$ proportional to likelihood $c_{i,j}$ leads to $p_{h_i, e_0 \wedge e_j}$ being equivalent to $p(h_i | t_j, I_0)$. This observation does not occur accidentally but follows the principle of likelihood [5], which states that two likelihood functions are regarded as the same if they are proportional to each other.

The principle of likelihood forms one of the two conditions for establishing the following method for acquiring $p_{i,j}$ from data. As mentioned before, Bayesian principle forms the other condition to construct the method. The satisfaction of these two principles leads to the following likelihood analysis method for acquisition of probabilistic evidence from imbalanced data without ambiguity.

Theorem 1: Suppose prior evidence e_0 is profiled as a BPD defined by (1), and new evidence e_j is acquired from the observed test result t_j and profiled by (8), with both e_0 and e_j having the highest weight, or $w_0 = w_j = 1$. The combination of e_0 and e_j by applying the ER rule leads to the same posterior probability as that generated by applying Bayes' rule, or $p_{h_i, e_0 \wedge e_j} = p(h_i | t_j, I_0)$, if and only if basic probability $p_{i,j}$ for evidence e_j is given as normalized likelihood as follows:

$$p_{i,j} = c_{i,j} / \sum_{n=1}^{N} c_{n,j} \quad \forall i = 1, \ldots, N \text{ and } j = 1, \ldots, L. \quad (14)$$

Proof: See Section S1.1 proof of Theorem 1 of the Supplementary Material part for this article. ■

Equation (14) asserts that the basic probability $p_{i,j}$ that test result t_j points to state h_i should be proportional to the likelihood $c_{i,j}$ that test result t_j is observed given state h_i. More precisely, $p_{i,j}$ should be the normalized $c_{i,j}$, obtained by dividing $c_{i,j}$ for any $i = 1, \ldots, N$ with the same constant $\sum_{n=1}^{N} c_{n,j}$. That is, basic probability vector $[p_{1,j}, \ldots, p_{N,j}]^T$ should be proportional to likelihood vector $[c_{1,j}, \ldots, c_{N,j}]^T$. According to the principle of likelihood [5], the evidential meaning contained in $[p_{1,j}, \ldots, p_{N,j}]^T$ is the same as that in $[c_{1,j}, \ldots, c_{N,j}]^T$. By substituting (14) to (13), it is straightforward to see that $p_{h_i, e_0 \wedge e_j}$ equals $p(h_i | t_j, I_0)$. It is therefore straightforward to assert that (14) is sufficient to ensure that the ER rule generates the same result as Bayes' rule yet in a symmetrical manner in the sense that both e_0 and e_j are probability distributions over system states.

To show that (14) is also a necessary condition, Bayesian principle ought to be followed. That is, $p_{i,j}$ must be given by (14) as normalized likelihood in order to make $p_{h_i, e_0 \wedge e_j}$ equal to the posterior probability that state h_j is true given prior evidence e_0 and the observation of test result t_j. To put it simply, if $p_{i,j}$ is not proportional to $c_{i,j}$, the combined results of e_0 and e_j will not be equal to the posterior probability generated by Bayes' rule, or $p_{h_i, e_0 \wedge e_j}$ generated by (13) will not be a principled or meaningful result.

Theorem 1 in essence establishes a mapping from sample data to evidence by identifying the likelihood of observing a test result given a system state. That is, the observation of a test result is transformed into a piece of evidence profiled as a BPD about which states the test result points to and to what extent. As explained later, this mapping is independent of the prior distribution of the sample data, or irrespective of whether the sample data is balanced or imbalanced. Theorem 1 is essential to ensure that the ER rule produces equivalent

Authorized licensed use limited to: University of Manchester. Downloaded on April 11,2023 at 09:43:10 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

outcomes to those of Bayesian inference when all pieces of evidence are probabilistic and prior distribution is also known. It also shows that the ER process constitutes a symmetrical likelihood inference process as it allows multiple pieces of independent evidence to be represented in the same format of BPD and combined in any order without having to rely on a known prior distribution, which is formally stated as follows.

Corollary 1: Suppose evidence e_j is acquired by assigning its basic probability $p_{\theta,j} = p_{i,j}$ using (14) for any $\theta = h_i \in \Theta$ and $p_{\theta,j} = 0$ for any other $\theta \subseteq \Theta$, and e_j is fully reliable. If e_j is combined with a completely unknown prior (or e_0), or total ignorance, which has the basic probability of one assigned to system space Θ, and zero to all other system states, the resultant posterior probabilities will be the same as the basic probabilities of e_j.

Proof: See Section S1.2 proof of Corollary 1 of the Supplementary Material part for this article. ■

From Theorem 1 and Corollary 1, three assertions can be made. The first one is that the ER rule constitutes a likelihood inference process in the sense of combing multiple pieces of probabilistic evidence acquired as likelihood distributions. This is because each piece of probabilistic evidence is so acquired that its basic probabilities are assigned as normalized likelihood or given by (14). The second assertion is that such a likelihood inference process does not depend on prior distribution, or whether data is balanced or imbalanced. The third assertion is that the ER process is a symmetrical probabilistic inference process and generates the same posterior probability as Bayesian inference does given the same prior probability distribution. Putting the three assertions together, it can be concluded that if probabilistic evidence is acquired as a normalized likelihood distribution and satisfy the assumptions in Corollary 1, the ER process constitutes a likelihood inference process and becomes equivalent to Bayesian inference when prior probability distribution is given and taken into account as independent evidence.

