ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096030

OPTIMAL COMPRESSION FOR MINIMIZING CLASSIFICATION ERROR PROBABILITY: AN INFORMATION-THEORETIC APPROACH

Jingchao Gao[1], *Ao Tang*[2], *Weiyu Xu*[1]

[1] University of Iowa, Iowa City, IA, 52242

[2] Cornell University, Ithaca, NY, 14853

ABSTRACT

We formulate the problem of performing optimal data compression under the constraints that compressed data can be used for accurate classification in machine learning. We show that this translates to a problem of minimizing the mutual information between data and its compressed version under the constraint on error probability of classification is small when using the compressed data for machine learning. We then provide analytical and computational methods to characterize the optimal trade-off between data compression and classification error probability. First, we provide an analytical characterization for the optimal compression strategy for data with binary labels. Second, for data with multiple labels, we formulate a set of convex optimization problems to characterize the optimal tradeoff, from which the optimal trade-off between the classification error and compression efficiency can be obtained by numerically solving the formulated optimization problems. We further show the improvement of our formulations over the information-bottleneck methods in classification performance.

Index Terms— classification, error probability, compression, mutual information, rate-distortion theory

1. INTRODUCTION

Machine learning plays an important role in science and engineering. Among machine learning tasks, classification is an important one which has many applications in communication and signal processing, for example, image recognition.

Machine learning needs sensor data to make inference or to perform classification [1, 2]. These sensor data are first collected, and then stored in storage or transmitted through communication channels to classifiers. However, the capacities of storage or communication channel are often limited. Thus, there is often a need to compress sensing data for more efficient storage or transmission. [3, 4, 5]. A fundamental question is hence how much compression one can achieve for sensing data such that machine learning tasks can still be executed with a certain given accuracy? In this paper, we propose a formulation of this problem, and try to answer this question for classification from an information-theoretic perspective.

In classification, we assume that labels (denoted by random variable Y) generate data (denoted by X) according to data generation distribution $\mathrm{P}(X|Y)$. Data X is fully known to the data compressor. The data compressor compresses X into compressed data $\tilde{X}$. The goal for the compressor is to compress X as much as possible for efficient communication or storage while allowing the classification task to be performed still with a specified fidelity: namely the label Y can still be sufficiently accurately recovered using only compressed data $\tilde{X}$. Towards this end, we propose to minimize the mutual information between X and $\tilde{X}$ while minimizing the error probability (or generalized costs associated with classification errors).

In classical rate-distortion theory for lossy data compression, data compression is performed so that the mutual information between data X and compressed data $\tilde{X}$ is minimized under the constraint on a distortion criterion between X and $\tilde{X}$ [6]. The distortion criterion in rate-distortion theory is often a direct distortion measure depending on the original data X and the compressed data $\tilde{X}$. In contrast, in this paper, for the classification task, we are considering the distortion between the original label and the recovered label ($\hat{Y}$) for classification, rather than the direct distortion between X and $\tilde{X}$.

Our research problem is connected with the information bottleneck principle [7][8][9][10], which was proposed to study data compression under the constraint of preserving classification labels to a certain fidelity. The information bottleneck principle uses the mutual information between label (Y) and compressed data ($\tilde{X}$) as a simple proxy for the fidelity in preserving the label information. However, mutual information may not be an accurate indicator of the distortion between the recovered label $\hat{Y}$ and the original label Y in the classification task. This is especially true if the distortion in classification is asymmetric: the distortion for mis-classifying an object with label "a" to label "b" is weighted higher than mis-classifying an object with label "b" to label "a". In addition, [11, 12] looked at rate-limited communication of training data in machine learning and derived performance limits of constructed predictors based on such rate-limited communication.

This research is supported by NSF awards 2000425, 2133205 and 2133403.

In this paper, we directly consider more relevant metrics for characterizing classification performances in determining optimal compression of sensing data. In particular, we study the problem of minimizing the mutual information between data and compressed data under constraints on classification error probability (or or generalized costs associated with classification errors), which are widely used performance metric for evaluating a classifier. The rest of this paper is organized as follows. In Section 2, we formulate the problem of optimally compressing data under classification error probability constraints. In Section 3, we analytically characterize the optimal compression strategy for binary symmetric channel connecting label and sensing data. In Section 4, we propose a general optimization framework to calculate the optimal compression and resulting minimum classification error probability. In Section 5, we present numerical results showing the optimal trade-off between data compression and classification error probability.

