Are We Training with The Right Data? Evaluating Collective Confidence in Training Data using Dempster Shafer Theory

Sangeeta Dey
deyesangeeta@ajou.ac.kr
Ajou University
Suwon, Gyeonggi-do, South Korea

Seok-Won Lee
leesw@ajou.ac.kr
Ajou University
Suwon, Gyeonggi-do, South Korea

ABSTRACT
The latest trend of incorporating various data-centric machine learning (ML) models in software-intensive systems has posed new challenges in the quality assurance practice of software engineering, especially in a high-risk environment. ML experts are now focusing on explaining ML models to assure the safe behavior of ML-based systems. However, not enough attention has been paid to explain the inherent uncertainty of the training data. The current practice of ML-based system engineering lacks transparency in the systematic fitness assessment process of the training data before engaging in the rigorous ML model training. We propose a method of assessing the collective confidence in the quality of a training dataset by using Dempster Shafer theory and its modified combination rule (Yager’s rule). With the example of training datasets for pedestrian detection of autonomous vehicles, we demonstrate how the proposed approach can be used by the stakeholders with diverse expertise to combine their beliefs in the quality arguments and evidences about the data. Our results open up a scope of future research on data requirements engineering that can facilitate evidence-based data assurance for ML-based safety-critical systems.

CCS CONCEPTS
- Software and its engineering → Risk management; Collaboration in software development; - Mathematics of computing → Hypothesis testing and confidence interval computation.

KEYWORDS
data uncertainty, safety, machine learning, Dempster Shafer theory

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Preprint: ICSE-NIER ’22, May 21–29, 2022, USA
© 2022 Association for Computing Machinery.
ACM ISBN XXX-X-XXX-XXXX-X/XX/XX...
https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

1 INTRODUCTION
With the recent advances of machine learning (ML) techniques software-intensive systems are now heavily relying on ML models. While the enhanced performance of ML techniques seems to be very promising, their black-box nature and data dependency have affected the quality assurance practice of software engineering, especially in safety-critical environments [1][7]. Many ML researchers are now focusing on explaining ML models to assure the safe behavior of ML-based systems. However, not enough attention has been paid to explain the training data quality that is one of the key contributors to the performance of such ML models. As rightly argued by Rasouli et al. [9], without having a proper understanding of the uncertainty in the training data, it will be futile to focus only on reducing model uncertainty to ensure safety. The current practice of ML-based system engineering lacks transparency in the process of a systematic fitness assessment of the training data before delving into the rigorous ML model training. Tacit knowledge of software engineers, lack of collaboration between the system-level experts and the ML component level experts, insufficient metrics to measure the collective confidence in the quality of the training data make the quality assurance process of such systems extremely obscure. This challenge motivates us to design a method that can be helpful for the experts with various expertise to collectively assess the training data quality.

Data quality assessment in the era of big data and ML has already been discussed in recent years [3][6][11]. While these approaches are effective to measure data uncertainty by statistical metrics, they do not consider the uncertainty in the quality arguments that can arise due to the conflicts in the judgments of the diverse stakeholders, such as, ML expert, data expert, domain expert, system engineer. We argue that an integrated engineering perspective is more effective to assess the overall fitness of training data from various perspectives of the diverse experts. In the case of ML models, there are numerous sources of uncertainty in training data; such as representation gap, lack of coverage of rare cases. We can attempt to minimize the uncertainty by guiding on collection and augmentation necessities of training data. Furthermore, by performing early exploratory analysis on a collected training dataset it is possible to support the claim that “data uncertainty has been minimized” by producing relevant pieces of evidence and arguments. However, not all the evidences and their corresponding arguments are fully reliable. Therefore, unlike a traditional assurance case that only depicts claims in binary states (satisfied and unsatisfied), we aim to introduce scope to express uncertainty and conflicts in the judgments of the diverse experts. We propose a step-wise method that uses Dempster Shafer (DS) Theory of Evidence [10] and its modified combination rule (Yager’s rule) [12] to combine the beliefs of component level experts (ML experts) and system-level experts (domain experts, systems engineers, etc.) in the quality of the training data. With the help of CALTECH [5] and JAAD [8] dataset for...
pedestrian detection of autonomous vehicles, we demonstrate how the proposed approach can help experts collaborate to evaluate the collective confidence in the training data quality.

