Chapter 1
Information Theory and Its Relation to Machine Learning

Bao-Gang Hu

Abstract In this position paper, I first describe a new perspective on machine learning (**ML**) by four basic problems (or levels), namely "What to learn?", "How to learn?", "What to evaluate?", and "What to adjust?". The paper stresses more on the first level of "What to learn?", or "Learning Target Selection". Toward this primary problem within the four levels, I briefly review the existing studies about the connection between information theoretical learning (**ITL** [1]) and machine learning. A theorem is given on the relation between the empirically-defined similarity measure and information measures. Finally, a conjecture is proposed for pursuing a unified mathematical interpretation to learning target selection.

> *From the Tao comes one, from one comes two, from two comes three, and from three comes all things.* [2].
> —by Lao Tzu (ca. 600–500 BCE).
> *Nature is the realization of the simplest conceivable mathematical ideas.* [3].
> —by Albert Einstein (1879–1955).

1.1 Introduction

Machine learning learning is the study and construction of systems that can learn from data. The systems are called *learning machines*. When Big Data emerges increasingly, more learning machines are developed and applied in different domains. However, the ultimate goal of machine learning study is *insight*, not machine itself. By the term insight I mean *learning mechanisms* in descriptions of

B.-G. Hu (✉)
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
e-mail: hubg@nlpr.ia.ac.cn

mathematical principles. In a loose sense, learning mechanisms can be regarded as the natural entity. As the "*Tao* (道)" reflects the most fundamental of the universe by Lao Tzu (老子), Einstein suggests that we should pursue the simplest mathematical interpretations to the nature. Although learning mechanisms are related to the subjects of psychology, cognitive, and brain science, this paper stresses on the exploration of mathematical principles for interpretation of learning mechanisms. Up to now, we human beings are still far away from deep understanding ourself of learning mechanisms in terms of mathematical principles. It is the author's belief that "mathematical-principle-based machine" might be more important and critical than "brain-inspired machine" in the study of machine learning.

The purpose of this position paper is to put forward a new perspective and a novel conjecture within the study of machine learning. In what follows I will present four basic problems (or levels) in machine learning. The study on information theoretical learning is briefly reviewed. A theorem between the empirically defined similarity measures and information measures are given. Based on the existing investigations, a conjecture is proposed in this paper.

1.2 Four Basic Problems (or Levels) in Machine Learning

For information processing by a machine, in the 1980s, Marr [4] proposed a novel methodology by three distinct yet complementary levels, namely "Computational theory", "Representation and algorithm", and "Hardware implementation", respectively. Although the three levels are "coupled" loosely, the distinction is of great necessity to isolate and solve problems properly and efficiently. In 2007, Poggio [5] described another set of three levels on learning, namely "Learning theory and algorithms", "Engineering applications", and "Neuroscience: models and experiments", respectively. Apart from showing a new perspective, one of the important contributions of this methodology is adding a *closed loop* between the levels. These studies are enlightening because they show that complex objects or systems should be addressed by decompositions with different, yet basic, problems. The methodology is considered to be *reductionism* philosophically.

In this paper, I propose a novel perspective on machine learning by four levels shown in Fig. 1.1. The levels correspond to four basic problems. The definition of each level is given below.

Definition 1.1 "What to learn" is a study on identifying learning target(s) to the given problem(s), which will generally involve two distinct sets of representations (Fig. 1.2) defined below.

Definition 1.2 "Linguistic representation" reflects a *high-level* description in a *natural language* about the expected learning information. This study is more related to linguistics, psychology, and cognitive science.

