P-T Probability Framework and Semantic Information G Theory Tested by Seven Difficult Tasks

Chenguang Lu

P-T Probability Framework and Semantic Information G Theory Tested by Seven Difficult Tasks

Chenguang Lu [0000-0002-8669-0094]

College of Intelligence Engineering and Mathematics,
Liaoning Technical University, Fuxin, Liaoning, 123000, China
lcguang@foxmail.com

Abstract. To applying information theory to more areas, the author proposed semantic information G theory, which is a natural generalization of Shannon's information theory. This theory uses the P-T probability framework so that likelihood functions and truth functions (or membership functions), as well as sampling distributions, can be put into the semantic mutual information formula at the same time. Hence, we can connect statistics and (fuzzy) logic. Rate-distortion function $R(D)$ becomes rate-verisimilitude function $R(G)$ (G is the lower limit of the semantic mutual information) when the distortion function is replaced with the semantic information function. Seven difficult tasks are 1) clarifying the relationship between minimum information and maximum entropy in statistical mechanics, 2) compressing images according to visual discrimination, 3) multilabel learning for obtaining truth functions or membership functions from sampling distributions, 4) feature classifications with maximum mutual information criterion, 5) proving the convergence of the expectation-maximization algorithm for mixture models, 6) interpreting Popper's verisimilitude and reconciling contradiction between the content approach and the likeness approach, and 7) providing practical confirmation measures and clarifying the raven paradox. This paper simply introduces the mathematical methods for these tasks and the conclusions. The P-T probability framework and the semantic information G theory should have survived the tests. They should have broader applications. Further studies are needed for combining them with neural networks for machine learning.

Keywords: Semantic information, Probability framework, Machine learning, Philosophy of science, Rate distortion, Boltzmann distribution, Maximum mutual information, Verisimilitude, Confirmation measure.

1 Introduction

Although Shannon's information theory [1] has achieved great success since 1948, we cannot use it to measure semantic information. It is also not easy to apply this theory to machine learning because we cannot put likelihood functions in the information measure. According to this theory, we can only use the distortion criterion instead of the information criterion to optimize detections and classifications. In 1949, Weaver first

proposed to research semantic information [1]. In 1952, Carnap and Bar-Hillel [2] presented an outline of semantic information theory. There exist multiple different information theories relating to semantic information [3].

In 1993, the author proposed a generalized information theory [4,5]. Now it is called semantic information G theory or G theory [3], where "G" means generalization. Early G theory was used mainly for semantic information evaluations and data compression [4,5]. Recently, the author developed the P-T probability framework for G theory [6] so that likelihood functions and (fuzzy) truth functions (or membership functions) can be mutually converted by a pair of new Bayes' formulas. G theory now can resolve more problems with machine learning [3] and philosophy of science [6,7].

This paper introduces the P-T probability framework and G theory; and uses them to complete seven difficult tasks for testing them.

2 P-T Probability Framework and Semantic Information G Theory

2.1 From Shannon's Probability Framework to P-T Probability Framework

The probability framework used by Shannon is defined as follows.

Definition 1. X is a discrete random variable taking a value $x \in U=\{x_1, x_2, \ldots, x_m\}$; $P(x_i)$ is the limit of the relative frequency of event $X = x_i$. Y is a discrete random variable taking a value $y \in V = \{y_1, y_2, \ldots, y_n\}$; $P(y_j) = P(Y = y_j)$. Shannon names $P(X)$ the source, $P(Y)$ the destination, and $P(Y|X)$ the channel. The latter consists of Transition Probability Functions (TPF): $P(y_j|x)$, j=1, 2, …, n.

The P–T probability framework is defined as follows.

