Quantifying Uncertainty of V-Information for Data Valuation

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Srinivasa Josyula
Spring, 2023

Technical Project Team Members
Hanjie Chen

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Srinivasa Josyula
Yangfeng Ji, Department of Computer Science

Quantifying Uncertainty of $\mathcal{V}$-Information for Data Valuation

Srinivasa Josyula, Hanjie Chen, Yangfeng Ji
Department of Computer Science
University of Virginia
Charlottesville, VA, USA
ssj2eq@virginia.edu, hc9mx@virginia.edu, yangfeng@virginia.edu

Abstract

Recent works have quantified the amount of information in a dataset over a model family. Predictive $\mathcal{V}$-information extends the foundations of Shannon Information theory by using mutual information and other concepts of informativeness. Specifically, $\mathcal{V}$-information takes into account the modeling power and computational constraints of the user (Xu et al., 2020). Further work has been done to calculate $\mathcal{V}$-usable information and pointwise $\mathcal{V}$-information (PVI). (Ethayarajh et. al., 2022) These metrics allow for comparison between different datasets or slices of datasets w.r.t. a model family. In this paper, we discuss the robustness issues faced by this metric. These issues are centered around $\mathcal{V}$-information being sensitive to small perturbations in input data. This produces inconsistent results and therefore limits the real-world usage of this metric. We also propose a method to estimate the uncertainty of $\mathcal{V}$-information using dropout layers. (Gal and Ghahramani, 2016). This will allow us to determine the Bayesian approximation and construct a confidence interval for PVI.

1 Introduction

Recent metrics have made it easier to compare the usefulness of various datasets across a model family. $\mathcal{V}$-information and pointwise $\mathcal{V}$-information quantify the usable information in a dataset. This is especially useful for natural language processing tasks because it helps to see the contribution of individual points in training. It can also be used to understand the "difficulty" of a dataset in training a model (Ethayarajh et al., 2021). However, these metrics are only useful in real-world scenarios if they are robust to small perturbations in input data. This allows for effective comparison between datasets or slices of the same dataset.

In this paper, we show that $\mathcal{V}$-information and pointwise $\mathcal{V}$-information suffer from robustness issues that decreases their usefulness. Specifically, these issues affect the downstream task performance indicated by metrics such as Kendall-Tau Correlation and top-k intersection. In this paper, we applied small perturbations on the input data. Then, we applied to $\mathcal{V}$-information algorithm on these models to highlight their robustness issues using the above metrics. These perturbations should have no effect on the information encoded in the input data. Still, $\mathcal{V}$-information indicates that there is a significant enough change to impact the downstream task metrics. These issues make it difficult to compare the $\mathcal{V}$-information across various datasets on the same model.

In this paper, we propose a method to estimate the uncertainty on $\mathcal{V}$-information and pointwise $\mathcal{V}$-information. This is achieved using dropout training on the original model to get a Bayesian approximation for $\mathcal{V}$-information. Dropout training has been proven to approximate to Bayesian inference in deep Gaussian processes (Gal and Ghahramani, 2016).

2 Related Works

Usable information in a dataset. $\mathcal{V}$-information allows for the estimation of usable information in different datasets over a model class. It also provides advantages over Shannon information (Shannon, 1948) that are relevant to machine learning applications. According to the Data Processing Inequality, post-processing of data cannot increase its mutual information with a random variable. However, in many examples such as encryption schemes and image classification this is not useful. Plain text will be much more useful information at predicting the label than encrypted text. $\mathcal{V}$-information seeks to provide a more intuitive metric for understanding the information in dataset. $\mathcal{V}$-information can be also provably estimated which makes it a more practical choice for usage (Xu et al., 2020).

Estimating the difficulty of data points.

Pertubation	Kendall-Tau	Top 10 Intersect	Top 25 Intersect	Bottom 10 Intersect	Bottom 25 Intersect
Seed 1	-3.96e-3	0.20	0.52	0.60	0.60
Seed 2	-6.49e-4	0.60	0.64	0.70	0.60
Repeat	-9.76e-3	0.50	0.44	0.30	0.48
Synonym 1	1.07e-2	0.40	0.52	0.70	0.52
Synonym 2	3.72e-3	0.40	0.52	0.60	0.48

Table 1: Summary of different metrics after applying various perturbations mentioned above.

A related metric known as pointwise $\mathcal{V}$-information(PVI) was introduced to measure the information from slices or specific points in the dataset. A high PVI score would indicate that the point provides a high amount of usable information that the model can use. In addition, it also indicates that the point has a high chance being correctly classified.

