entropy

MDPI

Article

A Note on Representational Understanding

Antal Jakovác [1] and András Telcs [1,2,3,*]

[1] Department of Computational Sciences, Wigner Research Centre for Physics, H-1121 Budapest, Hungary
[2] Department of Computer Science and Information Theory, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, H-1111 Budapest, Hungary
[3] Department of Quantitative Methods, Faculty of Business and Economics, University of Pannonia, H-8200 Veszprém, Hungary
* Correspondence: telcs.andras@wigner.hu

Abstract: In this paper, we explore a new approach in which understanding is interpreted as a set representation. We prove that understanding/representation, finding the appropriate coordination of data, is equivalent to finding the minimum of the representational entropy. For the control of the search for the correct representation, we propose a loss function as a combination of the representational entropy, type one and type two errors. Computational complexity estimates are presented for the process of understanding and using the representation found.

Keywords: data representation; coordinate systems; entropy; first and second error

check for updates

Citation: Jakovác, A.; Telcs, A. A Note on Representational Understanding. *Entropy* **2022**, *24*, 1313. https://doi.org/10.3390/e24091313

Academic Editor: Fernando Morgado-Dias

Received: 25 July 2022
Accepted: 14 September 2022
Published: 17 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Intelligence in general is the ability to respond quickly and adequately to challenges of the external world. Animals need to be able to recognise enemies and predators quickly. This can be considered as a one-sided classification, one class versus all the others. In [1] the authors argue that responding extremely quickly requires a different solution than classical classification or learning. There, they propose a new paradigm, fundamentally different from classification, to cover the cognitive process of understanding. The present paper shows the theoretical feasibility of the comprehension process and proposes practical development of the process of understanding.

The authors in [1] argue, following the strategy of representation learning [2,3], that understanding a topic is equivalent to finding the right representation of the data. According to this approach, understanding does not involve data compression, but merely the rearrangement of known facts/characteristics that fit best to the observed phenomena.

There is a subtle philosophical difference between understanding and learning, although they may seem technically similar and are part of our cognitive work. Classic AI/ML tasks such as classification, clustering, encoding, etc. go hand in hand with understanding, "model building". Before we are able build a proper model, we try to classify objects based on their characteristics, features, or try to cluster them. If we find a strikingly good arrangement, it can lead to a model idea. After some trials we find a model that fits the data well (after adjusting a few parameters) and meets our expectations of the model. We stress here that model building always involves some preliminary assumptions (which can be good or bad). Once you have a suitable model in hand, the task of classification or clustering is easy. As an example, let us consider a classification problem. If we have to classify a new item, we have to compare its features with those of the possible classes. If we have a model, a few relations between some features should be checked to identify the right class. This is the bird tweet phenomenon. Without a model, it is difficult (expensive in terms of algorithmic steps) to find the solution, while the model plays the role of the bird and tells us the correct answer. Imagine that we have sample points from several polynomials of order one, two and three. If we are given a new set of

sample points from an unknown polynomial, even if we know the three classes, we need to compare several class elements with the new one until we find a class element that is convincingly close, similar, to the new one. Contrary, if we know the laws of the three classes, a quick model fitting yields the answer. We will see that this quick solution comes at a price. While supervised or unsupervised learning algorithms have relatively low complexity, finding a model that well characterizes the classes, the relationship between the features, is a very difficult task with much higher complexity. The difference is typically exponential, or even hyper-exponential, depending on the task.

Let us take a closer look at the learning and model building processes. We talk about training when we have a feature set of data and we want to fine-tune a combination of these features (pick out a few key ones) to optimise some behaviour (usually classification or clustering, in some cases forecasting or filling in missing data). The task of understanding, on the other hand, is to find the best data model that reveals the relationships (functions, laws) between features and best describes a problem that is set prior the whole investigation. We call this the context of model building. Training always requires an explicit loss function that decides whether the trained system behaves correctly or not. This may be rather implicit, as in the case of auto-encoders or reinforcement learning. In the process of understanding, we consider only the input data set and try to figure out the feature functions that separate our data set from the rest of the world.