Definition 2. The y_j is a label or a hypothesis, $y_j(x_i)$ is a proposition, and $y_j(x)$ is a predicate. The θ_j is a fuzzy subset of universe U, $y_j(x)$ = "$x \in \theta_j$" = "x is in θ_j". The θ_j is also treated as a model or a set of model parameters. A probability that is defined with "= ", such as $P(y_j) = P(Y = y_j)$, is a statistical probability. A probability that is defined with "$\in$", such as $P(X \in \theta_j)$, is a logical probability denoted by $T(y_j) = T(\theta_j) = P(X \in \theta_j)$. $T(\theta_j|x) = P(x \in \theta_j) = P(X \in \theta_j|X = x) \in [0,1]$ is the truth function of y_j and the membership function of θ_j.

According to Davidson's truth condition semantics [8], the truth function of y_j ascertains the semantic meaning of y_j. Truth functions $T(\theta_j|x)$ (j = 1, 2, …, n) form a semantic channel. Bayes, Shannon, and the author used three types of Bayes' Theorem [3]. Two asymmetrical formulas can express the third type:

$$P(x|\theta_j) = T(\theta_j|x)P(x)/T(\theta_j),\ T(\theta_j) = \sum_i P(x_i)T(\theta_j|x_i), \tag{1}$$

$$T(\theta_j|x) = [P(x|\theta_j)/P(x)]/\max[P(x|\theta_j)/P(x)]. \tag{2}$$

2.2 From Shannon's Information Measure to Semantic Information Measure

Shannon's mutual information [1] is defined as:

$$I(X;Y)=\sum_j\sum_i P(x_i,y_j)\log\frac{P(x_i\mid y_j)}{P(x_i)}, \tag{3}$$

Replacing $P(x_i|y_j)$ with $P(x_i|\theta_j)$ after the log, we have semantic mutual information:

$$I(X;\Theta)=\sum_j\sum_i P(x_i,y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_j\sum_i P(x_i,y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}. \tag{4}$$

When $Y=y_j$, we have generalized Kullback-Leibler (KL) information:

$$I(X;\theta_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_i P(x_i\mid y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}, \tag{5}$$

Further, if $X=x_i$, we have semantic information of y_j about x_i (illustrated in Fig. 1):

$$I(x_i;\theta_j)=\log\frac{P(x_i\mid\theta_j)}{P(x_i)}=\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}. \tag{6}$$

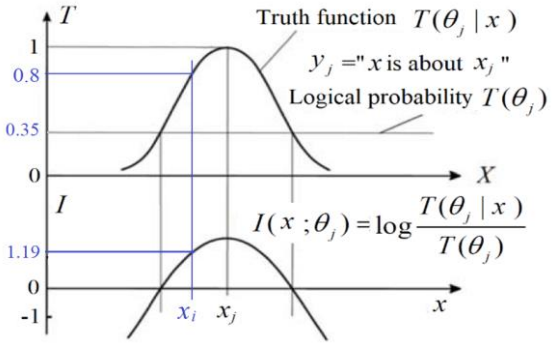


Fig. 1. Semantic information changes with x. When real x is x_i, the truth value is $T(\theta_j|x_i)$ = 0.8; information $I(x_i; \theta_j)$ is 1.19 bits. If x exceeds a certain range, the information is negative.

If the truth function is a Gaussian truth function, i.e., $T(\theta_j|x)=\exp[-(x-x_j)^2/(2\sigma_j^2)]$, $I(X; \Theta)$ is equal to the generalized entropy minus the mean relative squared error:

$$I(X;\Theta)=-\sum_j P(y_j)\log T(\theta_j)-\sum_i\sum_j P(x_i,y_j)(x_i-y_j)^2/(2\sigma_j^2). \tag{7}$$

We can find that the maximum semantic mutual information criterion is like the Regularized Least Square (RLS) criterion that is getting popular in machine learning.