2. MODEL FORMULATION

Suppose that we have m labels in the label set $\mathcal{Y}$, which is $\{y_1, y_2, \ldots, y_m\}$. We let the prior probability for the labels be $\mathrm{P}(y_i)$, $i = 1, 2, \ldots, m$. Then the label (Y) will generate data, and we denote the set of possible data as $\mathcal{X}$. We assume that $\mathcal{X}$ has n elements, and its elements are $x_1, x_2, \ldots, x_n$. We denote the transition probability between each label and any possible data as $\mathrm{P}(x_j|y_i)$, where $i = 1, 2, \ldots, m$; and $j = 1, 2, \ldots, n$. For efficient storage and communication, we want to compress data X to compressed data $\tilde{X}$, which are sampled from set $\tilde{\mathcal{X}}$ of cardinality l. To be exact, $\tilde{\mathcal{X}}$ includes $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_l$ as its elements. Furthermore, we define that the transition probability between each data X and its compressed data $\tilde{X}$ as $\mathrm{P}(\tilde{x}_k|x_j)$, where $j = 1, 2, \ldots, n$, and $k = 1, 2, \ldots, l$. We assume that the decoder or machine learning algorithms use the maximum a posteriori (MAP) decoder (or the minimum-cost decoder when general costs associated with decoding errors are considered) to decode compressed data $\tilde{x_k}$ to label $\hat{y}_{\tilde{x}_k}$, where $1 \leq k \leq l$. The job of the compressor is to design the transition probabilities $\mathrm{P}(\tilde{x}_k|x_j)$'s such that the mutual information $I(X, \tilde{X})$ is minimized for most efficient compression, while keeping the decoding error probability (the probability that the decoded label is not equal to the original label) smaller than a certain threshold.

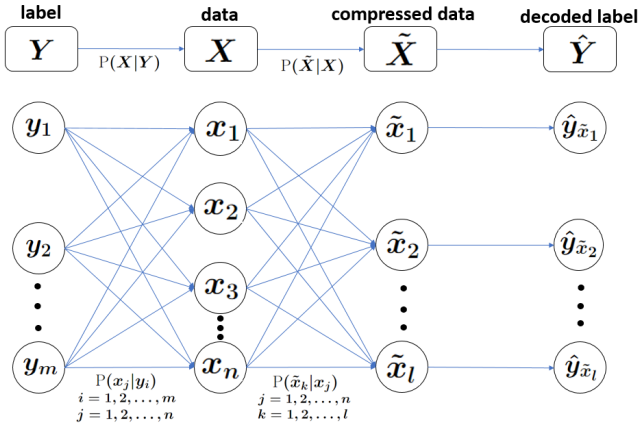

Fig. 1. Transition probabilities between labels, data, and compressed data

3. OPTIMAL COMPRESSION FOR BINARY SYMMETRIC CHANNEL: ANALYTICAL RESULTS

While it is difficult to obtain analytical solutions to the proposed problem in general, we are able to analytically derive analytical optimal compression strategies for binary labels and data. We consider the case of binary labels and we assume that there are also two elements in the alphabet for data and the alphabet for compressed data. We assume that $\mathrm{P}(Y = 0) = \frac{1}{2}$, and $\mathrm{P}(Y = 1) = \frac{1}{2}$. We try to minimize the mutual information between X and $\tilde{X}$ (subject to MAP decoding error threshold constraints) over the following transition probabilities p_1, p_2 and p_3: $\mathrm{P}(X = 1|Y = 0) = \mathrm{P}(X = 0|Y = 1) = p_1$, $\mathrm{P}(\tilde{X} = 1|X = 0) = p_2$, and $\mathrm{P}(\tilde{X} = 0|X = 1) = p_3$.