Our results present the efficacy of using evidence theory as it rightly depicts that upon collecting training data that can represent the problem domain better, the overall confidence in the data increases. Our research opens up a scope for future research on the field of data requirements engineering that can facilitate evidence-driven data assurance for ML-based safety-critical systems.

2 DEMPSTER SHAFFER THEORY AND COLLECTIVE CONFIDENCE

‘Intellectual diversity’ or the concept of ‘a different pair of eyes’ has long been used in engineering to reduce human error. Bev Littlewood has extensively studied the use of diverse arguments to increase confidence in the dependability claims of software systems [2]. The triad of uncertainty, diversity, and confidence still plays an important role in the days of AI to explain complex system behavior. Especially, for software-intensive systems that depend on data-driven ML models, data uncertainty is two-fold. (i) inherent uncertainty of the data (ii) uncertainty in the safety claims and arguments about the data. Although data experts and ML experts can assess the data uncertainty through data exploration, their judgment needs to be evaluated from the perspective of system-level experts like system engineers, domain expert, safety engineer, etc. as their expertise lies in the entirety of the system. We argue that DS Theory [10] can be useful to mathematically assess the collective confidence about the data quality as this theory is more accommodative than other probability-based approaches to embrace the subjective knowledge of the experts.

The objective is to quantify the experts’ beliefs in the uncertainty of the data regarding its use for the training purpose for a safety-critical functionality. In other words, it needs to be concluded whether the collected data is safe or dependable enough to be used for the training purpose or not.

Therefore, the frame of discernment

$$\Theta = \text{Safe}, \text{Unsafe}$$

(1)

And, the power set that contains all possible subsets of $\Theta$

$$P(\Theta) = \{\emptyset, \text{Safe}, \text{Unsafe}, (\text{Safe}, \text{Unsafe})\}$$

(2)

In the running example, we consider that empty set $\emptyset$ is not viable as the experts must give some judgment regarding the quality of the data. DS theory is flexible enough to not only consider the belief in ‘Safe’ and ‘Unsafe’ but also the cases where the experts cannot give a strict judgment due to epistemic uncertainty or insufficient proofs. This is denoted as (Safe, Unsafe) in the power set. For simplicity, we showcase the application of the DS theory by considering the perspectives of only two types of experts. The experts at the ML component level and the system level are responsible to judge the data quality and provide a degree of belief or belief mass denoted by $m_1$ and $m_2$ respectively to each of the possible subsets of the power set. We combine the individual belief masses using Yager’s rule (Modified DS combination rule) [12] as follows:

$$\text{Collective Belief(Safe)} = \sum_{A_1 \cap A_2 = \text{Safe}} m_1(A_1) m_2(A_2)$$

Here, $A_1$ and $A_2$ denote any two focal sets in the power set. We use the modified DS combination rule instead of the classical DS combination rule as the latter produces a counterintuitive result in the case of totally conflicting belief structures. Equation (3) suggests that the collective belief in ‘Safe’ is attributed to not only the individual beliefs in the singleton set ‘Safe’ but also in other focal sets that have a common element ‘Safe’ like (Safe, Unsafe). The detailed approach is described in the following section.

3 ASSESSMENT OF COLLECTIVE CONFIDENCE IN TRAINING DATA

In this section, we describe the proposed step-wise method for training data uncertainty evaluation using a running example from the autonomous car domain. The method is depicted in Figure 1.

Figure 1: Step-wise method to evaluate collective confidence in training data quality

Step-1: Elicit the source of uncertainty
At first, the experts need to analyze the sources of uncertainty that can potentially affect the data quality. Moreover, as ML components are often part of a complex system, it is essential to involve experts at both the system level and the ML component level to analyze the causes of uncertainty. For most of the ML techniques, the representation gap (missing out the relevant features), lack of coverage (not covering the corner cases), lack of data pre-processing (insufficient labeling, annotation, etc.), data corruption, etc. are the main factors to affect data quality. However, the causes of uncertainty can vary based on the data source and the ML technique.