Shannon [9] proposed information rate-distortion function $R(D)$ for lossy data compression limit. D is the upper limit of the average of distortion $d_{ij}=d(x_i, y_j)$. Replacing d_{ij} with $I_{ij}=I(x_i; y_j)$ and let G be the lower limit of $I(X; \Theta)$, we obtain $R(G)$ function:

$$G(s)=\sum_i\sum_j P(x_i)P(y_j \mid x_i)I_{ij}, \quad R(s)=sG(s)-\sum_i P(x_i)\log \lambda_i, \tag{8}$$

$$P(y_j \mid x_i)=P(y_j)P(x_i \mid \theta_j)^s / \lambda_i,\ \lambda_i=\sum_j P(y_j)P(x_i \mid \theta_j)^s. \tag{9}$$

Every $R(G)$ function has a point where information efficiency G/R=1 (see Fig. 2).

3 To Complete Seven Difficult Tasks

3.1 Clarifying the Relationship between Minimum Information and Maximum Entropy

Researchers found that Boltzmann's distribution can be used for machine learning [10] and is related to the rate-distortion function [11]. Using the P-T probability framework, we can explain this relationship better. The Boltzmann distribution is:

$$P(x_i \mid T)=\exp(-\frac{e_i}{kT}) / Z, \quad Z=\sum_i \exp(-\frac{e_i}{kT}), \tag{10}$$

where $P(x_i|T)$ is the probability of a particle in the ith state x_i with energy e_i, T is the absolute temperature, k is the Boltzmann constant, and Z is the partition function. If x_i is the state with the ith energy, G_i is the number of states with e_i, and G is the total number of all states, then $P(x_i) = G_i/G$. Hence, the above formula becomes

$$P(x_i \mid T)=P(x_i)\exp(-\frac{x_i}{kT}) / Z', \quad Z'=\sum_i P(x_i)\exp(-\frac{x_i}{kT}). \tag{11}$$

Now, we can find that $\exp[-e_i/(kT)]$ can be treated as a truth function or a Distribution Constraint Function (DCF), Z' as a logical probability, and Eq. (11) as a Bayes' formula (see Eq. (1)). A DCF means that there should be $1 - P(x|y_j) \le 1 - P(x|\theta_j)$. The author [4] has proved that for given $P(X)$ and a group of DCFs: $T(\theta_{xi}|y)$, i=1, 2, …, m, the minimum Shannon's mutual information $R(\Theta)$ is equal to the semantic mutual information $I(Y; \Theta)$, e.g. $R(\Theta) = R(D)$ [7]. From equation $F = E - TS$ (F is Helmholtz free energy, E is total energy, and S is entropy), we can derive [7]

$$R(\Theta)=-\sum_j P(y_j)\frac{e_j}{kT_j}-\sum_j P(y_j)\ln(Z_j / G)=\ln G - S/(kN), \tag{12}$$

This formula indicates the maximum entropy principle is equivalent to the minimum mutual information principle.

3.2 *R*(*G*) Function for Compressing Data according to Visual Discrimination

Suppose that x_i represents a gray level, a color, or a pixel with a certain position and color, y_j is the corresponding perception, which means "x is about x_j." The visual discrimination function is $T(\theta_j|x)=\exp[-(x-x_j)/(2d^2)]$. For given $P(x)$, the subjective information function is $I(x;\ \theta_j)=\log[T(\theta_j|x)/T(\theta_j)$. Then we can obtain the relationship between R, G, and d, as shown in Fig. 2. Fig. 2. reveals that the matching point (R=G) changes with the discrimination, and too high resolution is unnecessary. For more results, see [4,5].

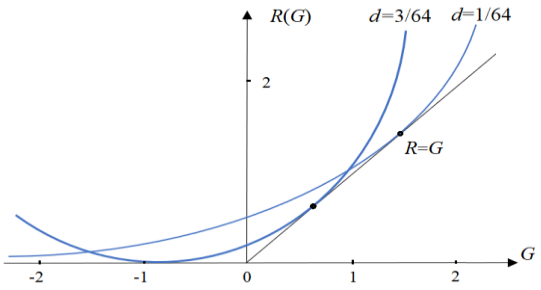


Fig. 2. Two $R(G)$ functions for different discrimination parameters d.