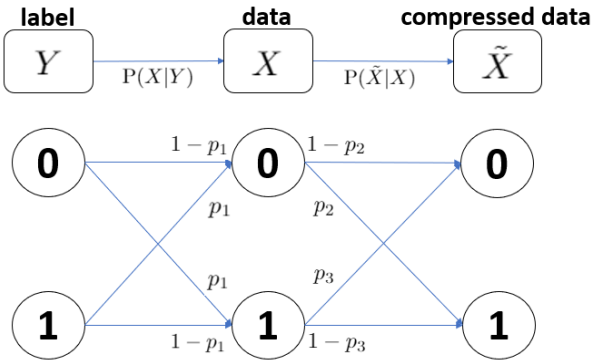

Fig. 2. Transition probabilities for binary data.

Theorem 1. *For binary data, where each label has equal probability, and with symmetric crossover transition probabilities that are less than $\frac{1}{2}$ between label and data, the optimal trade-off in terms of classification error probability and data compression is achieved by having symmetric transition probabilities between data and compressed data (namely $p_2 = p_3 \leq \frac{1}{2}$). Then the smallest achievable mutual information between X and $\tilde{X}$ is $I = 1 - p_2 \log \frac{1}{p_2} - (1 - p_2) \log \frac{1}{1-p_2}$ corresponding to an error probability no bigger than $Pe = p_1 + p_2 - 2p_1p_2$.*

Proof. In this proof, we show that if $p_2 \neq p_3$, we can always make the crossover probability symmetric and equal to the average of p_2 and p_3, without increasing $I(X, \tilde{X})$ and without increasing the MAP decoding error probability.

$$\mathrm{P}(Y=1|\tilde{X}=0) = \frac{\mathrm{P}(Y=1,\tilde{X}=0)}{\mathrm{P}(\tilde{X}=0)} = \frac{p_1+p_3-p_1p_2-p_1p_3}{1-p_2+p_3},$$

Similarly,

$$\mathrm{P}(Y=0|\tilde{X}=1) = \frac{p_1+p_2-p_1p_3-p_1p_2}{1+p_2-p_3}.$$

Then, $\mathrm{P}(Y=0|\tilde{X}=0) = 1-\mathrm{P}(Y=1|\tilde{X}=0)$, $\mathrm{P}(Y=1|\tilde{X}=1) = \mathrm{P}(Y=0|\tilde{X}=1)$. Since $p_1 < \frac{1}{2}$ and p_1 is fixed, if $p_2 < 1-p_3$, we have $\mathrm{P}(Y=1|\tilde{X}=1) > \mathrm{P}(Y=0|\tilde{X}=1)$ and $\mathrm{P}(Y=0|\tilde{X}=0) > \mathrm{P}(Y=1|\tilde{X}=0)$. This gives us

$$Pe = \frac{1}{2}\mathrm{P}(\tilde{X}=0|Y=1) + \frac{1}{2}\mathrm{P}(\tilde{X}=1|Y=0) = p_1(1-p_2-p_3) + \frac{p_2+p_3}{2}.$$

Otherwise, if $p_2 > 1-p_3$, similarly, it follows:

$$Pe = 1 - p_1(1-p_2-p_3) - \frac{p_2+p_3}{2}.$$

Next, we do the convex combination of p_2 and p_3, such that

$$\mathrm{P}(\tilde{X}=1|X=0) = \mathrm{P}(\tilde{X}=0|X=1) = \frac{p_2+p_3}{2},$$

Since $p_1 < \frac{1}{2}$ and p_1 is fixed, by the same process as above, if $p_2 < 1-p_3$, we have $Pe = \frac{1}{2}\mathrm{P}(\tilde{X}=0|Y=1) + \frac{1}{2}\mathrm{P}(\tilde{X}=1|Y=0) = p_1(1-p_2-p_3) + \frac{p_2+p_3}{2}$. Otherwise, if $p_2 > 1-p_3$, similarly, $Pe = \frac{1}{2}\mathrm{P}(\tilde{X}=0|Y=0) + \frac{1}{2}\mathrm{P}(\tilde{X}=1|Y=1) = 1 - p_1(1-p_2-p_3) - \frac{p_2+p_3}{2}$.