However, these works do not discuss the robustness or confidence of $\mathcal{V}$-information measurements. This makes it difficult to reliably use the metrics to decide upon the importance of each data point. Therefore, in some instances a high PVI may result in an erroneous label because the model is not confident with the PVI score.

Using Dropout as Bayesian Approximation Prior work has shown that for deep Gaussian processes, Dropout can be used as a Bayesian approximation (Gal and Ghahramani, 2016). In many cases, neural networks use dropout training to avoid the problem of overfitting. It has been shown that dropout can estimate model uncertainty without affecting computational complexity or test accuracy.

3 Method

Calculation of V-Information and PVI. The calculation of $\mathcal{V}$-information first requires the calculation of v-entropy on the given dataset. We take X and Y to be random variables and V to be a predictive family. V-entropy is defined as the smallest expected log-likelihood that can be achieved while predicting Y given the observation X using the model family V. $\mathcal{V}$-information is calculated by finding the difference in $\mathcal{V}$-entropy when the model family $\mathcal{V}$ is provided with observation X compared to a null set. $\mathcal{V}$-information allows for comparison between usable information in different datasets. Pointwise $\mathcal{V}$-information is calculated similar to $\mathcal{V}$-information. Instead of calculating the difference in v-entropy over the entire dataset XxY, it is done for every single point (x,y) in the dataset. Pointwise $\mathcal{V}$-information allows for comparison between points in one dataset or slices of a dataset.

	Original	Repeat	Synonym 1
Original	-	-	-
Repeat	91.33	-	-
Synonym 1	94.80	88.11	-
Synonym 2	94.59	87.87	98.43

Table 2: Percentage of overlap between PVI confidence intervals within 2 standard deviations

Calculating PVI confidence interval. A dropout layer is added after the BERT embeddings. This dropout layer is turned on during inference time, which allows it to compute a variable answer for each inference run. Using this, we can generate a sample of $\mathcal{V}$-entropy and pointwise $\mathcal{V}$-information values for each data point. Consequenty, the mean and standard deviation of the sample can be calculated. This will allow us to assess the spread of each PVI value in the dataset. It will also provide a provide a more robust measurement of the PVI of a data point.

4 Experiment

4.1 Experimental Setup

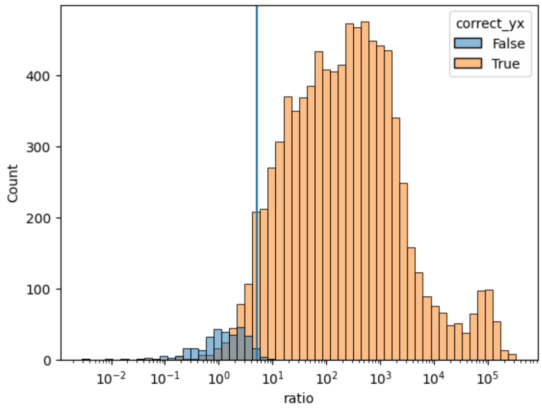
Dataset. We used the Stanford Natural Langugage Inference (SNLI) dataset. The dataset consists of sentence pairs that have been manually labelled for classification as entailment, contradiction, and neutral. **Model.** We used the BERT model to train on the SNLI dataset. BERT is a large language model that is primarily used for paired sentence classification. This makes it useful to observe the information gain across different data points.

4.2 Robustness Issues

Pertubation Methods. We first found the PVI for the data points in the original dataset trained on a BERT model. Then, we modified in different ways that intuitively do not change the usable information of the text. The text was modified by adding repetitive text and synonym replacement, both of which do not change the syntactic meaning of the sentence. The repetition was performed by repeating the input twice. The synonym replacement consisted of replacing one word in the sentence with a close synonym of that word. This was achieved by using the WordNet database in the Natural Language Toolkit (NLTK). Then, the PVI was found on the modified datasets to compare against the original dataset. We also found the PVI of the original dataset using two different starting seeds of the BERT model to study the affect of random initialization.

Evaluation Metrics. The results from the original and modified runs were compared using Kendall-Tau correlation, Top-k intersection and Bottom-k intersection. These rank based metrics would allow us to see how robust PVI was with respect to relative ranking after applying small perturbations. The results of these experiments are illustrated in Table 1. The results clearly illustrate that PVI results are not robust to small perturbations in the input text. The Kendall-Tau correlation is close to 0 for all the perturbations, indicating a weak ranking correlation. (A Kendall-Tau score of 1 represents identical ranking and -1 represents opposite ranking.) The bottom-k and top-k intersection are also close to 0.50 indicating a weak correlation near the top and bottom range of PVI values.

