 doi:10.20944/preprints201710.0076.v1

Addressing Complexities of Machine Learning in Big Data: Principles, Trends and Challenges from Systematical Perspectives

Qi Wang[1], *Xia Zhao**[2], *Jincai Huang*[1], *Yanghe Feng*[1], *Zhong Liu*[1], *Jiahao Su*[3], *Zhihao Luo*[1], *Guangquan Cheng*[1]

[1]Science and Technology on Information System and Engineering Laboratory, National University of Defense Technology, Changsha, China, 410073

[2] Department of Mathematics and Systems Science, College of Science, National University of Defense Technology, Changsha, Hunan, China, 410073.

[3]College of Information System and Engineering, National University of Defense Technology, Changsha, China, 410073

*Correspondence Author: zxmdi@163.com

Abstract

The concept of 'big data' has been widely discussed, and its value has been illuminated throughout a variety of domains. To quickly mine potential values and alleviate the ever-increasing volume of information, machine learning is playing an increasingly important role and faces more challenges than ever. Because few studies exist regarding how to modify machine learning techniques to accommodate big data environments, we provide a comprehensive overview of the history of the evolution of big data, the foundations of machine learning, and the bottlenecks and trends of machine learning in the big data era. More specifically, based on learning principals, we discuss regularization to enhance generalization. The challenges of quality in big data are reduced to the curse of dimensionality, class imbalances, concept drift and label noise, and the underlying reasons and mainstream methodologies to address these challenges are introduced. Learning model development has been driven by domain specifics, dataset complexities, and the presence or absence of human involvement. In this paper, we propose a robust learning paradigm by aggregating the aforementioned factors. Over the next few decades, we believe that these perspectives will lead to novel ideas and encourage more studies aimed at incorporating knowledge and establishing data-driven learning systems that involve both data quality considerations and human interactions.

Keywords: big data; machine learning; regularization; data quality; robust learning framework

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

1. Introduction

Big data, a novel concept born in the 21st century, has become a very active research domain that has stimulated revolutions in industry, education, business and other applications and serves as a catalyst for innovation[1]. One big data characteristic that significantly distinguishes it from previous data is the volume, as evidenced by the fact that the sizes and dimensions of examples in the popular machine learning repository hosted by the University of California, Irvine (UCI) are continuously increasing[2]. According to the International Data Corporation (IDC)'s Digital Universe forecasts, the data volume will rise to approximately 40 ZB by 2020[3]. As the ultimate source of knowledge[4], the potential value, patterns, trends and associations in big data have attracted more attention than ever, but some fundamental issues are still less systematically discussed. In big data, the analysis of data cannot keep up with data collection and storage. The aim of this paper is to clarify theoretical foundations in machine learning and to present recent progress with new perspectives on these issues in the era of big data. Before beginning our explorations, some essential topics are revisited in this section, including definitions of big data, applications of big data and the potential value of big data.