In conclusion, we notice that Pe remains the same before and after doing convex combination. Since the mutual information is convex function of the transition probability between X and $\tilde{X}$ for fixed $P(X)$ [6], mutual information is not increased after doing convex combination while Pe does not increase. This implies that the optimal transition probability should be symmetric.

Finally, with this conclusion, we can focus on a symmetric crossover probability p_2, namely, $\mathrm{P}(\tilde{X}=1|X=0) = \mathrm{P}(\tilde{X}=0|X=1) = p_2$. Then,

$$\mathrm{P}(Y=1|\tilde{X}=0) = \mathrm{P}(Y=0|\tilde{X}=1) = p_1+p_2-2p_1p_2,$$

Now suppose that $p_1 < \frac{1}{2}$, and we notice that if we also have $p_2 < \frac{1}{2}$, then, $\mathrm{P}(Y=0|\tilde{X}=0) > \mathrm{P}(Y=1|\tilde{X}=0)$ and $\mathrm{P}(Y=1|\tilde{X}=1) > \mathrm{P}(Y=0|\tilde{X}=1)$. This suggests that $Pe = p_1+p_2-2p_1p_2$ and the mutual information is given by $1-p_2\log\frac{1}{p_2} - (1-p_2)\log\frac{1}{1-p_2}$. □

Remarks: Our proof is different from showing that symmetric transition probabilities achieve optimal rate-distortion tradeoff involving $I(X,\tilde{X})$ and binary distortion between X and $\tilde{X}$. Here we consider the decoding error probability for label Y, making our proof arguably more involved.

4. OPTIMIZATION FORMULATION FOR COMPUTING OPTIMAL COMPRESSION

Suppose that we have m labels in the label set $\mathcal{Y}$, and we denote them by $y_1, y_2, \ldots, y_m$. We denote the prior probability for each label as $\mathrm{P}(y_i)$, $i = 1, 2, \ldots, m$. Then these labels generate data sampled from set $\mathcal{X}$ of cardinality n. Specifically, the elements in $\mathcal{X}$ are $x_1, x_2, \ldots, x_n$. We denote the transition probability between each label and possible element for data as $\mathrm{P}(x_j|y_i)$, where $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$. We want to map (compress) the data to l possible letters in the compressed data set $\tilde{\mathcal{X}}$ of cardinality l, which includes $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_l$ as its elements. Furthermore, we define the transition probability between x_j and compressed data $\tilde{x}_k$ as $\mathrm{P}(\tilde{x}_k|x_j)$, where $j = 1, 2, \ldots, n$; $k = 1, 2, \ldots, l$.

Our goal is to minimize the mutual information between X and $\tilde{X}$ by optimizing over the transition probabilities $\mathrm{P}(\tilde{x}_k|x_j)$, subject to the constraint that the classification error probability is smaller than a certain threshold ϵ. However, this optimization problem is a non-convex optimization problem. We propose to obtain global optimal solution by dividing this optimization problem into multiple convex optimization problems, based on different MAP decoding rules.

We assume that for a given letter $\tilde{x}_k$, the MAP rule decodes it to label $\hat{y}_{\tilde{x}_k}$, which is from the set $\mathcal{Y}$. We notice that there are m^l possible MAP maps from $\tilde{\mathcal{X}}$ to $\mathcal{Y}$. For each MAP decoding rule, we are trying to minimize the mutual information between X and $\tilde{X}$. So for a particular MAP decoding rule, minimizing $I(X;\tilde{X})$ is equivalent to the following convex programming:

$$\min_{\mathrm{P}(\tilde{x}_k|x_j)} I(X;\tilde{X})$$

$$\text{subject to} \quad Pe = \sum_{k=1}^{l}\sum_{y_i\neq\hat{y}_{\tilde{x}_k}}\sum_{j=1}^{n}\mathrm{P}(y_i)\mathrm{P}(x_j|y_i)\mathrm{P}(\tilde{x}_k|x_j) \le \epsilon,$$

$$\mathrm{P}(\tilde{x}_k|x_j) \ge 0, \quad \forall x_j \in \mathcal{X}, \tilde{x}_k \in \tilde{\mathcal{X}}$$