More formally, in the task of understanding the representative elements of the system (sub-universe) we want to understand, are presented one by one to the model-building apparatus. It has separating and descriptive coordinates (paper [1] called them relevant and irrelevant coordinates, respectively, based on observed features). These are either kept as they are or modified to better fit the new observed element. The separation coordinates are expected to be constant 1 for the elements of the investigated sub-universe. Meanwhile, the descriptive coordinates distinguish the elements of our subset piece-by-piece. After observing a large enough number of samples, we expect to have a good understanding, and later our understanding can also be the basis for a quick classification.

Again, we emphasize that we think that there is no understanding/model building without context, prior assumptions, questions and prior elimination of almost all the aspects of the universe, except some that are the particular subject of the investigation, the problem to be solved. In problem setting, we must specify the set of objects Ω and the function space that maps the "measurements" of the elements of the space to the new "coordinates". In our abstract model, the sub-space of binary functions which has $log_2(\Omega)$ variables should be specified. Overall, the context is given by Ω and the chosen function space.

One simplified example could be the predefined scale of the study. We may search for a model of sand in a desert. If we have bird's eye view photographs of the desert (10^6 mm scale), we can try to shift and overlap the dunes in the images and find that there is an almost periodic pattern that describes well the sand surface. The model can be formalised using trigonometric series. Time lapsed video recordings of the same scale can reveal wave-like behaviour and explain the periodicity (or falsify the static view). At a 10^1–10^2 mm scale, surface tension, friction, avalanche effects may contribute to the description of the observation. The formalism should be based on polynomials of a few order, chaotic systems, or stochastic dynamical systems. On a scale of 10^{-3}–10^{-0} mm, fragmentation, particle collision processes can be captured and the formalism can be based on branching processes or dynamical systems with external potentials (forces) from statistical physics.

In general, it is not easy to decide whether a coordination is appropriate in the above sense or not. Are we satisfied with the coordination (representation) or do we need further investigation, sample elements? To facilitate this, it was suggested in [4] to associate an entropy with each representation, which is minimal if the coordinates are chosen properly. In [4], some properties of the proposed entropy function were investigated and it was demonstrated in simple examples that it indeed performs the desired task.

In the present paper, we take a closer look at the practical understanding process. First, we recall the formal model of understanding. We then show that representation

entropy is minimal if and only if the representation is canonical, i.e., a representation that separates the subset and describes the elements in terms of "independent" coordinates. We propose a loss function for representations and provide theoretical and practical calculation of the type one and type two errors of representations. Finally, an estimate of the cost/complexity of the understanding procedure is given.

2. The Representational Entropy

We assume that Ω is a finite set, $\{\Omega, \mathcal{F}, P\}$ is a probability space, and P is uniform over Ω :$P(\omega) = \frac{1}{|\Omega|}$ for $\omega \in \Omega$.

For simplicity, we use sets with cardinality 2^k with $k > 0$ integers. $|\Omega| = 2^N$. Let $\Omega_1 \subset \Omega$ with $|\Omega_1| = 2^{N_1}$.

Definition 1. *A coordination of Ω is a bijection $x = \{x_i\}_{i=1}^N : \Omega \to \{0,1\}^N$. The coordinates are the binary functions: $x_i : \Omega \to \{0,1\}$ for $i = 1, \ldots, N$.*

For continuous variables see the discussion in [1] .

Definition 2. *A representation of $\Omega_1 \subset \Omega$ is a coordination in which the set Ω_1 and its complement $\Omega_2 = \Omega \backslash \Omega_1$ are distinguished by $N - I_1$ coordinates (binary functions), called separation coordinates (SCs), if all of them are 1 on Ω_1, and at least one of them is zero on Ω_2. The other I_1 coordinates called Descriptive Coordinates (DCs). Given that permutation of coordinates is irrelevant we may assume that the separation coordinates have the lower index:*

$$\begin{array}{lll} \forall i \in \{1, \ldots, N - I_1\} & x_i(\omega) = 1 & \text{if} \quad \omega \in \Omega_1, \\ \exists i \in \{1, \ldots, N - I_1\} & x_i(\omega) = 0 & \text{if} \quad \omega \notin \Omega_1. \end{array}$$

Definition 3. *A representation of Ω_1 is optimal (canonical) if all the DCs as random variables are independent over Ω_1 (with respect of the uniform distribution over Ω).*

Let us denote the descriptive coordinates by $y = \{y_i\}_{i=1}^{I_1}$, and the function they provide by $y : \Omega_1 \to \{0,1\}^{I_1}$ as well.