Step-2: Derive uncertainty reduction criteria
Based on the elicited sources of uncertainty, the next step is to derive the criteria that are to be met by the training data in order to reduce the scope of uncertainty. For a safety-critical feature like...
### Table 1: Fitness criteria to be met by the training data for pedestrian detection to reduce uncertainty

<table>
<thead>
<tr>
<th>Source of Uncertainty</th>
<th>Criteria</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Representation Gap</strong></td>
<td>Weather condition variability should be considered while collecting data.</td>
<td>CID-1</td>
</tr>
<tr>
<td></td>
<td>Brightness condition variability should be considered.</td>
<td>CID-2</td>
</tr>
<tr>
<td></td>
<td>Crowd density condition (less crowded and moderately crowded) should be considered.</td>
<td>CID-3</td>
</tr>
<tr>
<td></td>
<td>Spatial conditions variability (various cities) should be considered by collecting data in various cities.</td>
<td>CID-4</td>
</tr>
<tr>
<td></td>
<td>Scale of pedestrian should follow a realistic distribution</td>
<td>CID-5</td>
</tr>
<tr>
<td></td>
<td>Poses of pedestrian are to be marked from attributed testing.</td>
<td>CID-6</td>
</tr>
<tr>
<td></td>
<td>Accessories of pedestrian that can change shape of the pedestrian should be represented in data.</td>
<td>CID-7</td>
</tr>
<tr>
<td></td>
<td>Occlusion ratio variations (partial and heavy occlusion) should be part of data.</td>
<td>CID-8</td>
</tr>
<tr>
<td></td>
<td>Types of pedestrians considering age, sex, etc. need to be marked for future attributed testing.</td>
<td>CID-9</td>
</tr>
<tr>
<td><strong>Lack of data coverage</strong></td>
<td>Combination of risky and rare situations like outliers in height distribution (child), accessories that can change shape (big backpack or umbrella), heavily occluded pedestrian, severe weather condition (dark rainy night), etc. should be considered and attributed in dataset.</td>
<td>CID-10</td>
</tr>
<tr>
<td><strong>Data corruption</strong></td>
<td>Data source is not a third party and every preprocessing step should confirm the stability of the data.</td>
<td>CID-11</td>
</tr>
<tr>
<td><strong>Lack of data pre-processing</strong></td>
<td>Labelling should conform to the objective of the ML model (e.g. classes to be identified in case of classification) like ‘Person’, ‘Crowd/people’. Any annotation should not violate real-world constraints, e.g. height-width ratio of bounding box should be realistic.</td>
<td>CID-12</td>
</tr>
</tbody>
</table>