D. Illustration of ER for Symmetrical Bayesian Inference

Theorem 1 enables us to use the data of Table III as an example to illustrate how an ER can be performed for equivalent Bayesian inference in a symmetrical manner. We will also show why evidence acquired using Theorem 1 is independent of prior distribution.

By applying the ER rule, posterior probability $p_{h_1,e_0 \wedge e_1}$ can be calculated by combining the evidence of a positive test result e_1 and the sample prior e_0, where e_0 and e_1 are represented by $e_0 = \{(h_1, p_{1,0}), (h_2, p_{2,0})\}$ and $e_1 = \{(h_1, p_{1,1}), (h_2, p_{2,1})\}$. From the data of Table III, we have $e_0 = \{(h_1, 0.1), (h_2, 0.9)\}$ calculated by (3). From (14) and (4), the basic probabilities $p_{1,1}$ and $p_{2,1}$ for a positive test result are given as the following normalized likelihoods:

$$p_{1,1} = \frac{c_{1,1}}{c_{1,1} + c_{2,1}} = \frac{0.95}{0.95 + 0.15} = \frac{0.95}{1.1} = 0.8636$$
$$p_{2,1} = \frac{c_{2,1}}{c_{1,1} + c_{2,1}} = \frac{0.15}{1.1} = 0.1364. \qquad (15)$$

Note from (6) that the above basic probabilities are the same as the posterior probabilities generated by Bayesian inference based on the uniform prior distribution. This implies that the basic probabilities generated by (14) are equivalent to Bayesian posterior probabilities generated from the same likelihood function and a uniform prior distribution.

The above implication does not incur accidentally. In fact, from Tables III and IV the calculation of $p_{1,1}$ can be reformulated as the process of calculating the relative frequency of 190 cases of joint "positive test result and steroid user" out of 200 steroid users, over the sum of the same 190 cases and the number of the cases of joint "positive test result and steroid free individuals" out of an equal number of 200 steroid free individuals, which is equal to (270/1800) × 200, that is

$$p_{1,1} = \frac{190}{190 + (270/1800) \times 200} = \frac{(190/200)}{(190/200) + (270/1800)} = \frac{c_{1,1}}{c_{1,1} + c_{2,1}}. \qquad (16)$$

The above equation shows that the calculation of $p_{1,1}$ does not depend on the prior distribution of steroid user and steroid free, nor on any resampling. Similarly, the calculation of $p_{2,1}$ does not depend on the prior distribution or resampling either and can be reformulated as follows:

$$p_{2,1} = \frac{270}{(190/200) \times 1800 + 270} = \frac{(270/1800)}{(190/200) + (270/1800)} = \frac{c_{2,1}}{c_{1,1} + c_{2,1}}. \qquad (17)$$

The above calculations and interpretation show that vector $[p_{1,1}, p_{2,1}]^T$ is proportional to vector $[c_{1,1},\ c_{2,1}]^T$. That is, $p_{1,1}$ and $p_{2,1}$ are the basic probabilities that possess the same evidential meaning as $c_{1,1}$ and $c_{2,1}$. From the above discussions, basic probability calculated by (14) has the following interpretation.

Corollary 2: Basic probability $p_{i,j}$ calculated by (14) for test result t_j is the same as the posterior probability generated from the observation of t_j and an assumed uniform prior distribution e_0 as defined by (1) with $p_{i,0} = p_{n,0}$ for any $i, n = 1, \ldots, N$.

Proof: Given the conditions $p_{i,0} = p_{n,0}\ \ \forall i, n = 1, \ldots, N$, the proof is straightforward as follows:

$$p_{i,j} = \frac{c_{i,j}}{\sum_{n=1}^{N} c_{n,j}} = \frac{c_{i,j} p_{i,0}}{\sum_{n=1}^{N} c_{n,j} p_{n,0}} = p\big(h_i | t_j, I_0\big). \qquad (18)$$

If the prior distribution of the data in Table III is calculated by (3), using (13) to combine it with the evidence of observing a positive test result leads to the following outcome:

$$p_{h_1,e_0 \wedge e_1} = \frac{p_{1,1} p_{1,0}}{p_{1,1} p_{1,0} + p_{2,1} p_{2,0}} = \frac{0.8636 \times 0.1}{0.8636 \times 0.1 + 0.1364 \times 0.9} = 0.413. \qquad (19)$$

The outcome is the same as the posterior probability of (5) generated by using Bayes' rule, i.e., $p_{h_1,e_0 \wedge e_1} = p(h_1 | t_1, I_0) = 190/(190 + 270) = 0.413$.