Fig. 1.1 Four basic problems
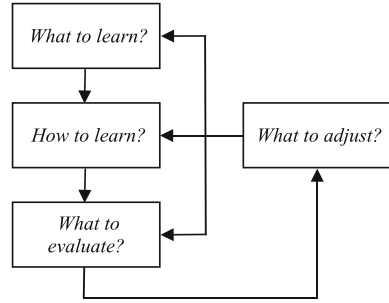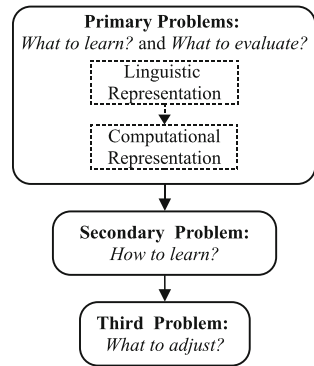(or levels) in machine
learning

Fig. 1.2 Design flow
according to the basic
problems in machine learning

Definition 1.3 "Computational representation" is to define the expected learning
information based on *mathematical notations*. It is a relatively *low-level* repre-
sentation which generally includes objective functions, constraints, and optimiza-
tion formations.

Definition 1.4 "How to learn?" is a study on learning process design and imple-
mentations. Probability, statistics, utility, optimization, and computational theories
will be the central subjects. The main concerns are generalization performance,
robustness, model complexity, computational complexity/cost, etc. The study may
include physically realized system(s).

Definition 1.5 "What to evaluate?" is a study on "evaluation measure selection"
where evaluation measure is a mathematical function. This function can be the same
or different with the objective function defined in the first level.

Definition 1.6 "What to adjust?" is a study on dynamic behaviors of a machine
from adjusting its component(s). This level will enable a machine with a func-
tionality of "evolution of intelligence".

The first level is also called "learning target selection". The four levels above are
neither mutually exclusive, nor collectively exhaustive to every problems in
machine learning. We call them *basic* so that the extra problems can be merged
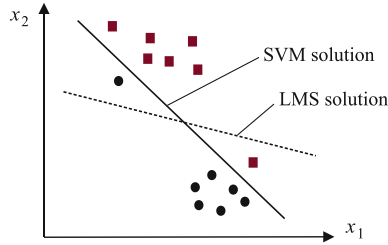
Fig. 1.3 Learning target selection within linearly separated dataset. (after [8] in Figs. 5–17). *Black Circle* Class 1, *Ruby Square* Class 2

within one of levels. Figures 1.1 and 1.2 illustrate the relations between each level in different contexts, respectively. The problems within four levels are all inter-related, particularly for "What to learn?" and "What to evaluate?" (Fig. 1.2). "How to learn?" may influence to "What to learn?", such as convexity of the objective function or scalability to learning algorithms [6] from a computational cost con-sideration. Structurally, "What to adjust?" level is applied to provide the multiple closed loops for describing the interrelations (Fig. 1.1). Artificial intelligence will play a critical role via this level. In the "knowledge driven and data driven" model [7], the benefits of utilizing this level are shown from the given examples by *removable singularity hypothesis* to "Sinc" function and *prior updating* to Mackey-Glass dataset, respectively. Philosophically, "What to adjust?" level remedies the intrinsic problems in the methodology of *reductionism* and offers the functionality power for being *holism*. However, this level receives even less attention while learning process holds a *self-organization* property.

I expect that the four levels show a novel perspective about the basic problems in machine learning. Take an example shown in Fig. 1.3 (after Duda et al. [8], Figs. 5–17). Even for the linearly separable dataset, the learning function using least mean square (**LMS**) does not guarantee a "minimum-error" classification. This example demonstrates two points. First, the computational representation of LMS is not compatible with the linguistic representation of "minimum-error" classification. Second, whenever a learning target is wrong in the computational representation, one is unable to reach the goal from Levels 2 and 3. Another example in Fig. 1.4 shows why we need two sub-levels in learning target selection. For the given character (here is *Albert Einstein*), one does need a linguistic representation to describe "(un)likeness" [9] between the original image and caricature image. Only when a linguistic representation is well-defined, is a computational measure of similarity *possibly proper* in caricature learning. The meaning of *possibly proper* is due to the difficulty in the following definition.