$$\sum_{k=1}^{l}\mathrm{P}(\tilde{x}_k|x_j) = 1, \quad \forall x_j \in \mathcal{X}$$

$$\sum_{j=1}^{n}\mathrm{P}(y_i)\mathrm{P}(x_j|y_i)\mathrm{P}(\tilde{x}_k|x_j) \le \sum_{j=1}^{n}\mathrm{P}(\hat{y}_{\tilde{x}_k})\mathrm{P}(x_j|\hat{y}_{\tilde{x}_k})\mathrm{P}(\tilde{x}_k|x_j) \quad \forall k, \forall y_i \neq \hat{y}_{\tilde{x}_k}$$

where ϵ is the given error probability tolerance threshold. *We have proved that the minimum objective value among these m^l such convex optimization problems give the globally optimal compression under a constraint on error probability.* This formulation also extends to asymmetrical cost for decoding error.

5. NUMERICAL RESULTS

In this section, we present numerical results for characterizing the optimal tradeoff between compression and classification accuracy.

In Fig.3, we calculate the curve of the allowed mutual information between data (X) and compressed data, against the classification error probability for the binary data under the parameters $p_1 = 0.3$. The plot is generated by using the result in Theorem 1. From the plotted curve, we can see that, when the mutual information between data X and compressed data $\tilde{X}$ is allowed to be large, the classification error probability can be reduced, but at the expense of compression efficiency.

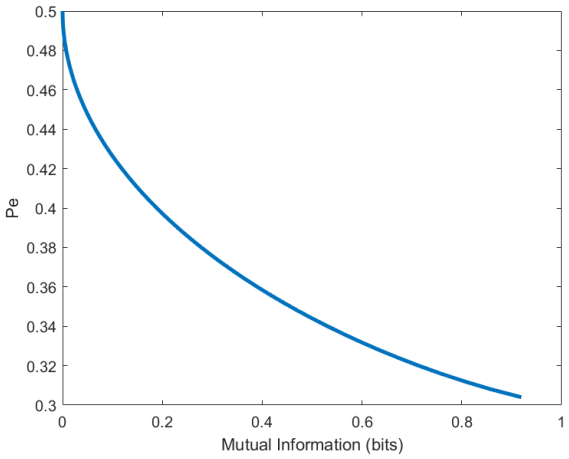

Fig. 3. Mutual information between data and compressed data against classification error probability for $p_1 = 0.3$.

We further consider the case where the costs of decoding to incorrect labels are asymmetrical. In Fig.4, we plot the optimal classification cost and data compression trade-off, for a classification task with 3 labels, 4 data letters and 3 compressed data letters, with transition probabilities in the first channel and costs of incorrectly decoding from each label to decoded label shown as follows (the prior probability for each label is $1/3$). Note that when the $c = 1$, the cost is equivalent to the decoding error probability.

$P(x_j\|y_i)$	y_1	y_2	y_3
x_1	0.995	0.001	0.002
x_2	0.001	0.996	0.002
x_3	0.002	0.001	0.994
x_4	0.002	0.002	0.002

cost	$\hat{y} = y_1$	$\hat{y} = y_2$	$\hat{y} = y_3$
y_1	0	c	c
y_2	1	0	1
y_3	1	1	0

Next, we consider the case with 3 labels, 3 data letters and 2 compressed data letters where costs of incorrectly decoding from each label to decoded label and transition probabilities between label and data are shown in the following tables.(the prior probability for each label is $1/4$, $1/4$ and $1/2$)

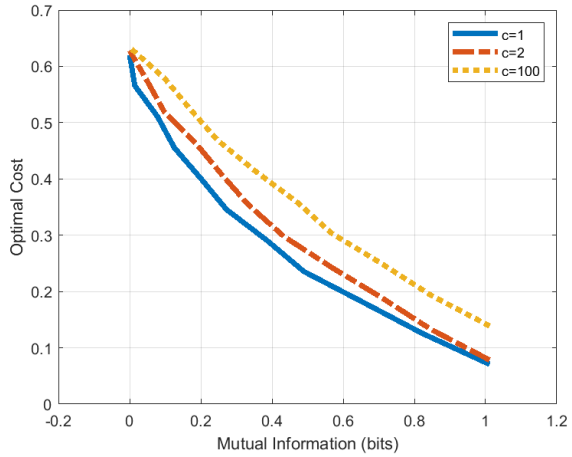

Fig. 4. Optimal cost against mutual information between data and compressed data for different c.