However, it should be noted that while they are equivalent when applied under the conditions where Bayes inference can also be applied, namely, with $w_0 = w_1 = 1$ in (10) and

Authorized licensed use limited to: University of Manchester. Downloaded on April 11,2023 at 09:43:10 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE V
LIKELIHOOD FROM EXPERIMENT

Likelihood		Random test result (x_k)	
		Defective ($x_k=1$)	Faultless ($x_k=0$)
State	Good lot (h_1)	p_1	$1-p_1$
	Bad lot (h_2)	p_2	$1-p_2$

TABLE VI
BASIC PROBABILITY ACQUIRED FROM LIKELIHOOD

Basic probability		Test result (x_k)	
		$e_{1,k}$ ($x_k=1$)	$e_{0,k}$ ($x_k=0$)
State	h_1	$p_{1,1k}=\frac{p_1}{p_1+p_2}$	$p_{1,0k}=\frac{1-p_1}{2-p_1-p_2}$
	h_2	$p_{2,1k}=\frac{p_2}{p_1+p_2}$	$p_{2,0k}=\frac{1-p_2}{2-p_1-p_2}$

ambiguous data cleaned, the ER and Bayesian inference are not the same in general. The meaning of $p(t_1|h_1,I_0)p(h_1|I_0)$, which is the numerator in (5) for calculating $p(h_1|t_1,I_0)$ in Bayesian inference, is not the same as $p_{1,1}p_{1,0}$, which is the numerator in (19) for calculating $p_{h_1,e_0\wedge e_1}$ in the ER rule. The former is the joint probability of the occurrence of both positive test result t_1 and the state "steroid user h_1" given the cleaned sample data of the experiment (I_0), while $p_{1,1}p_{1,0}$ is the joint probability mass that the positive test result and the prior distribution e_0 independently support the detection of steroid user h_1. More specifically, we can see that $p(h_1|I_0)=p_{1,0}=0.1$ but $p(t_1|h_1,I_0)=0.95\neq p_{1,1}=0.8636$. ■

E. Illustration of the ER Rule for Recursive Likelihood Inference

In the previous section, it was shown that Theorem 1 underpins the ER rule so that it can perform equivalent Bayesian inference symmetrically. In this section, another example is used to illustrate the recursive nature of ER for likelihood inference.

Example 2: Suppose a company supplies large quantities of parts by lots to its customers. Each lot contains some defective parts and can be classified as good lot (h_1) or bad lot (h_2). There is different cost implication if a good or bad lot is shipped to a customer, so the company conducts n independent random tests before any shipment to minimize cost. The likelihood of getting a defective part by randomly picking up a part from a good (or bad) lot is p_1 (or p_2). Suppose h_1 and h_2 are two mutually exclusive and collectively exhaustive states, x_k is the result from the kth random test, which can be defective ($x_k=1$) or faultless ($x_k=0$), as shown in Table V, and a part is put back to a lot after test. If n random tests are conducted for shipment decision making, with l of them being defective and the rest being faultless, what is the probability that h_1 (or h_2) is true?

Note that mathematically Example 2 is the same as the example analyzed in [30], where inference based on a carefully constructed belief function can only approximate Bayesian inference in general. In this section, we demonstrate that the ER rule generates the same result as Bayesian inference does when evidence is acquired by applying the proposed likelihood analysis method.

In this example, we have $\Theta=\{h_1,h_2\}$. Let $e_{j,k}$ stand for the evidence of observing $x_k=j$ and $p_{i,jk}$ for the basic probability that evidence $e_{j,k}$ points to state h_i, with $i\in\{1,2\}$ and $j\in\{0,1\}$. Applying Theorem 1, we can acquire evidences $e_{1,k}$ and $e_{0,k}$ from Table V, as shown in Table VI, and they are profiled as follows:

$$e_{1,k}=\{(h_1,p_{1,1k}),(h_2,p_{2,1k})\}\text{ and}$$
$$e_{0,k}=\{(h_1,p_{1,0k}),(h_2,p_{2,0k})\}.$$

Since the n test results are independently generated, they can be combined recursively using the ER rule in any order without changing the final result [46]. As such, without loss of generality, let us initially combine the first group of all the l test results that are defective or $x_k=1(k=1,\ldots,l)$, then the second group of all the $n-l$ test results that are faultless or $x_k=0(k=l+1,\ldots,n)$, and finally the two groups to generate the required probabilities. Let the outcomes be denoted by $p_{h_i,e_{1,1}\wedge\cdots\wedge e_{1,l}}$, $p_{h_i,e_{0,l+1}\wedge\cdots\wedge e_{0,n}}$, and $p_{h_i,(e_{1,1}\wedge\cdots\wedge e_{1,l})\wedge(e_{0,l+1}\wedge\cdots\wedge e_{0,n})}i\in\{1,2\}$, respectively.

Suppose each test result is regarded as fully reliable, meaning $w_0=w_1=1$ in (10). Note that from Table V, there is no ambiguity in data and therefore basic probability assigned to Θ is zero in $e_{1,k}$ and $e_{0,k}$ for any k. From (10) or (13), we then get

$$\begin{aligned}\tilde{m}_{h_i,e_{1,1}\wedge e_{1,2}}&=p_{i,11}\times p_{i,12}\\&=\frac{p_i}{p_1+p_2}\times\frac{p_i}{p_1+p_2}=\frac{(p_i)^2}{(p_1+p_2)^2}\end{aligned}$$

so $p_{h_i,e_{1,1}\wedge e_{1,2}}=[(p_i)^2/\bar{c}_{12}]$, where $\bar{c}_{12}=\sum_{t=1}^2(p_t)^2$ is a constant with respect to i.