Definition 1.5 "Semantic gap" is a difference between the two sets of represen-tations. The gap can be linked by two ways, namely a *direct way* for describing a connection from linguistic representation to computational representation, and an *inverse way* for a connection opposite to the direct one.
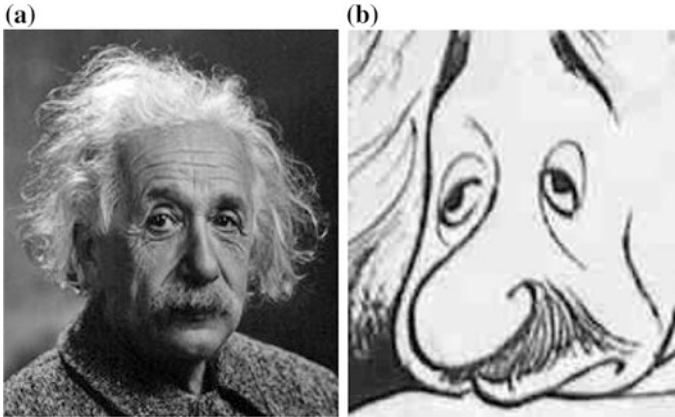
Fig. 1.4 Example of "What to learn?" and a need of defining a linguistic representation of similarity for the given character. **a** Original image (http://en.wikipedia.org/wiki/Albert_Einstein). **b** Caricature image drawn by A. Hirschfeld (http://www.georgejgoodstadt.com/goodstadt/hirschfeld.dca)

In this paper, I extend the definition of the gap in [10] by distinguishing two ways. The gap reflects one of the critical difficulties in machine learning. For the direct-way study, the difficulty source mostly comes from *ambiguity* and *subjectivity* of the linguistic representation (say, on *mental* entity), which will lead to an *ill-defined* problem. While sharing the same problem, an inverse-way study will introduce an extra challenge called *ill-posed* problem, in which there is no *unique* solution (say, from a 2D image to 3D objects).

Up to now, we have missed much studies on *learning target selection* if comparing with a study of *feature selection*. When "What to learn?" is the most primary problem in machine learning, we do need a systematic, or comparative, study on this subject. The investigations from [11, 12] into *discriminative* and *generative* models confirm the importance of learning target selection in the vein of computational representation. From the investigations, one can identify the advantages and disadvantages of each model for applications. A better machine gaining the benefits from both models is developed [13]. Furthermore, the subject of "What to learn?" will provide a strong driving force to machine learning study in seeking "the fundamental laws that govern all learning processes" [14].

Take a decision rule about "Less costs more"[1] for example. Generally, Chinese people classify object's values according to this rule. In *Big Data* processing, the useful information, which often belongs to a *minority class*, is extracted from massive datasets. While an English idiom describes it as "Finding a needle in a

---

[1] This rule is translated from Chinese saying, "Suddenly" in Pinyin "Wu Yi Xi Wei Gui". The translation is modified from the English phase "Less is more" which usually describes *simplicity* in design.

haystack", the Chinese saying refers to "Searching a needle in a sea (Needle in a haystack)". Users may consider that an error from a *minority class* will cost heavier than that from a *majority class* in their searching practices. This consideration will derive a decision rule like "Less costs more". The rule will be one of the important strategies in Big Data processing. Two questions can be given to the example. What is the *mathematical principle* (or *fundamental law*) for supporting the decision rule of "Less costs more"? Is it a Bayesian rule? Machine learning study does need to answer the questions.

1.3 Information Theoretical Learning

Shannon introduced "entropy" concept as the basis of information theory [15]:

$$H(Y) = -\sum_y p(y) \log_2 p(y), \tag{1.1}$$

where Y is a discrete random variable with *probability mass function $p(y)$*. Entropy is an expression of *disorder* to the information. From this basic concept, the other *information measures* (or *entropy functions*) can be formed (Table 1.1), where $p(t, y)$ is the *joint distribution* for the target random variable T and prediction random variable Y, and $p(t)$ and $p(y)$ are called *marginal distributions*. We call them *measures* because some of them do not satisfy the *metric* properties fully, like *KL divergence* (asymmetric). Other measures from information theory can be listed as learning criteria, but the measures in Table 1.1 are more common and sufficiently meaningful for the present discussion.