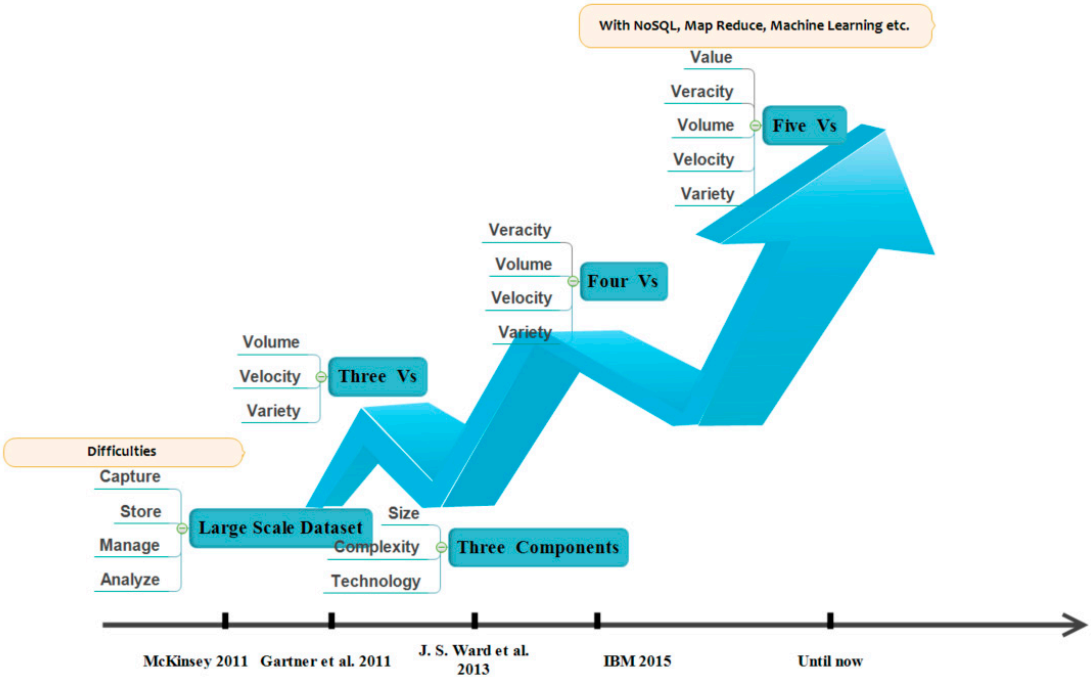


Fig1. The Evolution of Big Data in Concepts.

We first detail how the definition of big data has evolved over time just as **Fig1** shows. In 2011, the McKinsey consultancy corporation published a business report that focused on sensing and predicting both new rising areas and traditional domains of productivity, competitiveness and growth, in which the concept of big data was defined as datasets that were challenging to capture, store, manage and analyze with

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

typical or traditional software tools[5]. Gartner extracted three Vs—volume, velocity and variety—to conceptualize the aspects of big data for which cost-effective and innovative information processing techniques were required[6]. J. S. Ward et al. summarized the definitions of big data and deemed size (the volume of the dataset), complexity (the structure, behavior and permutations of the dataset) and technologies (the tools and techniques used to process the data) as three crucial components in the concept of big data[7]. However, M. Batty argued that the current obsession or puzzle of big data was not driven by the volume but rather the sources or methods of collection, which directly determine the potential value of the data[8]. Furthermore, IBM increased the 3-V concept of big data to four dimensions: volume, velocity, variety and veracity. S. Yin further expanded IBM's definition by adding "value"[9]. Finally, the term big data is denoted as the collection, storage and analysis of large-scale datasets characterized in terms of the 5 Vs with emerging techniques, including but not limited to NoSQL, MapReduce and machine learning.

Admittedly, the explosion of data originates from increasingly sophisticated sensor technology and the widespread Web 2.0[10]. The former brings more abundant informative observations while the latter allows opinion sharing and propagation online. Except for the massive increase in data, the utility of big data in applications and research is gradually being revealed in multiple domains[11, 12]. Serving government decision-making, big urban data streamed from sensors can scientifically guide resource allocation and smart city planning[8]. Business intelligence and analytics is another representative scenario. The business evolution tendency accompanies the source of the accessible dataset from the Database Management System (DBMS), Web to Mobile and Sensor-based contents. Meanwhile, analytics in various avenues encourage the boom in text analytics, web analytics, network analytics, mobile analytics, etc. Several pioneering companies have been pursuing potential interests that follow this tendency. Machine learning still leaves great room for smart manufacturing, and many companies do not realize the potential value of such generated or recorded data in process optimization or product design through predicting customers' underlying expectations and detecting failures or errors in advance[13]. With the capability to perform fine-grained analyses of increasingly available medical datasets, more accurate diagnoses methods are capturing attention from both doctors and computer scientists[14]. Undeniably, precise diagnosis systems would both serve us better and monitor our health status continually. However, these advanced data accessing techniques inspire another direction that presses us to take big data seriously. We enjoy the convenience of and harvest other economic value from big data, but simultaneously, sensors and analytic tools seem to invade our lives and privacy[5] [15, 16] [12]. Private personal information in the form of instantaneous locations, personal preferences, and recent interests are easier to access with mobile stations or service providers such as Ebay, Uber and Facebook. Even a trace click behavior on a search engine can expose people's underlying intentions; these are logged by Internet servers for further mining. Privacy research to find a compromise is ongoing. In [16], the existing privacy systems, research frameworks, several

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

privacy disciplines and mathematical representations, including challenges and opportunities with respect to privacy, were elaborated. Thus, the trade-off between two influential factors should be well managed in a valid manner, such as the design of privacy-enhancing machine learning systems[15]. Similar research issues with respect to applications of big data are uncountable. It has been suggested that big data has influenced the unlocking of other significant values, such as product development, operational efficiency, and market demand prediction[9].