3.3 Multilabel Learning for Labels' Extensions or Truth Functions

We need to obtain truth functions, membership functions, or similarity functions from samples or sampling distributions in multilabel learning. Multilabel learning is difficult [15]; we can only use a pair of Logistic functions for two complementary labels' learning. However, with the P-T probability framework, multilabel learning is also easy. From the people ages' prior and posterior distributions $P(x)$, $P(x|$" adult"), and $P(x|$" elder") (see Fig. 3). Can we find the extensions or the truth functions of two labels?

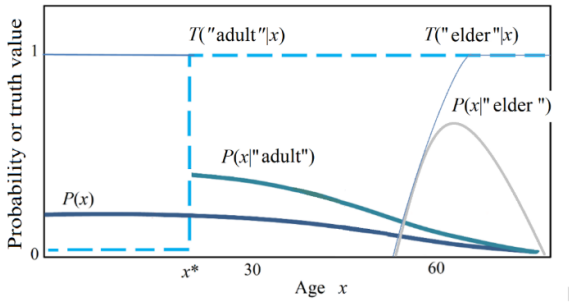


Fig. 3. Solving the truth functions of "adult" and "elder" using prior and posterior distributions. The human brain can guess T("adult" $|x$) and T("elder" $|x$).

Let **D** be a sample $\{(x(t), y(t))|t = 1 \text{ to } N;\ x(t) \in U;\ y(t) \in V\}$, where $(x(t), y(t))$ is an example. We can obtain sampling distribution $P(x|y_j)$ from **D**. According to Fisher's

maximum likelihood estimation, the optimized $P(x|\theta_j)$ is $P^*(x|\theta_j)= P(x|y_j)$. According to the third type of Bayes' Theorem, we have the optimized truth functions:

$$T^*(\theta_j|x) = [P^*(x|\theta_j)/P(x)]/\max(P^*(x|\theta_j)/P(x)) = [P(x|y_j)/P(x)]/\max(P(x|y_j)/P(x)). \quad (13)$$

We can use the above formula to solve two truth functions in Fig. 3. Further, we have

$$T^*(\theta_j|x) = P(y_j|x)/\max(P(y_j|x)),\ j = 1, 2, \ldots, n. \quad (14)$$

If samples are not big enough, we may use the generalized KL formula to obtain

$$T^*(\theta_j \mid x) = \arg\max_{T(\theta_j|x)} \sum_i P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)}. \quad (15)$$

We call the above method for the truth function Logical Bayesian Inference [3].

For classifications, we can use the maximum semantic information classifier:

$$y_j^* = f(x) = \arg\max_{y_j} \log I(x;\theta_j) = \arg\max_{y_j} \log[T(\theta_j \mid x) / T(\theta_j)], \quad (16)$$

which is compatible with the maximum likelihood classifier.

3.4 The Channels Matching Algorithm for Maximum Mutual Information (MMI) Classifications

In Shannon's information theory, the distortion criterion instead of the information criterion is used to detect and classify. Without the classification, we cannot express mutual information, whereas we cannot optimize the classification without mutual information's expression. However, G theory can avoid this loop [3].

Assume that we classify every instance with unseen true label x according to its observed feature $z \in C$. That is to provide a classifier $y=f(z)$ to get a label y (see Fig. 4).

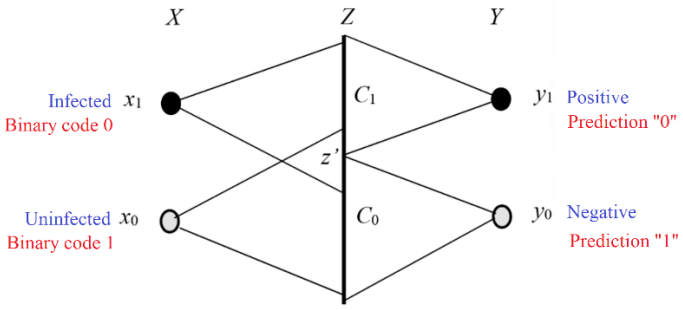


Fig. 4. Illustrating the medical test and the signal detection. We choose y_j according to $z \in C_j$.