Next, let us suppose $p_{h_i,e_{1,1}\wedge\cdots\wedge e_{1,l-1}}=(p_i)^{l-1}/\bar{c}_{1\cdots l-1}$ with $\bar{c}_{1\cdots l-1}=\sum_{t=1}^2(p_t)^{l-1}$ being a constant with respect to i. Applying the ER rule (10) again to combine $e_{1,1},\ldots,e_{1,l-1}$ with $e_{1,l}$ leads to

$$\tilde{m}_{h_i,e_{1,1}\wedge\cdots\wedge e_{1,l}}=\frac{(p_i)^{l-1}}{\bar{c}_{1\cdots l-1}}\times\frac{p_i}{p_1+p_2}=\frac{(p_i)^l}{\bar{c}_{1\cdots l-1}(p_1+p_2)}$$

so $p_{h_i,e_{1,1}\wedge\cdots\wedge e_{1,l}}=[(p_i)^l/\bar{c}_{1\cdots l}]$, where $\bar{c}_{1\cdots l}=\sum_{t=1}^2(p_t)^l$ is a constant with respect to i.

Following a similar process, we can combine $e_{0,l+1},\ldots,e_{0,n}$ and get $p_{h_i,e_{0,l+1}\wedge\cdots\wedge e_{0,n}}=(1-p_i)^{n-l}/\bar{c}_{l+1\cdots n}$, where $\bar{c}_{l+1\cdots n}=\sum_{t=1}^2(1-p_t)^{n-l}$ is a constant with respect to i.

Applying the ER rule again to combine $(e_{1,1}\wedge\cdots\wedge e_{1,l})$ with $(e_{0,l+1}\wedge\cdots\wedge e_{0,n})$ leads to

$$\begin{aligned}&\tilde{m}_{h_i,(e_{1,1}\wedge\cdots\wedge e_{1,l})\wedge(e_{0,l+1}\wedge\cdots\wedge e_{0,n})}\\&=\frac{(p_i)^l}{\bar{c}_{1\cdots l}}\times\frac{(1-p_i)^{n-l}}{\bar{c}_{l+1\cdots n}}=\frac{(p_i)^l(1-p_i)^{n-l}}{\bar{c}_{1\cdots l}\bar{c}_{l+1\cdots n}}.\end{aligned}$$

Note that $\bar{c}_{1\cdots l}$ and $\bar{c}_{l+1\cdots n}$ are each constant with respect to i. From (10), the required probabilities are finally generated by

Authorized licensed use limited to: University of Manchester. Downloaded on April 11,2023 at 09:43:10 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE VII
SAMPLE DATA WITH AMBIGUITY

Frequency		Test results					Total observation
		t_1	$\cdots$	t_j	$\cdots$	t_L	
Single state or a subset of states	θ_1	$s_{1,1}$	$\cdots$	$s_{1,j}$	$\cdots$	$s_{1,L}$	S_1
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	θ	$s_{\theta,1}$	$\cdots$	$s_{\theta,j}$	$\cdots$	$s_{\theta,L}$	S_θ
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	θ_{2^N-1}	$s_{2^N-1,1}$	$\cdots$	$s_{2^N-1,j}$	$\cdots$	$s_{2^N-1,L}$	S_{2^N-1}
Total test		T_1	$\cdots$	T_j	$\cdots$	T_L	S

the following normalization [46]:

$$\begin{aligned} &p_{h_i,e_{1,1}\wedge\cdots\wedge e_{1,l}\wedge e_{0,l+1}\wedge\cdots\wedge e_{0,n}} \\ &\quad= \frac{\widetilde{m}_{h_i,(e_{1,1}\wedge\cdots\wedge e_{1,l})\wedge(e_{0,l+1}\wedge\cdots\wedge e_{0,n})}}{\sum_{t=1}^{2}\widetilde{m}_{h_t,(e_{1,1}\wedge\cdots\wedge e_{1,l})\wedge(e_{0,l+1}\wedge\cdots\wedge e_{0,n})}} \\ &\quad= \frac{(p_i)^l(1-p_i)^{n-l}}{\sum_{t=1}^{2}(p_t)^l(1-p_t)^{n-l}}. \end{aligned}$$

III. LIKELIHOOD ANALYSIS AND INFERENCE WITH IMBALANCED AND AMBIGUOUS DATA

In the previous section, the likelihood method for analyzing imbalanced data without ambiguity was proposed to enable the ER rule to perform Bayesian inference symmetrically and equivalently. In this section, this method will be generalized to analyze imbalanced data with ambiguity, as shown in Table II, without resampling, deleting, or imputing ambiguous data.

A. Likelihood Analysis of Imbalanced and Ambiguous Data

In the previous section, ambiguous data was deleted for conventional Bayesian analysis. In this section, ambiguous data is taken into account as it stands with its impact on inference and decision making evaluated. Table VII illustrates a general inference problem with N mutually exclusive and collectively exhaustive singular states and L test results, which cover all possible outcomes of the test, categorical, or discretised, where data can be ambiguous in the sense that a test result can point to any subset of singular states. In Table VII, θ stands for state, which can be a singular state or a subset of singular states, t_j for the jth test result, $s_{\theta,j}$ for the number of observations where both state θ is true and test result t_j is observed, S_θ for the number of all observations where state θ is true, T_j for the number of all observations where test result t_j is observed, and S for the total number of all observations in the sample. So, $S_\theta = \sum_{j=1}^{L} s_{\theta,j}$, $T_j = \sum_{\theta\subseteq\Theta} s_{\theta,j}$, and $S = \sum_{\theta\subseteq\Theta} S_\theta = \sum_{j=1}^{L} T_j$.