Before continuing, domain researchers are cautioned that this work does not intend to provide comprehensive and complete reviews of these topics. Rather, our initial motivation is to identify the potential causes, understandable methodologies and challenges for a wider scope of readers, not just domain experts. Some of these summaries may be partial and restrictive because authors with pertinent works are unfortunately omitted.

The remainder of this paper is arranged as follows. Section 2 focuses on the nature of machine learning and describes typical learning styles, stressing the advantages of machine learning. The dominant principles of learning from naïve and statistical learning perspectives are revisited in Section 3, where relations between generalization bound, model complexity and example capacity are discussed. As the core of the research, Section 4 includes learning confusions due to data quality, reasons why data drives machine learning, tendencies for learning tasks and a proposed robust learning paradigm to overcome challenges in big data. Finally, the Conclusions section summarizes the research and guides further research in the area.

2. Machine Learning and Concerning Tasks

Abundant information is contained in large dataset collections. How to take advantage of these resources and mine their potential value is the core problem of big data research. Most of the time, it is impossible for human beings to consume and understand big data[5]. Machine learning is acknowledged as a powerful technique to achieve this goal and provides plausible methodologies to gain deep comprehension and better perception; consequently, machine learning can be viewed as a type of machine intelligence.

2.1 The Nature of Machine Learning

A recent survey regarding machine learning defined machine learning as a discipline that focuses mainly on two critical issues[15]. One is to construct computer systems that automatically evolve through the experience embedded in a collected dataset. Another is to capture the fundamental statistical computational information

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

encapsulated in the theoretical laws governing the mentioned system. Roughly speaking, machine learning involves learning from a collected dataset with the goal of addressing problems automatically. The ultimate goal of machine learning is to automate automation itself by analyzing the related data. The more data we have, the more explicit the knowledge we can mine.

Universally employed in real life, machine learning is an interdisciplinary subject aggregating optimization theory, information theory, decision theory, control theory and neuroscience. Compared with subjects such as mathematics, chemistry or physics, machine learning relies more on experience and resembles a black box. Machine learning is a powerful technique for combating the complexity of problems in real life by utilizing the power of data and directly modeling the mechanism from the input to the output[17]. The intuitive approach for learning is extremely dependent on data and involves observing possibly related attributes; the decision of what to observe is primarily based on either the researchers' experience or domain knowledge. A general manipulation is to test a set of collected attributes in experiments to assess their effectiveness. Thus, our induction for the first step in machine learning is "Just guess and try to fit it." All the hypotheses or heuristics serve the prediction, and analyzing or disentangling the relationships between factors is of great importance, as shown in **Fig 2**.

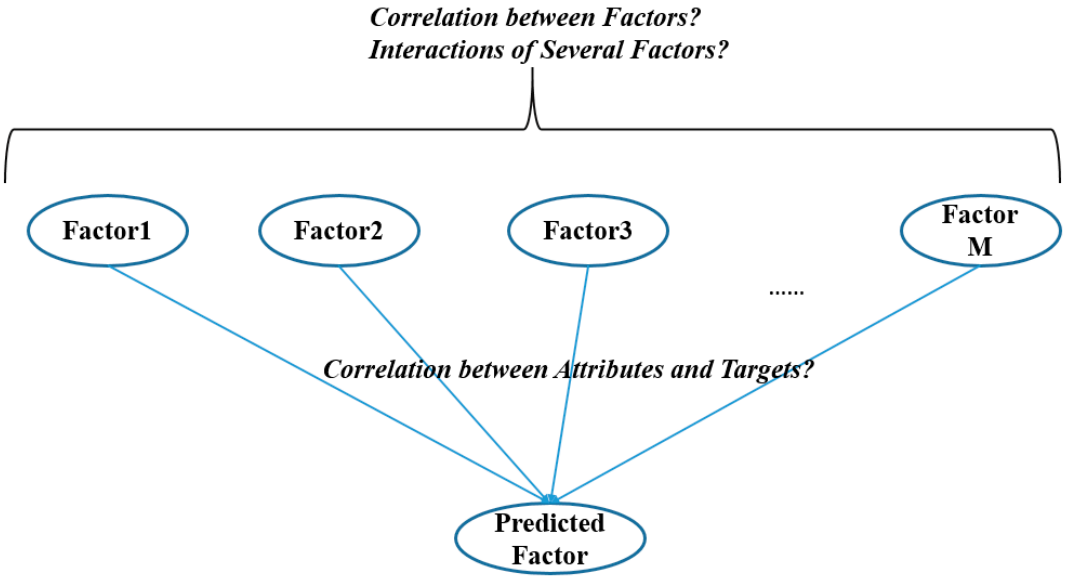


Fig 2. Disentangling Relationships between Factors. All factors can be taken as collected attributes; however, exploring the correlation relationships is difficult because of randomness or information redundancy. Determining causality is more challenging in this domain.