Let C_j be a subset of C and $y_j=f(z|z \in C_j)$; hence $S=\{C_1, C_2, \ldots\}$ is a partition of C. Our aim is, for given $P(x, z)$, to find the optimized S:

$$S^* = \arg\max_S I(X;\theta|S) = \arg\max_S \sum_j \sum_i P(C_j)P(x_i \mid C_j)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)}. \quad (17)$$

Matching I: We obtain the Shannon channel for given S:

$$I(X;\theta_j|z) = \sum_i P(x_i \mid z)I(x_i;\theta_j)\ ,\ j=0,1,\ldots,n, \quad (18)$$

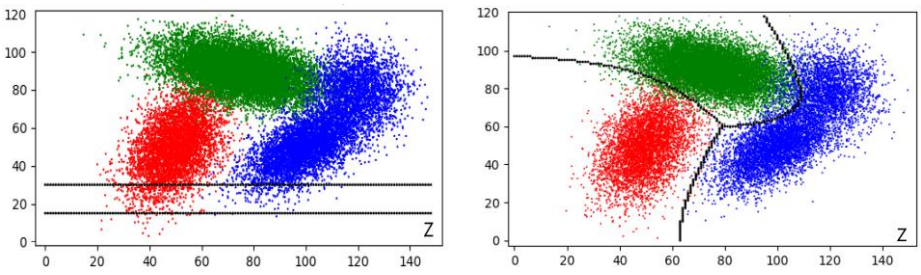
From this channel, we can obtain $T^*(\theta|x)$ (see Eq. (14)) and the semantic information $I(x_i; \theta_j)$. For given z, we have conditional information or reward functions:

$$I(X;\theta_j|z) = \sum_i P(x_i \mid z)I(x_i;\theta_j)\ ,\ j=0,1,\ldots,n. \quad (19)$$

Matching II: Let the Shannon channel match the semantic channel by the classifier:

$$y_j^* = f(z) = \arg\max_{y_j} I(X_i;\theta_j|z), j=0,1,\ldots,n. \quad (20)$$

Repeat Matching I and II until S converges to S^*. Fig. 5 shows an example.



(a) (b)

Fig. 5. The MMI classification. (a) Bad initial partition; (b) After two iterations.

After two iterations, the MMI is 1.0434 bits. The convergent MMI is 1.0435 bits. The result indicates that this algorithm is high-speed.

Using the $R(G)$ function, we can easily prove the above algorithm converges [3].

3.5 The Convergence Proof and the Improvement of the EM algorithm for Mixture Models

If a probability distribution $P_\theta(x)$ comes from n likelihood functions' mixture, e. g.,

$$P_\theta(x) = \sum_{j=1}^{n} P(y_j)P(x \mid \theta_j), \quad (21)$$

then we call $P_\theta(x)$ a mixture model. If every predictive model $P(x|\theta_j)$ is a Gaussian function, then $P_\theta(x)$ is a Gaussian mixture model. Assume that sampling distribution $P(x)$ comes from the mixture of two true models with ratios $P^*(y_1)$ and $P^*(y_2)=1-P^*(y_1)$. That is $P(x)=P^*(y_1)P(x|\theta_1^*)+P^*(y_2)P(x|\theta_2^*)$. Our task is to find the true model parameters and mixture proportions θ_1^*, θ_2^*, and $P^*(y_1)$ from $P(x)$.

The EM algorithm includes two steps [13,14]:

E-step: Write the conditional probability functions (e. g., the Shannon channel):

$$P(y_j \mid x) = P(y_j)P(x \mid \theta_j) / P_\theta(x), P_\theta(x) = \sum_j P(y_j)P(x \mid \theta_j). \tag{22}$$

M-step: Improve $P(y)$ and θ to maximize the complete data log-likelihood:

$$\begin{aligned} Q &= \sum_i \sum_j P(x_i)P(y_i \mid x_i) \log P(x_i, y_j \mid \theta) \\ &= L_X(\theta) + \sum_i \sum_j P(x_i)P(y_j \mid x_i) \log P(y_j \mid x_i), \end{aligned} \tag{23}$$

where $L_X(\theta)=\sum_i P(x_i)\log P_\theta(x_i)$ is the log-likelihood as the objective function. If Q cannot be improved further, then end iterations; otherwise, go to the E-step.