It is shown in [1] that the optimal (canonical) coordination always exists and coincides with the binary labeling of the elements of Ω_1 by descriptive coordinates, that is, by labeling them with the N_1-length binary vectors. One should observe immediately that $y : \Omega_1 \to \{0,1\}^{N_1}$ is a bijection for the canonical representation.

Definition 4. *The representation entropy is defined as follows: let $p_i(\sigma) = P_{\Omega_1}(x_i(\omega_i) = \sigma)$, $\sigma \in \{0,1\}$:*

$$S_{rep}(x, \Omega_1) = -\sum_{i=1}^{N} \sum_{\sigma=0}^{1} p_i(\sigma) \log_2 p_i(\sigma).$$

For convenience, let $H(x) = S_{Shannon}(x)$ denote the classical Shannon entropy:

$$H(\xi) = \sum_k -p_k \log(p_k),$$

where $p_k = P(\xi = x_k)$. Using that notation we have the following observation.

Lemma 1. *The representation has minimal representational entropy if and only if the coordinates are independent (or in other words, the representation is complete).*

Proof. At first, let us note that the entropy $H(y|\Omega_1)$ of the DCs is constant, equal to N_1, which justifies the search for a minimum. The statement of the Lemma is just reformulation of the well-known fact that

$$H(z_1, z_2, \ldots, z_n) \leq \sum_{i=1}^{n} H(z_i) \tag{1}$$

and equality holds if and only if all the z_i-s are independent. We give a short proof of this statement since it is in several textbooks (e.g., [5]) but not all the proofs are clear or complete.

If the family of random variables $z = \{z_i\}_{i=1}^n$ is independent then we have the equality in (1) by the chain rule. Let us assume that the family of random variables z is not independent. In that case, there are I_1 and $I_2 \subset \{1, 2, \ldots, n\}$ such that $I_1 \cap I_2 = \emptyset$, $I_1 \cup I_2 \subset \{1, 2, \ldots, n\}$, $I_1, I_2 \neq \emptyset$ and $x = \{z_i\}_{i \in I_1}, y = \{z_j\}_{j \in I_2}$ such that the variables x and y are not independent and

$$H(z) = H(x, y) \lneqq H(x) + H(y).$$

For convenience let us reorder $z_i - s$ so that $I_1 = \{1, ..k\}, I_2 = \{k+1, \ldots, n\}$. Using that notation we have that

$$\begin{aligned} H(z_1, z_2, \ldots, z_n) &= H(z) \lneqq H(x) + H(y) \leq \sum_{i=1}^{k} H(z_i) + \sum_{i=k+1}^{n} H(z_i) \\ H(z_1, z_2, \ldots, z_n) &\lneqq \sum_{i=1}^{n} H(z_i), \end{aligned}$$

which shows the reverse implication. □

3. Search

Let us assume that we have a given coordination x. Without loss of generality, we can assume that the first $N - I_1$ coordinates are going to represent Ω_1 with value 1-s. The type one error is:

$$\alpha = \alpha(x, \Omega_1) = P\Big(\cup\{x_i(\omega) = 0\}_{i=1}^{N-N_1} | \omega \in \Omega_1\Big)$$

and the type two error is:

$$\beta = \beta(x, \Omega_1) = P\Big(\{x_i(\omega) = 1\}_{i=1}^{N-I_1} | \omega \notin \Omega_1\Big).$$

The loss function is then defined as

$$L(x, \Omega_1) = S_{rep}(x, \Omega_1) + \lambda\alpha(x, \Omega_1) + \mu\beta(x, \Omega_1)$$

where $\lambda, \mu \geq 0$ regularizing meta parameters.