To more precisely describe the role played by machine learning, we review four prevailing learning styles based on the uncertainty of annotated labels in the training dataset. **Table 1** shows four styles by summarizing the forms of dataset, annotation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

information, objective in learning, as well as some typical techniques.

Table 1. Characteristics of Some Dominant Learning Styles.

	Forms of Dataset	Annotation	Objective in Learning	Typical Techniques
Supervised Learning	$\{(x_i, y_i)\|i = 1,2..,N\}$	Full Annotation	Design approximators as function f(x) or conditional distribution P(y|x)	Decision trees, Naïve Bayes, support vector machines (SVMs), artificial neural networks, etc.
Semi-supervised Learning	$\{(x_i, y_i)\|i = 1,2..,N\} \cup \{x_i\|i = 1,2..,N\}$	Partial Annotation	Design approximators as function f(x) or conditional distribution P(y|x) with exploitation of the unlabeled dataset	Transductive SVMs[18], co-training[19], label propagation[20], etc.
Unsupervised Learning	$\{x_i\|i = 1,2..,N\}$	No Annotation	Mine patterns into clusters and excavate geometric or statistical traits behind data	Clustering, dimension reduction, density estimation, visual example generation, data visualization, and manifold learning[21]
Reinforcement Learning	$\{(S_t, A_t, R_t)\|t = 1,2..,N\}$ Notation: S_t, A_t, R_t respectively represent state, action and reward at time t.	Delayed Annotation	Design policy mapping from state to action to maximize the accumulated rewards: policy evaluation and control	Dynamic programming, temporal difference[22], Monte Carlo, deep reinforcement learning[23]

Rather than being limited to these four styles, there are multiple learning categorizations due to various perspectives. From the perspective of predicting targets, the learning styles can be simplified to transductive learning and inductive learning, where the goal of the former is to predict the D_U examples, whereas the goal of the latter is to perform prediction on an uncollected dataset. **Fig 3** illustrates some learning styles.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

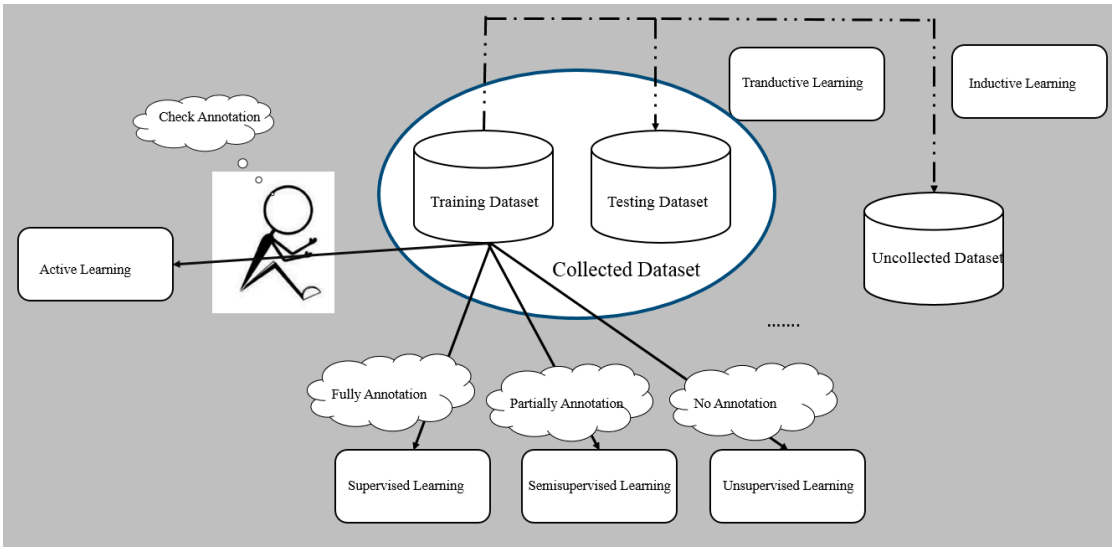


Fig 3. Some Popular Learning Styles.

Other well-known learning styles also exist, such as transfer learning (domain transfer)[24], online learning (batch of training dataset)[25], zero-shot learning (extension of label vectors)[26], and active learning (human involvement)[27].