The M-step can be divided into two steps: the M1-step for optimizing $P(y)$ and the M2-step for maximizing semantic mutual information $I(X;\Theta)$ by letting $P(x|\theta_j)=P(x|y_j)$. The iterative method for the rate-distortion function reminds us that we can repeat Eq. (23) and $P^{+1}(y_j)=\sum_i P(x_i)P(y_j|x_i)$ until $P^{+1}(y_j)=P(y)$. The improved EM algorithm is called the Channels Matching EM (CM-EM) algorithm [3,15], which repeats the M1-step many or several times so that $P^{+1}(y)\approx P(y)$.

Many researchers believe that Q and $L_X(\theta)$ are always positively correlated; we can achieve maximum $L_X(\theta)$ by maximizing Q. However, the author found that this is not true. We need reliable convergence proof to avoid blind improvements.

Using the semantic information method, we can derive [3]

$$H(P \| P_\theta)=H(X) - H_\theta(X) = R - G + H(Y \| Y^{+1}), \tag{24}$$

$$H(Y^{+1}||Y) = \sum_j P^{+1}(y_j) \log[P^{+1}(y_j) / P(y_j)]. \tag{25}$$

Using the variational and iterative methods that Shannon [9] and others used for analyzing the rate-distortion function $R(D)$, we can prove that both the E-step and the M1-step minimize $H(P||P_\theta)$ or maximize $L_X(\theta)$ [15].

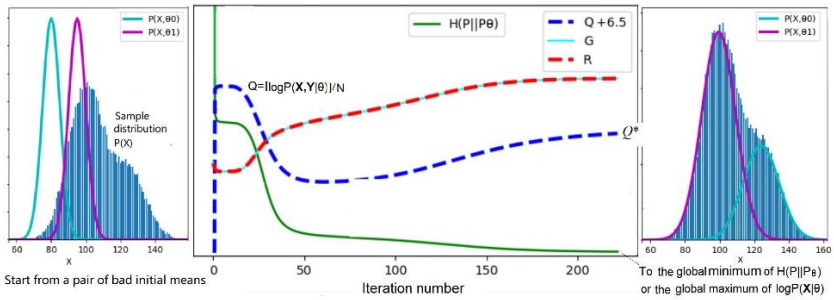


Fig. 6. Q and $H(P||P_\theta)$ change with iterations. The EM algorithm needs about 340 iterations, whereas the CM-EM algorithm needs about 240 iterations. The sample size is 50000.

Fig. 6 shows an example used for the Deterministic Annealing EM (DAEM) algorithm in [14]. The true model is (μ_1*, μ_2*, σ_1*, σ_2*, P*(y_1))= (80, 95, 5, 5, 0.5)); the initial model is (μ_1, μ_2, σ_1, σ_2, P*(y_1))= (125, 100, 10, 10, 0.7)). This example indicates that Q does not always increase while $H(P||P_\theta)$ decreases or $L_X(\theta)$ increases; the CM-EM algorithm is better in this case. For the detailed convergence proof and the improvement, see [15].

3.6 To Resolve the Problem with Popper's Verisimilitude and Explain the RLS Criterion

To evaluate hypotheses, Popper [16] replace trueness with verisimilitude (or truthlikeness [17]). Researchers use two approaches to interpret verisimilitude: the content approach and the consequence approach [17]. The former emphasizes tests' severity, unlike the latter that emphasizes hypotheses' truth or closeness to truth. Some researchers think that the content approach and the likeness approach are irreconcilable.