### Table 2: Evidence collection and assessment for CALTECH pedestrian dataset

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Evidence ID</th>
<th>Evidences collected by data experts</th>
<th>Weight</th>
<th>Assessment of evidence by ML Expert</th>
<th>Assessment of evidence by Domain Expert</th>
<th>Reason for conflict</th>
</tr>
</thead>
<tbody>
<tr>
<td>CID-1</td>
<td>EV-1</td>
<td>Weather: Data is collected in one season only.</td>
<td>3</td>
<td>Negative</td>
<td>Negative</td>
<td>NA</td>
</tr>
<tr>
<td>CID-2</td>
<td>EV-2</td>
<td>Brightness: Data is collected in daytime only.</td>
<td>3</td>
<td>Negative</td>
<td>Negative</td>
<td>NA</td>
</tr>
<tr>
<td>CID-3</td>
<td>EV-3</td>
<td>Crowd: Both single pedestrians and large group of Pedestrians are labelled (‘Person’ and ‘People’)</td>
<td>1</td>
<td>Positive</td>
<td>Positive</td>
<td>NA</td>
</tr>
<tr>
<td>CID-4</td>
<td>EV-4</td>
<td>Spatial diversity: Data is collected in one city only.</td>
<td>3</td>
<td>Negative</td>
<td>Negative</td>
<td>NA</td>
</tr>
<tr>
<td>CID-5</td>
<td>EV-5</td>
<td>Scale: Divided in three categories. Near (80 or more pixels), Medium (30-80 pixels), far (30 or less pixels) based on speed limit 55km/hr</td>
<td>2</td>
<td>Positive</td>
<td>Negative</td>
<td>Considering spatial diversity speed limit is not 55km/hr everywhere as per ODD.</td>
</tr>
<tr>
<td>CID-6</td>
<td>EV-6</td>
<td>Pose: No concrete evidence. Not explicitly attributed.</td>
<td>2</td>
<td>Cannot say</td>
<td>Cannot say</td>
<td>NA</td>
</tr>
<tr>
<td>CID-7</td>
<td>EV-7</td>
<td>Accessories: No concrete evidence. Not explicitly attributed.</td>
<td>2</td>
<td>Cannot say</td>
<td>Cannot say</td>
<td>Accessories like umbrella, backpack can change the appearance of pedestrians.</td>
</tr>
<tr>
<td>CID-8</td>
<td>EV-8</td>
<td>Occlusion: Partial occlusion (1-35%) and heavy occlusion (35-80%).</td>
<td>3</td>
<td>Positive</td>
<td>Positive</td>
<td>NA</td>
</tr>
<tr>
<td>CID-9</td>
<td>EV-9</td>
<td>Age, gender: Not explicitly considered and marked. No concrete evidence.</td>
<td>3</td>
<td>Cannot say</td>
<td>Cannot say</td>
<td>NA</td>
</tr>
<tr>
<td>CID-10</td>
<td>EV-10</td>
<td>Rare case: Although heavy occlusion cases are considered, no markup for accessories, not coverage of the outliers in the height distribution, severe weather conditions.</td>
<td>3</td>
<td>Negative</td>
<td>Negative</td>
<td>NA</td>
</tr>
<tr>
<td>CID-11</td>
<td>EV-11</td>
<td>Data source: Video data is collected by the dataset creator (not the third party), no data corruption is reported.</td>
<td>3</td>
<td>Positive</td>
<td>Positive</td>
<td>NA</td>
</tr>
<tr>
<td>CID-12</td>
<td>EV-12</td>
<td>Labelling: Data is labeled as ‘person’ and ‘people’ for individual and group of people respectively.</td>
<td>3</td>
<td>Positive</td>
<td>Positive</td>
<td>NA</td>
</tr>
<tr>
<td>CID-13</td>
<td>EV-13</td>
<td>Annotation: No clear evidence of the aspect ratio used by the annotators.</td>
<td>2</td>
<td>Cannot say</td>
<td>Cannot say</td>
<td>NA</td>
</tr>
</tbody>
</table>

The pedestrian detection training data should cover the Operational Design Domain (ODD) factors. Moreover, in the case of safety-critical systems, it is necessary to explicitly cover the rare yet risky scenarios in the data as much as possible. For instance, video data for pedestrian detection training should represent the environmental variety (brightness variation, weather variation, various types of pedestrians, etc.). Table 1 shows the sources of uncertainty and the corresponding evaluation criteria for such training data to increase the confidence in the data. The given list is not exhaustive and needs to be updated based on the data type (structured tabular data, image data, video data, etc.) and the problem domain. Although not shown in Table 1, it is possible to distribute weightage among the criteria or the evidences (in Step-4) according to their importance.

**Step 3: Collect pieces of evidence from training data**

In our approach, we assume that after collecting training data, data experts perform the early exploratory data analysis (EDA) on the training dataset and collect evidence that can be used as supporting/refuting regarding the satisfaction of the elicited criteria. In this example, to avoid any bias from our side, we referred to the statistical analysis done by the data creators of the CALTECH dataset to collect evidences. By evidence collection, we mean explicitly looking for proof that the criteria have been met by the training data. A few guidelines on EDA can be found in the replication package [4]. The detailed process of EDA is not in the scope of the paper.

**Step 4: Assess the contribution and weights of the evidences**

The ML expert and the domain expert need to evaluate individually each of the pieces of collected evidence and label them as positive (supports the satisfaction of the quality criteria), negative (refutes the satisfaction of the quality criteria), or cannot say (insufficient evidence). If required, data expert may come up with more evidences for that criteria. For instance, the distribution of the ODD factors in the dataset can be measured and produced as evidence.
to argue its representativeness. We evaluated the evidences from the perspectives of ML expert and domain expert. Weightage was also given on the scale of 1-3 to the evidence as not all the pieces of evidence will be equally important as shown in Table 2. Any conflict between the experts needs to be documented for further discussions. These discussions can eventually help to update either the quality criteria or the data augmentation requirements to achieve increased confidence in the data.