Let $p_{\theta,0}$ stand for the sample prior probability of state θ, and $c_{\theta,j}$ for the likelihood that the jth test result (t_j) is expected to occur given θ. Both are calculated from the data of Table VII by

$$p_{\theta,0} = S_\theta/S \text{ and } c_{\theta,j} = s_{\theta,j}/S_\theta \quad \forall\theta\subseteq\Theta \text{ and } j = 1,\ldots,L \tag{20}$$

Let $p_{\theta,j}$ stand for the basic probability that the jth test result points to state θ, with $\sum_{\theta\subseteq\Theta} p_{\theta,j} = 1$ for any $j = 1,\ldots,L$. Suppose e_0 stands for the sample prior distribution with weight w_0, and e_j for the evidence mapped from the jth test result with weight w_j, profiled as follows:

$$e_0 = \left\{(\theta,\ p_{\theta,0}),\ \theta\subseteq\Theta,\ \sum_{\theta\subseteq\Theta} p_{\theta,0} = 1\right\} \tag{21}$$

$$e_j = \left\{(\theta,\ p_{\theta,j}),\ \theta\subseteq\Theta,\ \sum_{\theta\subseteq\Theta} p_{\theta,j} = 1\right\} j = 1,\ldots,L. \tag{22}$$

The first question is how to generate basic probability $p_{\theta,j}$ from imperfect data so that the ER process constitutes a probabilistic inference process even when data are imperfect. To answer this question, it is asserted that both the principle of likelihood and Bayesian principle be followed. This assertion leads to the following likelihood analysis method to acquire evidence from imperfect data.

Theorem 2: Suppose two pieces of evidence e_0 and e_j ($j = 1,\ldots,L$) are defined in (21) and (22), respectively. The joint probability that both e_0 and e_j support state θ, generated by applying the ER rule, has the same evidential meaning as the posterior probability that state θ is true given e_0 and the observation of test result t_j if and only if basic probability $p_{\theta,j}$ is acquired as follows:

$$p_{\theta,j} = c_{\theta,j}\Big/\left(\sum_{A\subseteq\Theta} c_{A,j}\right) \quad \forall\theta\subseteq\Theta \text{ and } j = 1,\ldots,L. \tag{23}$$

Proof: See Section S1.3 proof of Theorem 2 of the Supplementary Material part for this article. ■

Similar to the interpretation for Theorem 1, the rationale of Theorem 2 can also be interpreted in terms of relative frequency by noting that $p_{\theta,j}$ is taken as the ratio of $s_{\theta,j}$ over the sum of $s_{A,j}$ rescaled to S_θ for all $A\subseteq\Theta$, or

$$\begin{aligned} p_{\theta,j} &= \frac{s_{\theta,j}}{s_{\theta,j} + \sum_{A\subseteq\Theta,A\neq\theta}(s_{A,j}/S_A)\times S_\theta} \\ &= \frac{(s_{\theta,j}/S_\theta)}{(s_{\theta,j}/S_\theta) + \sum_{A\subseteq\Theta,A\neq\theta}(s_{A,j}/S_A)} = \frac{c_{\theta,j}}{\sum_{A\subseteq\Theta} c_{A,j}}. \end{aligned} \tag{24}$$

This means that evidence e_j is generated without relying on the prior distribution of the sample data, or the former is independent of the latter. It is also worth noting that the acquisition of evidence e_j using (23) makes full use of all sample data without resampling. Theorem 2 establishes a mapping from imperfect data to evidence in the sense that the observation of a test result is transformed into a piece of evidence, independent of the prior and other evidence from subsequent tests if the tests are conducted independently.

Theorem 2 asserts that in the ER framework evidence must be acquired as normalized likelihoods in order that the combination of such evidence using the ER rule constitutes a probabilistic inference process. The following corollary shows that such an ER process is a likelihood inference process and is independent of sample prior, or irrespective of whether data is imbalanced or not.

Authorized licensed use limited to: University of Manchester. Downloaded on April 11,2023 at 09:43:10 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Corollary 3: Suppose evidence e_j is acquired by assigning basic probability $p_{\theta,j}$ using (23) for any $\theta \subseteq \Theta$, and is fully reliable. If e_j is combined with an unknown prior (e_0) that is also fully reliable and has the basic probability of one assigned to the system space and zero to any other subsets of singular states, the resultant probabilities are the same as the basic probabilities of e_j.

Proof: See Section S1.4 proof of Corollary 3 of the Supplementary Material part for this article. ■

B. Illustration of Likelihood Inference With Imperfect Data

This section is moved to the Supplementary Materials part for this article due to the page limit. The purpose of this section is to analyze imbalanced data with ambiguity to support evidence-based decision making and come back to investigate Example 1. The details of this section can be found from Section "S2. Illustration of likelihood inference with imperfect data" of the Supplementary Material part.