2.2 The Need for Machine Learning

After clarifying what machine learning does, it is necessary to investigate the reasons why machine learning is placed in such a crucial position and has experienced such high activity.

In a 2017 MIT technology review, reinforcement learning, paying with your face and self-driving trucks were listed as 10 breakthrough technologies. As Will Knight commented in an MIT technology review, “Computers are figuring out how to do things that no programmer could teach them”. Ralf Herbrich, the director of machine learning at Amazon, regards machine learning as the science of efficiently automatically mining potential patterns in datasets to make predictions about future data[28]. As catalyzers in technology, machine learning techniques offer promising prospects in a variety of application domains that include health care, manufacturing, education and financial markets[15]. The need for machine learning corresponds to intensive requirements, ubiquitous application situations and attempts to alleviate workloads for human beings.

The inherent advantages of machine learning relative to human experience or traditional mechanism models are mainly reflected in three aspects.

First, machine learning is superior at performing deep searches at high speeds. A notable instance in which machine learning overwhelmed human beings’ cognition is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

the success of the AlphaGo agent in playing the game Go. The agent benefits from the trial-and-error mechanism through reinforcement learning and from Monte Carlo tree searches. Although these techniques are computationally expensive, they seem to make machines better at performing deep thinking and deriving new findings. A proof of the evolution of machine intelligence is that the state-of-art AlphaGo's records of games it plays with itself are broadening our understanding of Go even though Go has a long playing history. In other words, machines with advanced artificial intelligence can perform wider and deeper searches for solutions more swiftly, perceive the future in advance and acquire additional knowledge—even though such tasks are challenging even for human experts.

Second, machine learning is superior at addressing uncertainty. Technically speaking, machine learning allows fewer constraints to be imposed in tasks, and reveals some hypotheses as generally weaker. Mechanistic models are quite dependent on experts' cognition of specific phenomenon. For example, pollution diffusion is generally modeled statistically based on wind speed, wind direction, location of radiation sources and release rate[29]. The hypothesis restricts the universal employment of the model. In particular, large-scale datasets have properties that small datasets do not share and that only machines, not humans or limited-mechanism models, can capture well.

Third, machine learning is superior at understanding relationships and abstracting concepts from datasets. Some faint but crucial details can be detected by machines, which use statistical tests to do so conveniently. Moreover, some proposed algorithms, such as principle component analysis (PCA) and tSNE[30], provide more intuitive relationships in lower-dimensional spaces. In concept or pattern abstraction, many models have been developed that range from unsupervised learning, such as K-means, which reduces a dataset to several patterns, to supervised learning, such as a decision tree, which learns concepts via feature space partitioning and rules for pattern recognition. Although humans' abilities to abstract entities are not completely understood, machine learning techniques provide multiple detection methods.

3. Principles of Machine Learning

Inappropriate employment of machine learning would deteriorate the performance in addressing problems, and some understandings of learning can be generally condensed in the form of principles. To facilitate the proper use of machine learning, this section presents a critical overview of the principles that guide the design of learning models.

The following subsections investigate the evolution of principles and disentangle the relationships between model complexity, volume of examples and capability of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

generalization. Additionally, two topics will be discussed throughout this section: What are the dominant principles of machine learning, and how have they evolved? What is the nature of regularization, and how can it be conceived?

3.1 Occam's Razor: the Naïve Principle

The capability of generalization, which measures the performance of a trained model using a collected dataset on a future or unlabeled dataset, is of universal concern[31]. As the fundamental and most influential principle in modern science, Occam's razor, which was formulated by William Occam in the Middle Ages, has been ubiquitously recognized as a prior guideline for generalization capability analysis and model design or adaptation during knowledge discovery. The naïve principle tries to clarify the relationship between model complexity and generalization.

However, the ambiguity of the razor's descriptions and exaggerated scope of applications have incurred substantial criticism. Pedro Domingos theoretically and empirically analyzed the two widely used razors. The two typical razors after refinement are as follows[32].