**Step-5: Consolidate the assessment of the experts**
We use DS theory to mathematically assess the collected evidences. Based on the evaluation by the experts, individual belief masses for ‘safe’, ‘unsafe’, ‘uncertain’ are calculated by considering the contribution and weightage of each of the pieces of evidence.

\[
Individual \text{ Belief (Safe)} = \sum \frac{\text{Weights of the positive evidence}}{\text{Total weight of all the evidence}} \tag{4}
\]

Similarly, the ‘negative’ and ‘cannot say’ cases contribute to the beliefs in ‘Unsafe’ and (Safe, Unsafe) as shown in Table 3.

**Step-6: Combine the evidence using modified Dempster-Shafer combination rule (Yager’s rule)**
Finally, the combined belief or the collective confidence about the data quality is calculated by using Equation (3). The final results of the collective belief, plausibility, and uncertainty about the data quality are shown in Table 3.

<table>
<thead>
<tr>
<th>Individual Mass Belief</th>
<th>ML Expert ( (m_1) )</th>
<th>Domain Expert ( (m_2) )</th>
<th>Collective Belief</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safe</td>
<td>12/33</td>
<td>16/33</td>
<td>0.26</td>
</tr>
<tr>
<td>Unsafe</td>
<td>12/33</td>
<td>16/33</td>
<td>0.38</td>
</tr>
<tr>
<td>(Safe, Unsafe)</td>
<td>9/33</td>
<td>7/33</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Belief\((\text{Safe})\)=0.26
Plausibility\((\text{Safe})\)=1-Belief\((\text{Unsafe})\)=1-0.38=0.62
Uncertainty=Plausibility\(-\)Belief\(=0.62-0.26=0.36\)

If the derived uncertainty is not acceptable then either more EDA needs to be done to collect more supportive evidences or the training data needs to be augmented to meet the criteria. To show the efficacy of our idea, we re-run the same method with a newer version of the pedestrian detection dataset, JAAD [8]. This dataset is more diverse as it covers the environmental diversity, diversity in poses, accessories, etc. This helps to collect more positive evidences. Therefore, the uncertainty window got reduced in the case of the newer dataset compared to the old CALTECH dataset (Figure 2). However, due to a lack of evidence about annotation protocol and other pre-processing steps, it shows a little uncertainty. The entire list of evidences is provided in the replication package [4].

**4 CONTRIBUTION**
We proposed a step-wise method that can help the component-level experts and the system-level experts to collaborate and assess the fitness of the training data before delving into a rigorous training process. Various sources of data uncertainty and the corresponding criteria to reduce them have been discussed. We show that it is possible to collect evidences from the training data to support/refute the quality claims. We used DS theory to embrace the uncertainty in the beliefs of the diverse experts about the collected evidences.

**Figures**

Figure 2: Comparison of collective confidence in data quality of CALTECH and JAAD training dataset

Finally, we derived the collective confidence in the quality of the data by applying the combination rule. This approach can contribute in **enhanced explainability** and **traceability**. Through the proposed method, the conflicts among the experts can come to the surface. This encourages more discussions and elevates confidence in data. Moreover, this approach also encourages focusing on evidence-based requirements engineering from a data perspective. Requirements engineers can collaborate with ML experts, systems engineers, and domain experts to elicit **verifiable data requirements**. The proposed approach will be useful for safety assurance where data quality can be assured by the claims, evidences, and the corresponding arguments.

**5 FUTURE ROADMAP**
We envision ML-based software-intensive systems engineering process as a systematic transition from problem space to solution space through a data space. Our focus is on first ensuring that the data space has been explored well before exploring the ML-based solution space. We believe that requirements engineering process can mitigate many challenges related to data dependency of ML models. We are currently working on a multi-layered framework to facilitate data requirements engineering that can support evidence-driven data quality verification. We are also designing a semi-formal data requirements specification language to express the verifiable data requirements, relevant evidences, and their assessment criteria. In the future, we have a plan to automate the confidence assessment process and perform an empirical study involving diverse experts.

**6 DATA AVAILABILITY**
Detailed guidelines, results calculation, rationales, the scope for improvement, etc. can be found in the replication package [4].

**ACKNOWLEDGMENTS**
This work was supported by the BK21 FOUR program of the National Research Foundation (NRF) of Korea funded by the Ministry of Education (NRF5199991014091) and the Basic Science Research Program through the NRF funded by the Ministry of Science and ICT (NRF-2020R1F1A1075605).
REFERENCES