The truth function $T(\theta_j|x)$ is also the confusion probability function; it reflects the likeness between x and x_j. The x_i (e. g., $X = x_i$) is the consequence, and the distance between x_i and x_j in the feature space reflects the likeness. The $\log[1/T(\theta_j)]$ represents the testing severity and potential information content. Using the formula for $I(x_i; \theta_j)$, we can easily explain an often-mentioned example: why "the sun has 9 satellites" (8 is true) has higher verisimilitude than "the sun has 100 satellites" [17].

Now, we can explain why the RLS criterion is getting popular. It is similar to the maximum mean verisimilitude criterion and the maximum semantic mutual information criterion.

3.7 To Provide Practical Confirmation Measures and Clarify the Raven Paradox

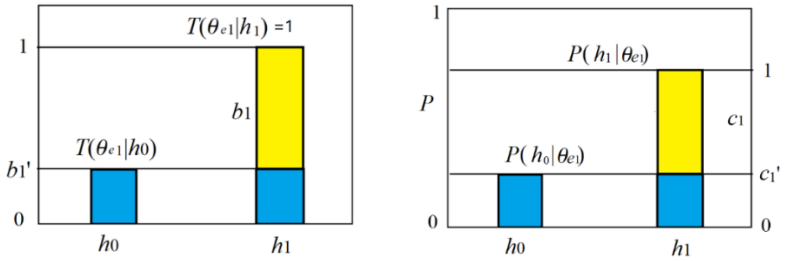
There have been many confirmation measures [7]. Researchers wish that 1) a confirmation measure can be used to evaluate tests and predictions like likelihood ratio; 2) it can be used to clarify the raven paradox.

We use the medical test, shown in Fig. 4, as an example to explain the two measures. Now x becomes h, and y becomes e. A major premise to be confirmed is "if e_1 then h_1" (e.g., $e_1 \rightarrow h_1$). A confirmation measure is denoted by $c(e \rightarrow h)$. We can obtain four examples' numbers a, b, c, and d (see Table 1) to construct confirmation measures for a given classification. We wish that a confirmation measure $c(e_1 \rightarrow h_1)$ changes between -1 and 1 and possesses consequence symmetry: $c(e_1 \rightarrow h_1) = -c(e_1 \rightarrow h_1)$ [7].

Table 1. The numbers of four examples of confirmation measures.

	e_0 **(negative)**	e_1 **(positive)**
h_1 (infected)	b	a
h_0 (uninfected)	d	c

We regard the truth function of $e_1(h)$ as a believable part plus an unbelievable part, as shown in Fig. 7.



(a) (b)

Fig. 7. The truth function (a) or the likelihood function (b) can be regarded as a believable part plus an unbelievable part.

The optimized b_1 is the degree of confirmation of major premise $e_1 \rightarrow h_1$ [7]. It is

$$b_1^* = b^*(e_1 \rightarrow h_1) = \frac{P(e_1|h_1) - P(e_1|h_0)}{\max(P(e_1|h_1), P(e_1|h_0))} = \frac{ad - bc}{\max(a(c+d), c(a+b))} = \frac{LR^+ - 1}{\max(LR^+, 1)}, \quad (268)$$

where LR^+ is the positive likelihood ratio. Since b_1^* reflects how well the test serves as a means or a channel, we call $b^*(e \rightarrow h)$ the channel confirmation measure, which is compatible with the likelihood ratio measure.

Suppose that likelihood function $P(h|\theta_{e1})$ includes a believable part and an unbelievable part, as shown in Fig. 7 (b), we derive the prediction confirmation measure

$$c_1^* = c^*(e_1 \rightarrow h_1) = \frac{P(h_1, e_1) - P(h_0, e_1)}{\max(P(h_1, e_1), P(h_0, e_1))} = \frac{a - c}{\max(a, c)}, \quad (27)$$

which indicates how well the test serves as a prediction.