We can divide the learning machines, in view of "mathematical principles", within two groups. One group is designed based on the empirical formulas, like *error rate* or *bound*, *cost* (or *risk*), *utility*, or *classification margins*. The other is on information theory [1, 16]. Therefore, a systematic study seems necessary to answer the two basic questions below [17]:

Table 1.1 Some information formulas and their properties as learning measures

Name	Formula	(Dis)similarity	(A)symmetry
Joint information	$H(T, Y) = -\sum_t \sum_y p(t, y) \log_2 p(t, y)$	Inapplicable	Symmetry
Mutual information	$I(T, Y) = \sum_t \sum_y p(t, y) \log_2 \frac{p(t,y)}{p(t)p(y)}$	Similarity	Symmetry
Conditional entropy	$H(Y|T) = -\sum_t \sum_y p(t, y) \log_2 p(y|t)$	Dissimilarity	Asymmetry
Cross entropy	$H(T; Y) = -\sum_z p_t(z) \log_2 p_y(z)$	Dissimilarity	Asymmetry
KL divergence	$KL(T, Y) = \sum_z p_t(z) \log_2 \frac{p_t(z)}{p_y(z)}$	Dissimilarity	Asymmetry

Q1. When one of the principal tasks in machine learning is to process data, can we apply entropy or information measures as a generic learning target for dealing with uncertainty of data in machine learning?

Q2. What are the relations between information learning criteria and empirical learning criteria, and the advantages and limitations in using information learning criteria?

Regarding the first question, Watanabe [18, 19] proposed that "learning is an entropy-decreasing process" and pattern recognition is "a quest for minimum entropy". The principle behind entropy criteria is to transform disordered data into ordered one (or *pattern*). Watanabe seems to be the first "to cast the problems of learning in terms of minimizing properly defined entropy functions" [20], and throws brilliant light on the learning target selection in machine learning.

In 1988, Zellner theoretically proved that Bayesian theorem can be derived from the optimal information processing rule [21]. This study presents a novel, yet important, finding that Bayesian theory is rooted in information and optimization concepts. Another significant contribution is given by Principe and his collaborators [1, 22] for the proposal of Information Theoretical Learning (**ITL**) as a generic learning target in machine learning. We consider ITL will stimulate us to develop new learning machines as well as "theoretical interpretations" of learning mechanisms. Take again the example of the decision rule about "Less costs more". Hu [23] demonstrates theoretically that Bayesian principle is unable to support the rule. When a minority class approximates to a zero population, Bayesian classifiers will tend to misclassify the minority class completely. The numerical studies [23, 24] show that *mutual information* provides positive examples to the rule. The classifiers based on *mutual information* are able to protect a minority class and automatically balance the error types and reject types in terms of population ratios of classes. These studies reveal a possible mathematical interpretation of learning mechanism behind the rule.

1.4 (Dis)similarity Measures in Machine Learning

When mutual information describes similarity between two variables, the other information measures in Table 1.1 are applied in a sense of dissimilarity. For better understanding of them, their graphic relations are shown in Fig. 1.5. If we consider the variable T provides a ground truth statistically (that is, $p(t) = (p_1, \ldots, p_m)$ with the *population rate* $p_i(i = 1, \ldots, m)$ is known and fixed), its entropy $H(T)$ will be the baseline in learning. In other words, when the following relations hold:

$$I(T, Y) = H(T; Y) = H(Y; T) = H(Y) = H(T), \text{or}$$
$$KL(T, Y) = KL(Y, T) = H(T|Y) = H(Y|T) = 0, \tag{1.2}$$

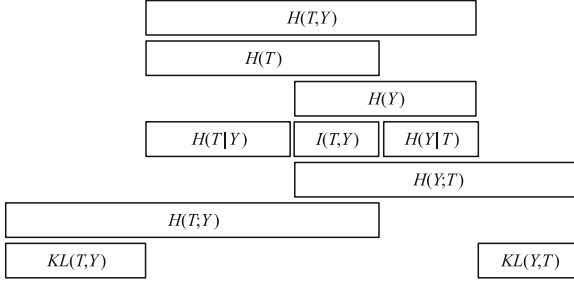we call the measures *reach* the baseline of $H(T)$.