C. Likelihood Inference for Fault Diagnosis

This section is moved to the Supplementary Materials part for this article due to the page limit. The purpose of this section is to use a case study about fault diagnosis for rail track maintenance management to demonstrate how the ER rule can be implemented for likelihood inference with ambiguous data collected from an engineering system. The details of this section can be found from Section "S3. Likelihood inference for fault diagnosis" of the Supplementary Material part.

IV. Conclusion

In this article, the two questions were investigated: 1) how to acquire evidence from imbalanced data without resampling and 2) how to deal with ambiguous data in the processes of acquiring and combining evidence in a principled manner for probabilistic inference and evidence-based decision making. After a brief introduction to Bayes' rule and the ER rule, the relationship between them was investigated. By following the principle of likelihood and Bayesian principle, a new sufficient and necessary condition was established to acquire evidence from imbalanced data without ambiguity, showing that basic probability must and must only be generated as normalized likelihood. Under this condition, the combination of prior probability and basic probability using the ER rule leads to the same posterior probability as what Bayesian inference generates when there is no ambiguity in data. It was also shown that when there is no prior information available, the ER rule operates as a likelihood inference process with prior taken as total ignorance. The outcome of the likelihood inference is the same as that of Bayesian inference by assuming uniform prior. Two examples in decision analysis under uncertainty were analyzed to demonstrate how the ER rule can be used for symmetrical Bayesian inference when prior distribution is known and for likelihood inference when prior is unknown.

The above investigation was then extended to construct a new sufficient and necessary condition for probabilistic inference with both imbalanced and ambiguous data. Underpinned by this condition, the inference process based on the ER rule is established as a framework for likelihood inference, where the basic probability can be assigned to singular states or any subset of singular states to deal with ambiguity in data. In this general framework, imbalanced and ambiguous data from different sources can be used for probabilistic inference without resampling, deleting, or imputing data. An example in decision making under uncertainty and a case study for fault diagnosis in railway track maintenance management were examined to elaborate how this general likelihood inference process can be implemented to acquire and combine evidence when there is ambiguity in data and for identification of low probability events such as system faults.

In summary, the research finds that in principle evidence must and must only be acquired as a normalized likelihood distribution for probabilistic inference in the ER framework. The proposed likelihood method provides a principled and practical means for analyzing imperfect data in situations where data is routinely recorded from different sources, instead of from carefully designed and strictly controlled laboratory conditions, and thus may not capture true system prior conditions. The case study demonstrated the application of the likelihood method to analyze data for fault diagnosis in engineering systems, with more applications of the method reported in other papers from healthcare to financial services, although its potential applications are by no means limited to these areas.

The likelihood method developed in this article can be used to acquire evidence from data sources individually under the assumption that one data source is independent of another in the process of evidence acquisition. However, this assumption may not always be true. There is therefore a need to investigate how to acquire evidence from data sources that are not independent of each other and how to use such evidence to support probabilistic inference and evidence-based decision analysis.

References

[1] K. AbuDahab, D.-L. Xu, and Y.-W. Chen, "A new belief rule base knowledge representation scheme and inference methodology using the evidential reasoning rule for evidence combination," *Exp. Syst. Appl.*, vol. 51, pp. 218–230, Jun. 2016.

[2] M. Aickin, "Connecting Dempster–Shafer belief functions with likelihood-based inference," *Synthese*, vol. 123, no. 3, pp. 347–364, 2000.

[3] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[4] Y.-W. Chen, J.-B. Yang, D.-L. Xu, and S.-L. Yang, "On the inference and approximation properties of belief rule based systems," *Inf. Sci.*, vol. 234, pp. 121–135, Jun. 2013.

[5] A. Birnbaum, "On the foundations of statistical inference," *J. Amer. Stat. Assoc.*, vol. 57, no. 298, pp. 269–306, 1962.

[6] C. P. Robert, N. Chopin, and J. Rousseau, "Harold Jeffreys's theory of probability revisited," *Stat. Sci.*, vol. 24, no. 2, pp. 141–172, 2009.

[7] N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "SMOTE: Synthetic minority oversampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[8] R. T. Cox, "Probability, frequency and reasonable expectation," *Amer. J. Phys.*, vol. 14, no. 1, pp. 1–10, 1946.

[9] R. T. Cox, *The Algebra of Probable Inference*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1961.

[10] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Ann. Math. Stat.*, vol. 38, no. 2, pp. 325–339, 1967.

[11] A. P. Dempster, "A generalization of Bayesian inference," *J. Roy. Stat. Soc. B Methodol.*, vol. 30, no. 2, pp. 205–247, 1968.

Authorized licensed use limited to: University of Manchester. Downloaded on April 11,2023 at 09:43:10 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.

[13] A. P. Dempster, "The Dempster–Shafer calculus for statisticians," *Int. J. Approx. Reason.*, vol. 48, no. 2, pp. 365–377, 2008.

[14] A. Donner, "The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values," *Amer. Stat.*, vol. 36, no. 4, pp. 378–381, 1982.

[15] R. A. Fisher, *Statistical Methods and Scientific Inference*, 3rd ed. New York, NY, USA: Hafner, 1973.

[16] D. A. S. Fraser, N. Reid, E. Marras, and G. Y. Yi, "Default priors for Bayesian and frequentist inference," *J. Roy. Stat. Soc. B Stat. Methodol.*, vol. 72, no. 5, pp. 631–654, 2010.