First Razor: *Given two models with the same generalization error, the more comprehensible one should be preferred. (Substituting simplicity with comprehensibility, which is largely domain-dependent)*

Second Razor: *Given two models with the same generalization error, the simpler one should be preferred only when the target phenomenon is believed to be simple. (Restricting the condition to be a simple phenomenon)*

In these two razors, model interpretability and simplicity (for simple phenomena) are regarded as two crucial decisive factors in model design. Here, we propose another explanation for the preference for simple models. Reasoning by analogy is one of the prevailing schools of thought in machine learning, in which items similar in representation resemble patterns. Because the training dataset is sampled from the whole dataset, the pattern of the predicted example can be approximated using its neighborhoods in feature space. This is called the local approximation principle, and this principle has been verified for both supervised learning, such as K-nearest neighborhoods (KNN) and SVM, and for unsupervised learning, such as kernel density estimation. Moreover, the penalties on the parameters such as the L1 derivation norm (continuity) and L2 derivation norm (smoothness)[33] in the structural risk minimization (SRM) principle conditionally prove this. These penalties are regularizations, upon which we will elaborate. Obviously, the more data we have, the more accurately we can approximate. Complex models tend to fluctuate intensively around some local points, especially when the examples are noisy, as in **Fig 4**. Thus, for simple phenomena, complex models violate the local approximation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

principle; consequently, simplicity is preferred.

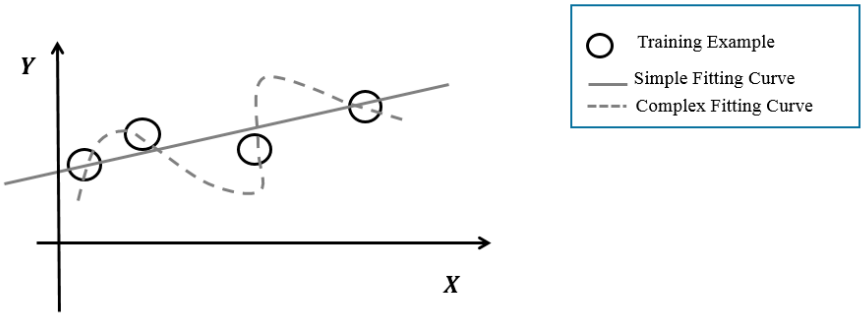


Fig 4. Curve Fitting Problem. The actual curve is a line, while a complex model captures noise.

Now a further question arises: does simplicity imply accuracy? Unfortunately, the answer is negative most of the time. For complicated datasets or sophisticated phenomena, models with greater complexity are preferable because they reduce the fitting errors. A specific example is Vapnik's SVM, in which the dimensionality of transformed features is usually high. The classification abilities of SVMs that dramatically increase dimensionality using kernel tricks seems to violate the simplicity rule even when dealing with examples of small size[34]. Further evidence lies in the widespread adoption of deep neural networks, whose interpretability is questioned by many researchers.

3.2 Risk Reduction: Dominant Principles

The razor theorems highlight the relationship between model complexity and generalization capability from a philosophical perspective. However, this perspective is insufficient to guide the learning process. Establishing these relationships in a quantitative manner would be more acceptable. Earlier, we reviewed the development of principles from statistical learning. Statistical learning focuses on the properties of machine learning from mathematical and probabilistic perspectives. Aldrich & Auret stated that the broad goal of statistical learning is to mine relations between variables[35]. Vapnik decomposed learning theory into four fundamental issues: the consistency of a learning process based on empirical risk minimization (ERM) principle, the non-asymptotic convergence rate in learning, the ability to control generalization, and the theory of algorithm construction[34]. Here, we successively review these four issues and connect them to principles.

Before analyzing the principles, we revisit the classical framework of supervised learning systems in **Fig 5**. The hypothesis space is a family of models, and the objective is to select the hypothesis that best matches the training dataset. After proper parameters have been found with an optimizer, a hypothesis is obtained in the form of a function or a conditional distribution that well explains the mechanism of prediction. A risk function is a target that the optimizer attempts to minimize, which is simply the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

principle of learning. In this sense, the learning task is mainly decomposed into three key components, models to fit the dataset, evaluations of the models, and techniques for objective optimization[31]. The process of optimization is generally transformed into minimization of a risk function.