Fig. 1.5 Graphic relations among joint information, mutual information, marginal information, conditional entropy, cross entropy, and *KL* divergences (modified based on [25] by including cross entropy and *KL* divergences)
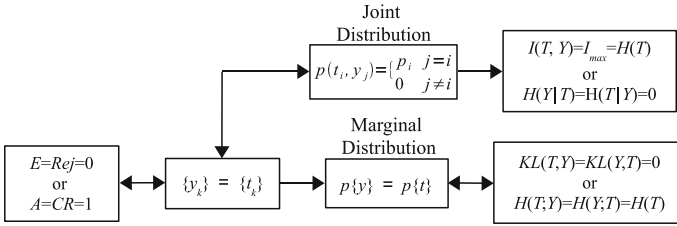

Fig. 1.6 Relations between exact classifications and mutual information, conditional entropy, cross entropy, and *KL* divergences

Based on the study in [26], further relations are illustrated in Fig. 1.6 between exact classifications and the information measures. We apply the notations of *E*, *Rej*, *A*, *CR* for the *error, reject, accuracy, and correct recognition rates*, respectively. Their relations are given by:

$$CR + E + Rej = 1,$$
$$A = \frac{CR}{CR + E}. \tag{1.3}$$

The form of $\{y_k\} = \{t_k\}$ in Fig. 1.6 describes an equality between the label variables in every sample. For a finite dataset, the empirical forms should be used for representing the distributions and measures [26]. Note that the link using "↔" indicates a *two-way* connection for equivalent relations, and "→" for a *one-way* connection. Three important aspects can be observed from Fig. 1.6:

  I. The *necessary* condition of exact classifications is that all the information measures reach the baseline of $H(T)$.
 II. When an information measure reaches the baseline of $H(T)$, it does not *sufficiently* indicate an exact classification.
III. The different locations of one-way connections result in the interpretations *why* and *where* the sufficient condition exists.

Although Fig. 1.6 only shows the relations to the information measures listed in Table 1.1 for the classification problems, its observations may extend to other information measures as well as to the other problems, like clustering, feature selection/extraction, image registrations, etc. When we consider machine learning or pattern recognition to be a process of data in a *similarity* sense (any dissimilarity measure can be transformed into similarity one [26]), one important theorem exists to describe their relations.

Theorem 1.1 *Generally, there is no one-to-one correspondence between the empirically-defined similarity measures and information measures.*

The proof is neglected in this paper, but it can be given based on the study of bounds between entropy and error (cf. [27] and references therein). The significance of Theorem 1.1 implies that an optimization of information measure may not guarantee to achieve an optimization of the empirically defined similarity measure.

1.5 Final Remarks

Machine learning can be exploited with different perspectives depending on the study goals of researchers. For in-depth understanding of the learning mechanisms mathematically, we can take learning machines as *human's extended sensory perception*. This paper stresses on identifying the primary problem in machine learning from a novel perspective. I define it as "What to learn?" or "learning target selection". Furthermore, two sets of representations are specified, namely "linguistic representation" and "computational representation". While a wide variety of computational representations have been reported in learning targets, we can argue if there exists a unified, yet fundamental, principle behind them. Towards this purpose, this paper extends the Watanabe's proposal [18, 19] and the studies from Zellner [21] and Principe [1] to a "conjecture of learning target selection" in the following descriptions.