[17] R. M. Hamer and P. M. Simpson, "Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials," *Amer. J. Psychiat.*, vol. 166, no. 6, pp. 639–641, 2009.

[18] A. Gelman, J.-B. Carlin, H.-S. Stern, D.-B. Dunson, A. Vehtari, and D.-B. Rubin, *Bayesian Data Analysis*. New York, NY, USA: CRC Press, 2014.

[19] C. Howson and P. Urbach, *Scientific Reasoning: The Bayesian Approach*, 3rd ed. Chicago, IL, USA: Open Court, 2005.

[20] H. Jeffreys, *Theory of Probability*, 3rd ed. Oxford, U.K.: Oxford Univ. Press, 1961.

[21] H. Kang, "The prevention and handling of the missing data," *Korean J. Anaesthesiol.*, vol. 64, no. 5, pp. 402–406, 2013.

[22] G. L. Kong et al., "Combining principal component analysis and the evidential reasoning approach for healthcare quality assessment," *Ann. Oper. Res.*, vol. 271, no. 2, pp. 679–699, 2018.

[23] G. L. Kong, D.-L. Xu, J.-B. Yang, T. B. Wang, and B. G. Jiang, "Evidential reasoning rule-based decision support system for predicting ICU admission and in-hospital death of trauma," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 11, pp. 7131–7142, Nov. 2021, doi: 10.1109/TSMC.2020.2967885.

[24] S. Kotsiantis and P. E. Pintelas, "Mixture of expert agents for handling imbalanced data sets," *Ann. Math. Comput. TeleInformat.*, vol. 1, no. 1, pp. 46–55, 2003.

[25] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 30, pp. 1–12, Nov. 2005.

[26] J. S. Lee, S. Choi, S.-S. Kim, C. Park, and Y. G. Kim, "A mixed filtering approach for track condition monitoring using accelerometers on the axle box and bogie," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 3, pp. 749–758, Mar. 2012.

[27] N. K. Malhotra, "Analyzing marketing research data with incomplete information on the dependent variable," *J. Market. Res.*, vol. 24, no. 1, pp. 74–84, 1987.

[28] S. Sachan, J.-B. Yang, D.-L. Xu, D. Erasobenavides, and Y. Li, "An explainable AI decision-support-system to automate loan underwriting," *Exp. Syst. Appl.*, vol. 144, Apr. 2020, Art. no. 113100.

[29] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ, USA: Princeton Univ. Press, 1976.

[30] G. Shafer, "Belief functions and parametric models," *J. Roy. Stat. Soc. B Stat. Methodol.*, vol. 44, no. 3, pp. 322–352, 1982.

[31] G. Shafer and J. Pearl, *Readings in Uncertain Reasoning*. San Mateo, CA, USA: Morgan Kaufmann, 1990.

[32] G. Shafer, "A betting interpretation for probabilities and Dempster–Shafer degrees of belief," *Int. J. Approx. Reason.*, vol. 52, no. 2, pp. 127–136, 2011.

[33] S. Sinharay, H. S. Stern, and D. Russell, "The use of multiple imputation for the analysis of missing data," *Psychol. Methods*, vol. 6, no. 4, pp. 317–329, 2001.

[34] F. Smarandache, J. Dezert, and J.-M. Tacnet, "Fusion of sources of evidence with different importances and reliabilities," in *Proc. 13th IEEE Conf. Inf. Fusion (FUSION)*, 2010, pp. 1–8.

[35] P. Smets, "Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem," *Int. J. Approx. Reason.*, vol. 9, no. 1, pp. 1–35, 1993.

[36] K. S. Van Horn, "Constructing a logic of plausible inference: A guide to Cox's theorem," *Int. J. Approx. Reason.*, vol. 34, no. 1, pp. 3–24, 2003.

[37] D.-L. Xu, J.-B. Yang, and Y.-M. Wang, "The ER approach for multi-attribute decision analysis under interval uncertainties," *Eur. J. Oper. Res.*, vol. 174, no. 3, pp. 1914–1943, 2006.

[38] X. Xiaobin, Z. Jin, Y. Jian-Bo, X. Dong-Ling, and C. Yu-Wang, "Data classification using evidence reasoning rule," *Knowl.-Based Syst.*, vol. 116, pp. 144–151, Jan. 2017.

[39] X. J. Xu, X. P. Yan, C. X. Sheng, C. Q. Yuan, D. L. Xu, and J. B. Yang, "A belief rule based expert system for fault diagnosis of marine diesel engines," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 2, pp. 656–672, Feb. 2020.

[40] J.-B. Yang and M. G. Singh, "An evidential reasoning approach for multiple attribute decision making with uncertainty," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 24, no. 1, pp. 1–18, Jan. 1994.

[41] J.-B. Yang and P. Sen, "A general multi-level evaluation process for hybrid MADM with uncertainty," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 24, no. 10, pp. 1458–1473, Oct. 1994.

[42] J.-B. Yang, "Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainties," *Eur. J. Oper. Res.*, vol. 131, no. 1, pp. 31–61, 2001.

[43] J.-B. Yang and D.-L. Xu, "On the evidential reasoning algorithm for multiattribute decision analysis under uncertainty," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 32, no. 3, pp. 289–304, May 2002.