$P(x_j\|y_i)$	y_1	y_2	y_3
x_1	0.9	0.1	0.05
x_2	0.1	0.9	0.05
x_3	0	0	0.9

cost	$\hat{y} = y_1$	$\hat{y} = y_2$	$\hat{y} = y_3$
y_1	0	1	1
y_2	1	0	1
y_3	0.0001	0.0001	0

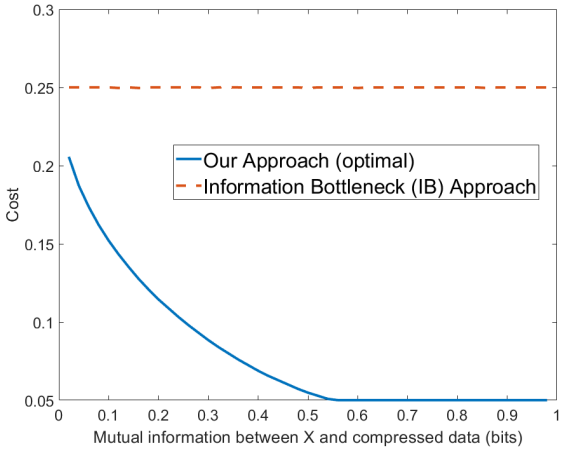

Fig. 5. Optimal cost and compression tradeoff for our approach, and comparison with the performance of information bottleneck approach.

In Fig.5, compared with Information Bottleneck Principle (IBP, which directly maximizes mutual information between label and compressed data), we get a curve of decoding cost against the mutual information between data (X) and compressed data ($\tilde{X}$). As we can see, our newly proposed approach can significantly outperform the IBP approach in achieving minimum decoding cost and highest compression efficiency. The reason is that the information bottleneck approach was not optimized for minimizing the cost.

6. REFERENCES

[1] Kevin P. Murphy, *Machine learning : a probabilistic perspective*, MIT Press, Cambridge, Mass. [u.a.], 2013.

[2] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[3] Robert Calderbank, Sina Jafarpour, and Robert Schapire, "Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain," Tech. Rep., 2009.

[4] E. Zisselman, A. Adler, and M. Elad, "Chapter 1 - compressed learning for image classification: A deep neural network approach," in *Processing, Analyzing and Learning of Images, Shapes, and Forms: Part 1*, Ron Kimmel and Xue-Cheng Tai, Eds., vol. 19 of *Handbook of Numerical Analysis*, pp. 3–17. Elsevier, 2018.

[5] Jiangnan Cheng, Marco Pavone, Sachin Katti, Sandeep Chinchali, and Ao Tang, "Data sharing and compression for cooperative networked control," *accepted to NeurIPS 2021*.

[6] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*, Wiley-Interscience, USA, 2006.

[7] Naftali Tishby, Fernando Pereira, and William Bialek, "The information bottleneck method," *Proceedings of the 37th Allerton Conference on Communication, Control and Computation*, vol. 49, 07 2001.

[8] Naftali Tishby and Noga Zaslavsky, "Deep learning and the information bottleneck principle.," *CoRR*, vol. abs/1503.02406, 2015.

[9] Anton Bardera, Jaume Rigau, Imma Boada, Miquel Feixas, and Mateu Sbert, "Image segmentation using information bottleneck method," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1601–1612, 2009.

[10] Bernhard Geiger and Gernot Kubin, "Information bottleneck: Theory and applications in deep learning," *Entropy*, vol. 22, pp. 1408, 12 2020.

[11] Maxim Raginsky, "Achievability results for statistical learning under communication constraints," *CoRR*, vol. abs/0901.1905, 2009.

[12] Maxim Raginsky, "Empirical processes, typical sequences, and coordinated actions in standard borel spaces," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1288–1301, 2013.