Conjecture 1.1 *In a machine learning study, all computational representations of learning target(s) can be interpreted, or described, by optimization of entropy function(s).*

I expect that the proposal of the conjecture above will provide a new driving force not only for seeking fundamental laws governing all learning processes [14] but also for developing improved learning machines [28] in various applications.

References

1. Principe JC (2010) Information theoretic learning: Renyi's entropy and kernel perspectives. Springer, New York
2. Lao T (ca. 500 BCE) Tao Te Ching "Tao Te Ching"
3. Norton JD (2000) Nature is the realization of the simplest conceivable mathematical ideas: Einstein and the canon of mathematical simplicity. Stud Hist Philos Mod Phys 31:135–170
4. Marr D (2010) Vision. A computational investigation into the human representation and processing of visual information. The MIT Press, Cambridge
5. Poggio T (2007) How the brain might work: The role of information and learning in understanding and replicating intelligence. In: Jacovitt G et al (eds) Information: science and technology for the new century. Lateran University Press, New York, pp 45–61
6. Bengio Y, LeCun Y (2007) Scaling learning algorithms towards AI. Large-Scale Kernel Mach 34:1–41
7. Hu BG, Qu HB, Wang Y, Yang SH (2009) A generalized constraint neural networks model: associating partially known relationships for nonlinear regressions. Inf Sci 179:1929–1943
8. Duda RO, Hart PE, Stork D (2001) Pattern classification, 2nd edn. Wiley, New York
9. Brennan SE (1985) Caricature generator: the dynamic exaggeration of faces by computer. Leonardo 40:392–400
10. Smeulders AW, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22:1349–1380
11. Rubinstein YD, Hastie T (1997) Discriminative vs informative learning. KDD 5:49–53
12. Ng A, Jordan MI (2002) On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: NIPS
13. Bishop CM, Lasserre J (2007) Generative or discriminative? Getting the best of both worlds. In: Bernardo JM et al (eds) Bayesian Statistics, vol 8. Oxford University Press, Oxford, pp 3–23
14. Mitchell TM (2006) The discipline of machine learning. Technical Report CMU-ML-06-108, Carnegie Mellon University
15. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423
16. Yao YY (2003) Information-theoretic measures for knowledge discovery and data mining. In: Karmeshu (ed) Entropy measures, maximum entropy principle and emerging applications. Springer, Berlin, pp 115–136
17. Hu BG, Wang Y (2008) Evaluation criteria based on mutual information for classifications including rejected class. Acta Automatica Sinica 34:1396–1403
18. Watanabe S (1980) Pattern recognition as a quest for minimum entropy. Pattern Recognit 13:381–387
19. Watanabe S (1981) Pattern recognition as conceptual morphogenesis. IEEE Trans Pattern Anal Mach Intell 2:161–165
20. Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 21:660–674
21. Zellner A (1988) Optimal information processing and Bayes's theorem. Am Stat 42:278–284
22. Principe JC, Fisher JW III, Xu D (2000) Information theoretic learning. In: Haykin S (ed) Unsupervised adaptive filtering. Wiley, New York, pp 265–319
23. Hu BG (2014) What are the differences between Bayesian classifiers and mutual-information classifiers? IEEE Trans Neural Netw Learn Syst 25:249–264
24. Zhang X, Hu BG (2014) A new strategy of cost-free learning in the class imbalance problem. IEEE Trans Knowledge Data Eng 26:2872–2885
25. Mackay DJC (2003) Information theory, inference, and learning algorithms. Cambridge University Press, Cambridge
26. Hu BG, He R, Yuan XT (2012) Information-theoretic measures for objective evaluation of classifications. Acta Automatica Sinica 38:1170–1182

27. Hu BG, Xing HJ (2013) A new approach of deriving bounds between entropy and error from joint distribution: case study for binary classifications. arXiv:1205.6602v1[cs.IT]
28. He R, Hu BG, Yuan XT, Wang L (2014) Robust recognition via information theoretic learning. Springer, Heidelberg