[44] J.-B. Yang, J. Liu, J. Wang, H.-S. Sii, and H.-W. Wang, "A belief rule-base inference methodology using the evidential reasoning approach—RIMER," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 36, no. 2, pp. 266–285, Mar. 2006.

[45] J.-B. Yang, J. Liu, D.-L. Xu, J. Wang, and H. W. Wang, "Optimization models for training belief-rule-based systems," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 37, no. 4, pp. 569–585, Jul. 2007.

[46] J.-B. Yang and D.-L. Xu, "Evidential reasoning rule for evidence combination," *Artif. Intell.*, vol. 205, pp. 1–29, Dec. 2013.

[47] J.-B. Yang and D.-L. Xu, "A study on generalising Bayesian inference to evidential reasoning," in *Belief Function: Theory and Applications* (Lecture Notes in Artificial Intelligence), F. Cuzzolin, ed. Cham, Switzerland: Springer, 2014, pp. 180–189.

Jian-Bo Yang received the B.Eng. and M.Eng. degrees in control engineering from North Western Polytechnic University, Xi'an, China, in 1981 and 1984, respectively, and the Ph.D. degree in decision science and systems engineering from Shanghai Jiao Tong University, Shanghai, China, in 1987.

He is the Chair Professor of Decision and Systems Sciences and the Director of the Decision and Cognitive Sciences Research Centre with the Alliance Manchester Business School, The University of Manchester, Manchester, U.K. He is also the Advisory Professor with the Management School, Hefei University of Technology, Hefei, China, and the Distinguished Visiting Professor with the Ji Mei University, Xiamen, China. He is the Principal Investigator or Co-Investigator for over 70 research projects at a total value of over £21M, funded by many organizations, including U.K. Engineering and Physical Science Research Council, U.K. Economic and Social Research Council, Innovate U.K., U.K. Department of Environment, Food and Rural Affairs, European Commission, Natural Science Foundation of China, and Hong Kong Research Grant Council and industry. He has published four books, over 270 journal papers and book chapters, and a similar number of conference papers, with extensive citations, such as in Web of Science and Google Scholar, and developed several software packages in optimization and decision support with wide applications worldwide. Over the last three decades, he has conducted theoretical and methodological research in many areas, including the evidential reasoning (ER) theory; multiple criteria decision analysis under uncertainties; probabilistic inference and decision analysis using both data and judgments; multiple objective optimization; intelligent system modeling; explainable artificial intelligence (AI) and interpretable machine learning; hybrid decision methodologies and technologies combining concepts and techniques from decision science, systems science, operational research, and AI. His current applied research covers a wide range of applications driven by data, powered by AI and enabled by ER, including modeling and decision support for professional services, such as finance, insurance, and healthcare; diagnosis and prognosis, design and operation decision analysis in healthcare, engineering, and social systems; pattern identification and analysis of consumer behaviors; analysis of public sentiments and system risks (financial or nonfinancial); new product development; aggregated production management; system maintenance management; risk and security modeling and analysis; and performance analysis and improvement of products, processes, and organizations.

Authorized licensed use limited to: University of Manchester. Downloaded on April 11,2023 at 09:43:10 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Dong-Ling Xu received the B.Eng. degree in electrical engineering from Hunan University, Changsha, China, in 1983, and the M.Eng. and Ph.D. degrees in system control engineering from Shanghai Jiao Tong University, Shanghai, China, in 1986 and 1988, respectively.

She is the Chair Professor of Decision Science and Systems with the Alliance Manchester Business School, The University of Manchester, Manchester, U.K. Over the past 30 years, she has been conducting research in data analysis, statistical inference, machine learning, and decision support systems under uncertainty; system and process modeling; statistical fault detection, system development, and their applications in a wide range of areas. She has developed several interactive web-based decision support systems and co-developed a Windows-based Software Package called Intelligent Decision System (IDS) via Evidential Reasoning for general purpose multiple criteria decision analysis, and several statistical fault detection systems. Those systems are used in a wide range of decision making and risk assessment activities by organizations, such as General Motors, Tesco, NHS, Ford, Shell, BP, and CNOOC in areas, such as healthcare and finance; modeling and analysis of system safety and security; and organizational self-assessment in quality management. The Evidential Reasoning approach and the IDS software is used by practitioners and researchers from over 50 countries. She has published over 100 peer reviewed journal papers, book chapters, and books.

Xiaobin Xu received the Ph.D. degree in power electronics and power transmission from Shanghai Maritime University, Shanghai, China, in 2009.

He is a Professor with the Department of Automation and the Belt and Road Information Technology Research Institute, Hangzhou Dianzi University, Hangzhou, China. His research interests include evidence theory, fuzzy-set theory, and applications in the processing of uncertain information, the reliability analysis, safety evaluation, and condition monitoring of complex industrial systems.

Chao Fu received the Ph.D. degree from Hefei University of Technology, Hefei, China, in 2009.

He is currently a Professor with the School of Management, Hefei University of Technology. He has published about 60 articles in international journals. He has conducted theoretical and applicable research in many areas, including multiple criteria-decision analysis based on evidential reasoning theory, data-driven decision analysis, data-driven behavioral analysis, and medical auxiliary diagnosis.

Authorized licensed use limited to: University of Manchester. Downloaded on April 11,2023 at 09:43:10 UTC from IEEE Xplore. Restrictions apply.