We can use measure c^* to clarify the Raven Paradox. Hemple [18] proposed the Raven Paradox. According to the Equivalence Condition (EC) in the classical logic, "if

x is a raven, then *x* is black" (Rule I) is equivalent to "if *x* is not black, then *x* is not a raven" (Rule II). A piece of white chalk supports Rule II; hence it also supports Rule I. However, according to the Nicod Criterion (NC), a black raven supports Rule I, a non-black raven undermines Rule I, and a non-raven thing, such as a black cat or a piece of white chalk, is irrelevant to Rule I. Hence, there exists a paradox between EC and NC.

Almost all confirmation measures [7], only measure c^* supports the NC and objects the EC because c^* is only affected by a and c. For example, when a=6, c=1 and b=d=10, $c^*(e_1 \rightarrow h_1) = (6 - 1)/6 = 5/6$. The author has demonstrated that except for c^*, all popular confirmation measures cannot explain that a black raven can confirm "ravens are black" better than a piece of white chalk [7].

4 Summary

This paper has shown that using the P-T probability framework can obtain truth function or membership functions from sampling distribution and connect statistics and logic (fuzzy logic). It has been explained that the semantic information criterion is the verisimilitude criterion and similar to the RLS criterion. Using the semantic information criterion instead of the distortion criterion, we can solve the MMI classification better.

This paper has introduced how to apply the P-T probability framework and the G theory to semantic communication, machine learning, and philosophy of science for completing some challenging tasks. These applications reveal that the P-T probability framework and the G theory have great potential. We should be able to find more applications. We need further studies for combining the P-T probability framework and the G theory with neural networks for machine learning.

References

1. Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. The University of Illinois Press, Urbana (1949).
2. Bar-Hillel Y, Carnap R.: An outline of a theory of semantic information. Tech. Rep. No. 247, Research Lab. of Electronics, MIT (1952).
3. Lu, C.: Semantic information G theory and logical Bayesian inference for machine learning. Information 10(8), 261 (2019).
4. Lu, C. A Generalized Information Theory. China Science and Technology University Press, Hefei (1993).
5. Lu, C. A generalization of Shannon's information theory. Int. J. Gen. Syst. 28(6), 453–490 (1999).
6. Lu, C. The P–T Probability Framework for Semantic Communication, Falsification, Confirmation, and Bayesian Reasoning. Philosophies 5(4), 25 (2020).
7. Lu, C. Channels' confirmation and predictions' confirmation: From the medical test to the raven paradox. Entropy *22*(4), 384 (2020).
8. Davidson, D. Truth and meaning. Synthese 17(3), 304–323 (1967).
9. Shannon, C.E.: Coding theorems for a discrete source with a fidelity criterion. IRE Nat. Conv. Rec. 4, 142–163 (1959).

10. H. A. David, Hinton, G. E.; Sejnowski, T. J.: A learning algorithm for Boltzmann machines. Cognitive Science 9(1), 147–169 (1985).
11. Li, Q., Chen, Y.: Rate Distortion Via Restricted Boltzmann Machines, 56th Annual Allerton Conference on Communication, Control, and Computing, Monticello, 1052–1059 (2018).
12. Zhang, M. L., Zhou, Z. H.: A review on multilabel learning algorithm. IEEE Transactions on Knowledge and Data Engineering 26(8), 1819-1837 (2014).
13. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society, Series B, 39(1), 1–38 (1997).
14. Ueda, N., Nakano, R.: Deterministic annealing variant of the EM algorithm. In G. Tesauro et al. (eds.), Advances in NIPS 7, pp. 545– 552. Cambridge, MA: MIT Press, (1995).
15. Lu, C.: From the EM Algorithm to the CM-EM Algorithm for Global Convergence of Mixture Models. https://arxiv.org/abs/1810.11227. last accessed 2020-12-10.
16. Popper, K.: Conjectures and Refutations. Repr. Routledge, London and New York (1963/2005)
17. Oddie, G.: Truthlikeness, the Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/truthlikeness/, last accessed 2020-12-1
18. Hempel, C. G.: Studies in the Logic of Confirmation. Mind**,** 54, 1–26 and 97–121 (1945).