One should take into consideration when choosing the meta-parameters λ, μ that

$$\begin{aligned} \log_2|\Omega_1| &\leq S_{rep}(x, \Omega_1) \leq N, \\ 0 &\leq \alpha, \beta \leq 1. \end{aligned}$$

The theoretical values of the error probabilities are cumbersome, but interesting on their own. Their detailed calculation is given in Appendix A. The empirical estimates are easy. Similarly to the ML procedure, we split the sample into two. We use the first to create a coordination of the learning set, assuming that we have large enough sample then with the second part we calculate the estimate of the type one error. If we have access to non Ω_1 elements, then we can calculate the estimate of the probability of type two error as well.

Costs

The representation cost can also be a factor in the design of the coordination. There are different costs.

1. Creation of the coordinates that include the presentation of ω-s, building of the x_i coordinates and, at least once, the calculation of the ready coordinates.
2. Storage of the coordinate functions.
3. Calculation of the representation entropy, type one and type two errors.
4. Cost of decision if $\omega \in \Omega_1$ and full coordination of ω.

Let $\Omega_2 \subset \Omega_1$ be the set of presented items. The creation of a complete coordination of Ω_2 needs the presentation of $|\Omega_2| = 2^{N_2}$ elements. The canonical, complete representation contains $r_2 = N - N_2$ functions $x_i|_{\Omega_2} \equiv 1$ and further N_2 functions which maps to the lexicographical ordered binary vectors of length N_2. All the functions should be different. A very conservative estimate of the construction cost is the typical length $l(N)$ of a binary function of N inputs: (c.f. [6])

$$l(x_i) = l(N) \sim 2^N / \log_2 N$$

Of course if the function class is restricted, much lower cost is possible, but without the explicit knowledge of that we can not incorporate into our estimate. We need N such functions (where the choice of the value of N is based on expert guess, suggestion, external parameter of the process, can not be specified by the procedure). The cost of learning C_L is estimated by

$$\begin{aligned} C_L &= \text{number of items} \times \text{number of functions} \times \text{cost of a function} \\ &\sim c 2^{N_2} N 2^N / \log_2 N, \end{aligned}$$

where $c \geq 1$ given that each constructed function should be calculated at least once for each $\omega \in \Omega_2$.

The storage and recall of the representation needs at least

$$C_L \sim c' N 2^N / \log_2 N$$

steps.

The representation entropy for a complete representation is fixed ($1/M_2$), no calculation needed. Type one error can be estimated with the observed relative frequency which is based on the investigation of M_3 number of Ω_1 elements. That needs

$$\begin{aligned} C_{test} &= \text{number of items} \times \text{number of functions} \times \text{cost of a calculation of a function} \\ &\sim c' M_3 (N - N_2) 2^N / \log_2 N \end{aligned}$$

steps. If we have access to, or can create, $\Omega \backslash \Omega_1$ elements for test purposes, a similar estimate can be given for the cost of the estimate of type two error.

Finally, from the above consideration, it is clear that for an ω the decision if $\omega \in \Omega_1$ needs

$$C_{decision} \sim (N - N_2) 2^N / \log_2 N$$

steps and the full representation needs

$$C_{full} \sim N 2^N / \log_2 N$$

steps.

4. Summary and Discussion

The goal of understanding is to build a data model by which we can make a distinction between the "essential" an "unimportant" items.

Following [1], we identify the set of essential data through their characteristic features. This means that we calculate some quantities that have a given value for our distinguished set. We should build our data representation based on the characteristic and irrelevant features, which are called separation and descriptive coordinates in this paper. Once we know all the separation coordinates, we can identify our singled out set with 100% certainty.

In practice, however, it is very tedious to find this coordination and verify if it is optimal. It is much easier if we have a real valued function (loss function) that is minimal for the correct data representation. In this work, we have proven with mathematical rigour that the representation entropy defined in Definition 4 has exactly this property.