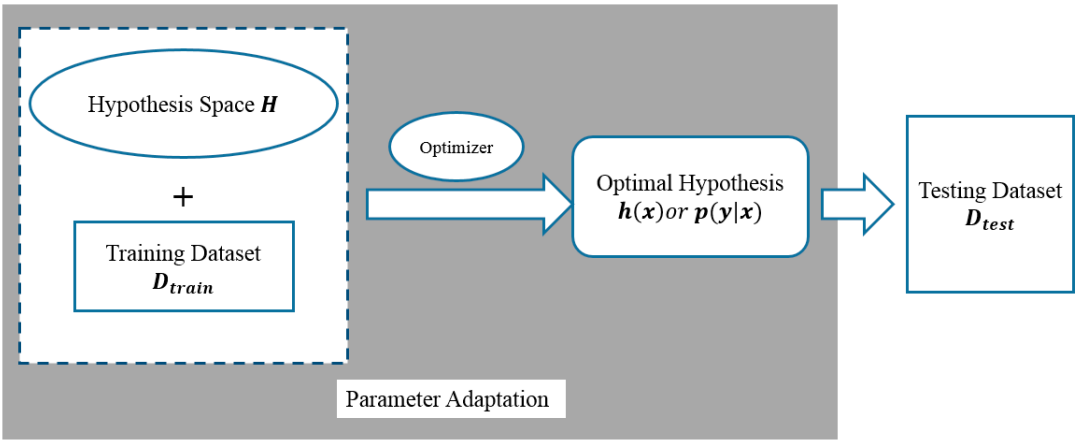


Fig 5. Supervised Learning System.

Expected Risk Function Minimization

In supervised learning, the expected risk function in learning is formulated as

$$R(\alpha) = \int L\big(y, f(x, \alpha)\big) dF(x, y) \quad (1)$$

where L measures the expected loss throughout the whole input-output probability space as $F(x, y)$. The objective is to search for the parameter α^* value that best minimizes $R(\alpha)$, thus deriving the optimal hypothesis $f(x, \alpha^*)$.
In another unified expression, the general form of learning in minimizing the expected risk function can be expressed as

$$R(\alpha) = \int Q(z, \alpha) dF(z) \quad (2)$$

where the distribution $F(z)$ defined on the space Z is generally unknown and α is the parameter to learn with the algorithm. Luckily, we have some i.i.d examples, which are denoted as $\{z_1, z_2, z_3.., z_N\}$.

Notation: We use the same expression for the expected risk function as in[34] to induce a uniform analysis for classical supervised and unsupervised learning tasks such as classification, regression and density estimation. In supervised learning especially, the space Z is equivalent to (X, Y).

Empirical Risk Function Minimization

The expected risk function, which is theoretically based on the Monte Carlo theorem and supported with the law of large numbers in functional space, can be approximated

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

by an empirical risk function constructed from limited examples:

$$\mathrm{R_{emp}}(\alpha) = \frac{1}{N}\Sigma_{i=1}^{N} Q(z_i, \alpha) \quad (3)$$

More specifically, three representative problems—pattern classification, regression and density estimation—in the empirical risk minimization framework can be detailed as follows.

Pattern classification:

$$\mathrm{Q(x, y, \alpha) = L\big(y, f(x, \alpha)\big)} = \begin{cases} 0, if\ y = f(x, \alpha) \\ 1, if\ y \neq f(x, \alpha) \end{cases} \quad (4)$$

Regression:

$$\mathrm{Q(x, y, \alpha) = L\big(y, f(x, \alpha)\big) = \big(y - f(x, \alpha)\big)^2} \quad (5)$$

Density estimation:

$$\mathrm{Q(x, \alpha) = L\big(p(x, \alpha)\big)} = -\log p(x, \alpha) \quad (6)$$

Although empirically, approximation has theoretical foundations, and minimization based on the empirical risk function indicates potential optimality, the relationship between two risk functions requires a more precise characterization.

Vapnik and Chervonenkis introduced the consistency in the probabilistic method that connects an empirical risk function to the expected risk function[36]. The consistency measures the distance between two functions probabilistically, and the empirical risk approximates the expected risk as closely as possible under some necessary conditions, with the capacity of examples increasing infinitely[34].

$$\lim_{N\to\infty} P\left\{\sup_{\alpha}\Big(R(\alpha) - R_{emp}(\alpha)\Big) > \epsilon\right\} = 0\,, \forall \epsilon > 0 \quad (7)$$

Evidently, the ERM principle is appropriate for learning tasks with large example capacity.

Structural Risk Function Minimization

Over some time, learning tasks based on the ERM principle may be ill-posed in science and engineering areas because there may be infinite hypotheses that satisfy the constraints[37] and even slight perturbations to the parameters of the model lead to very different solutions[34]. The principle of minimizing the errors on a training dataset is not self-evident, and another inductive principle and algorithms with higher generalization ability are required. The overfitting phenomenon, which describes a learning machine fitting the observations well but performing badly on unseen datasets[35], is a severe test of ERM.