Figure 1: Mean/SD of examples based on classification



4.3 Dropout to estimate confidence

Adding Dropout to embedding space. In practice, dropout layers are commonly used in training model to overcome the problem of overfitting. This is achieved by randomly dropping nodes during training to decrease the reliance on a particular node. Dropout layers are also an integral part of the BERT model. During inference time, dropout layers are typically frozen in order to provide consistent results across runs. In this paper, we propose adding a dropout layer during inference time to estimate the confidence of PVI values. Specifically, we modified the existing BERT model to add a dropout layers after the word embeddings layer. This allowed us to collect a large sample of PVI values for each sentence pair in the dataset, dependent of the randomness of the dropout layer. Using this sample, the mean and standard deviation of each sentence pair's PVI were calculated. This process was conducted on the original dataset which had not gone through any perturbation.

PVI confidence interval. A sample of 50 PVI values consisting of the mean and standard deviation were used to create a confidence interval for the PVI values. We suggest that this confidence interval captures a more robust measurement of the PVI for sentence pairing. We compared the PVI confidence interval across various perturbations to assess the level of overlap that existed. Table 2 shows the percentage of overlap within 2 standard deviations for the various perturbations. The overlap between most datasets is close to 90%. The lowest level of overlap was found between the repetition and synonym replacement datasets, which was around 88%. The two synonym replacement datasets had a high level of overlap which was around 98%. These results show that the PVI confidence interval captures the variations from perturbations very effectively.

Assessing data point "Hardness". The mean and standard deviation PVI of each point can also be effectively used to determine the "hardness" of calculating the PVI of that point. For instance, a high PVI mean and small standard deviation would indicate that the point is useful for the model and also easy to calculate. On the other hand, a small PVI mean and a high standard deviation would indicate that the example is not useful for the model and very hard to calculate.

We propose that using the ratio of mean and standard deviation is a better way to understand the usefulness of a data point over simply using the PVI. We calculated the median PVI grouped over correctly and incorrectly classified examples. The median over correctly classified examples is 1.544 and incorrectly classified is -0.567. On the other hand, the ratio between mean and standard deviation yielded a median value of -1.492 for in-

Premise	Hypothesis	Label	Mean/SD
A group of people stand near and on a large black square on the ground with some yellow writing on it.	a group of people stand	1	-172.935
A happy woman quite stands in front of a business that displays a closed sign and looks very animated.	A woman stands by a business.	1	-94.317
An older gentleman looks at the camera while he is building a deck.	An older gentleman in overalls looks at the camera while he is building a stained red deck in front of a house.	0	-72.032
Two men are shirtless on the beach on a beautiful day.	Two men are shirtless.	1	-70.903
A man balances a unicycle using his face and both hands.	A man is using his hands.	1	-46.545

Table 3: Top 5 hardest examples in the SLNI test set according to the ratio of mean and standard deviation

correctly classified examples and 214.008 for correctly classified examples. Therefore, the ratio offers a better defined threshold between hard and easy examples for the evaluator to classify. Figure 1 also illustrates that high ratio values yield correct classifications, whereas low ratio values yield incorrect classifications. This provides an extremely clear and reliable way of differentiating "hard" and "easy" examples.

4.4 Qualitative Analysis

Below we highlight the top 5 points with the lowest mean and standard deviation ratios. These points have the lowest mean values and very low standard deviation values. These values have the highest chance of being misclassifed.

5 Conclusion

In this work, we identify robustness issues with the pointwise $\mathcal{V}$-information metric. We found the PVI of the SNLI dataset trained on a BERT model and compared it to the PVI of various modified datasets. These issues make it difficult to effectively use PVI to compare the usefulness of points across datasets. Then, we proposed a method of using dropout layers after the BERT embedding layer to generate a confidence interval for each PVI calculation. The confidence intervals achieved high overlap between the original and modified datasets. We also showed that the ratio of the mean and standard deviation of PVI values can be used to effectively determine the difficulty of predicting that example.

References

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2021. Information-theoretic measures of dataset difficulty. *CoRR*, abs/2110.08420.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *International Conference on Learning Representations*.

(Ethayarajh et al., 2021) (Xu et al., 2020) (Gal and Ghahramani, 2016)