It is worth to emphasize that this representation quality measure is of statistical nature. While the generally used loss functions can be computed for each input, the use of representation entropy requires the knowledge of the bitwise probability distribution of the outputs, and thus it is available only after having seen enough examples from the given set. Once we know the bitwise probabilities, we can compute the representation entropy with the psuedo-code given by Algorithm 1.

Algorithm 1 Representation Entropy from bitwise probability distributions

procedure $S(p \in [0,1]^N)$ ▷ p are the bitwise probabilities of 1.
  $S \leftarrow 0$
  for $i = 0, 1, \ldots N-1$ **do**
    $S \leftarrow S - p_i \log_2 p_i - (1-p_i)\log_2(1-p_i)$
  end for
  return S
end procedure

In this paper, we worked in an abstract world with complete information, meaning that we know all possible inputs, and we can approach all elements of our specific set. In this case, we can find the best data representation using the psuedo-code of Algorithm 2, based on evolution algorithm.

Algorithm 2 Evolution Algorithm to find the proper representation

Require:
  $N \leftarrow$ number of bits of the input
  $\Omega_1 \leftarrow$ set to be represented
Ensure: initialization
  $R_0 \leftarrow X \to \{0,1\}^N$ original coordination (e.g., pixels)
  $S_{best} \leftarrow N$

for a given number of iteration **do**
  $P \leftarrow$ random $\{0,1\}^N \to \{0,1\}^N$ bijection (permutation)
  $x \leftarrow (0, 0, \ldots 0)$ N dimensional zero vector
  for ω **in** Ω_1 **do**
    $x \leftarrow x + P(R_0(\omega))$.
  end for
  $p \leftarrow x/|\Omega_1|$ ▷ These are the bitwise probabilities of 1.
  if $S(p) < S_{best}$ **then**
    $P_{best} \leftarrow P$
  end if
end for
return $P_{best} \circ R_0$

This algorithm works because the minimal representation entropy corresponds to the best coordination of the data.

In fact, in the abstract world we can also construct the best representation (c.f. [1]) as follows. The original R_0 representation associates an integer number to all elements,

interpreting its binary coordinates as digits of a binary number. Here the Ω_1 set usually shows up randomly. However, we can find a permutation P that sends the element of Ω_1 to the top of the list. Then $P \circ R_0$ associates the numbers $0, 1, \ldots 2^{N_1} - 1$ with the elements of Ω_1, i.e., only its first N_1 bits can be different from one. In this representation, therefore, the first N_1 bits are uniformly distributed, the higher bits are all one, thus the representation entropy is $N_1 = S_{Shannon}$. The type one and type two errors are all zero. Therefore $P_{best} = P$.

This solution, however, is not feasible in practice, since we do not have complete information. Then we shall use the given pseudo-code, with two modifications. The first is that we can not sample the whole $\{0,1\}^N \to \{0,1\}^N$ bijections, since the cardinality of this space is $2^N!$, which is usually too big to be parametrized. In practice, therefore, we shall be content with a relatively small subset of all the representations, which we should choose very carefully. For an appropriate choice of the function class, we can take into account a number of arguments, we shall examine the symmetry of the system, we may consider simple function classes (such as linear or low order polynomials), and we shall rely on the field knowledge, accumulated common wisdom in the given area.

The other reason is that usually we do not know the complete Ω and Ω_1 sets, we only see some samples of it. In fact we do not even know the cardinality of Ω_1 in general. To mitigate the problem of the limited number of observations in Ω_1, we may apply methods that potentially improve the convergence. In this paper we have suggested to incorporate into the cost function the type one and type two errors. That may be useful from this point of view.

Even with the best strategy, if we do not have complete information, the fixed values of the separation coordinates determine an $\Omega' \neq \Omega_1$ subset. Then a number of questions arise, such as what is the relation of Ω' and Ω_1, how to choose the set of available coordinations, how many separation and descriptive coordinates are worth to keep, and so on. These questions are the subject of our future studies.

Author Contributions: Conceptualization, A.J. and A.T.; Formal analysis, A.T.; Writing—original draft, A.T.; Writing—review & editing, A.J. All authors have read and agreed to the published version of the manuscript.

Funding: A.T. was partially supported by the Ministry of Innovation and Technology NRDI and the National Research, Development and Innovation Office, within the framework of the MILAB, Artificial Intelligence National Laboratory of Hungary and by the Hungarian Brain Research Program (2017-1.2.1-NKP-2017-00002). A.J. had a support from the Ministry of Innovation and Technology NRDI Office within the framework of the MILAB, Artificial Intelligence National Laboratory Program and the Hungarian Research Fund NKFIH (OTKA) under contract No. K123815.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors acknowledge useful discussions with T.S. Biró.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Let us calculate the error probabilities in coordination of the set Ω_1. Let $M = 2^N$ and $M_1 = 2^{N_1}$ First, let us investigate the process in which we develop the representation of Ω_1. Let us assume that we have shown $\Omega_2 \subset \Omega_1$ (with $M_2 = 2^{N_2} = |\Omega_2|$). Based on the given information, we can not represent Ω_1, but we can create as it is shown in [1] the complete representation of Ω_2. We use that as "approximation" for the representation of Ω_1. We analyse the relationship of the representation of Ω_1 and Ω_2 in particular from the point of view how the type one and type two errors behave.

Let R_i $(i = 1, 2)$ denote the index set of SCs for Ω_i-s and assume that $l = |R_1 \cap R_2|$. First, we calculate P_l the probability that is based on the Ω_2 data that we have a representation which has l common SCs with R_1.

$$P_l = \frac{1}{M_1 M_2} \binom{N}{l} \binom{N-l}{r_1-l} \binom{N-r_1}{r_2-l} 2^m,$$

where $m = N - r_1 - r_2 + l = N_2 - r_1 + l$. The type one error is for an $\omega \in \Omega_1$ is that we decide oppositely based on R_2.

$$\begin{aligned} & P\left(x|_{R_2} \equiv 1 | \ x|_{R_1} \equiv 1, l = |R_1 \cap R_2|\right) \\ = & \frac{1}{2^{N_2}} \binom{r_1 + r_2 - l}{r_2} 2^{N_2 - (r_1 - l)}, \end{aligned}$$

we have

$$\alpha = 1 - \sum_{l=0}^{r_1} \frac{1}{2^{N_2}} \binom{r_1 + r_2 - l}{r_2} 2^{N_2 - (r_1 - l)} P_l$$

The type two error takes place if $x|_{N_1} \equiv 1$ but $\omega \notin \Omega_1$. We now count the possible cases when the selective bits of Ω_2 and of Ω_1 overlap at most in $N - N_1 - 1$ places and produce false positive result. Let again $m = N - r_1 - r_2 + l$, then

$$\begin{aligned} \beta &= P\left(\{x_i(\omega) = 1\}_{i=1}^{N-N_1} | \omega \notin \Omega_1\right) \\ &= \frac{1}{2^{N_1}} \sum_{l=0}^{r_2} \sum_{k=1}^{N_1} \binom{N-m}{r_1-k} \binom{m}{k} 2^k P_l \end{aligned}$$

Given that the true size of Ω and Ω_1 is typically unknown, the theoretical values of the type one and two errors can not be calculated. On the other hand, the empirical values can be obtained based on a proper sample.

References

1. Jakovác, A.; Berényi, D.; Pósfay, P. Understanding understanding: A renormalization group inspired model of (artificial) intelligence. *arXiv* **2020**, arXiv:2010.13482.
2. Le-Khac, P.H.; Healy, G.; Smeaton, A.F. Contrastive representation learning: A framework and review. *IEEE Access* **2020**, *8*, 193907–193934. [CrossRef]
3. Yoshua, B.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828.
4. Biró, T.S.; Jakovác, A. Entropy of Artificial Intelligence. *Universe* **2022**, *8*, 53. [CrossRef]
5. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
6. Paterson, M.S. *An Introduction to Boolean Function Complexity*; Computer Science Department, School of Humanities and Sciences, Stanford University: Stanford, CA, USA, 1976.