Considering the loose generalization bounds derived from the ERM, the structural risk minimization (SRM) principle was developed to provide a trade-off between the ERM

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

and model complexity by adding restrictions in the hypothesis space. SRM proposes to capture a tighter confidence range in the hypothesis space, while simultaneously minimizing the empirical risk.

The structural risk function can be expressed as follows:

$$\mathrm{R}_{\mathrm{struct}}(\alpha) = \frac{1}{N}\Sigma_{i=1}^{N} Q(z_i, \alpha) + \lambda\Omega(H) \quad (8)$$

The term $\Omega(H)$ characterizes the model's complexity and can be viewed as a type of penalty, while λ is an adjustable parameter that controls the trade-off between model fitting and model complexity.
For a deeper understanding of the penalty term, we review two prevailing explanations for SRM.

- **Constraint Optimization Perspective**

When the norm of parameters acts as the measure of model complexity, the structural risk function is

$$\mathrm{R}_{\mathrm{struct}}(\alpha) = \frac{1}{N}\Sigma_{i=1}^{N} Q(z_i, \alpha) + \lambda\left|\left|w\right|\right|^2 \quad (9)$$

With the penalty in the normal form of the parameters, the minimization of the above regularized risk function [38] corresponds to the ERM under the following constraints:

$$\min \frac{1}{N}\Sigma_{i=1}^{N} Q(z_i, \alpha) \quad (10)$$

$$\text{st.} \left|\left|\mathrm{w}\right|\right|^2 < \beta$$

- **Bayesian Perspective**

Another well-acknowledged explanation for the validity of SRM comes from the Bayesian perspective.
The general form of the regularized optimization framework for the SRM is

$$\mathrm{x}^* \in \arg\min_{\mathrm{x}} f(x) + \lambda r(x) \quad (11)$$

in which $\mathrm{f(x)}$ is the traditional empirical risk function and $\mathrm{r(x)}$ is the regularizer.
The framework is equivalent to

$$\mathrm{x}^* \in \arg\max_{\mathrm{x}} \exp\big(-f(x)\big) * \exp(-\lambda r(x)) \quad (12)$$

In some cases, this is simply the Bayesian framework obtained by decomposing the objective into the likelihood and the prior. Commonly used algorithms such as ridge regression and lasso regression can be explained in this manner.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

3.3 Links between Generalization Bounds, Model Complexity and Example Capacity

Statistical learning theory attempts to bound the prediction risk function R(α) with the complexity of the hypothesis and the capacity of training examples in a probabilistic manner, which can characterize the generalization bound and guide model selection.

Given a certain probability $1-\delta$, a hypothesis space H, a function f parameterized with α^* in H and N examples for training, the ideal bound in statistical learning can be described in the empirical form

$$R(\alpha) \leq R_{emp}(f(\alpha^*)) + \Omega(H, N, \delta) \quad (13)$$

In reality, the complexity is not equivalent to the number of parameters in the hypothesis. However, what metrics can measure the complexity of a hypothesis well? This is a challenging but valuable research topic, and several researchers have proposed measurement methods, among which the Vapnik-Chevornenkis (VC) dimension and Rademacher complexity[39] are the most widely accepted.

In the binary case, the VC generalization-bound theorem is as follows[40]:
For any tolerance $\delta > 0$, example capacity N and VC dimension d_{vc}, the following probability inequality holds

$$P\left[R(\alpha) \leq R_{emp}(f(\alpha^*)) + \left(\frac{8}{N}\left(d_{vc}\ln\frac{2eN}{d_{vc}} + \ln\frac{4}{\delta}\right)\right)^{\frac{1}{2}}\right] \geq 1-\delta \quad (14)$$

Briefly, the VC dimension d_{vc} is defined in terms of the classification function to shatter a specific number of data points. The greater the number of data points with random labels that can be shattered by the function in the hypothesis space, the greater the complexity of the hypothesis.

According to the empirical relationships between the generalization bound, model complexity and example capacity, we can deduce both that efficient regularization embeds deep understanding about the task into the model as prior knowledge and that more training examples contribute to promoting generalization. The regularization is engineering-specific, and its capabilities are attracting great attention owing to the possibility of generalization.

Although the concept of big data is currently a focus of much attention, learning from small examples while domain knowledge is limited or difficult to abstract as regularizations is also a difficult problem; one challenging example is classifying medical image.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

When faced with such a problem, adaptation to the training sample is in great demands. Here, we emphasize two main plausible approaches, virtual example generation and cross domain transfer technique shown in **Fig 6**.

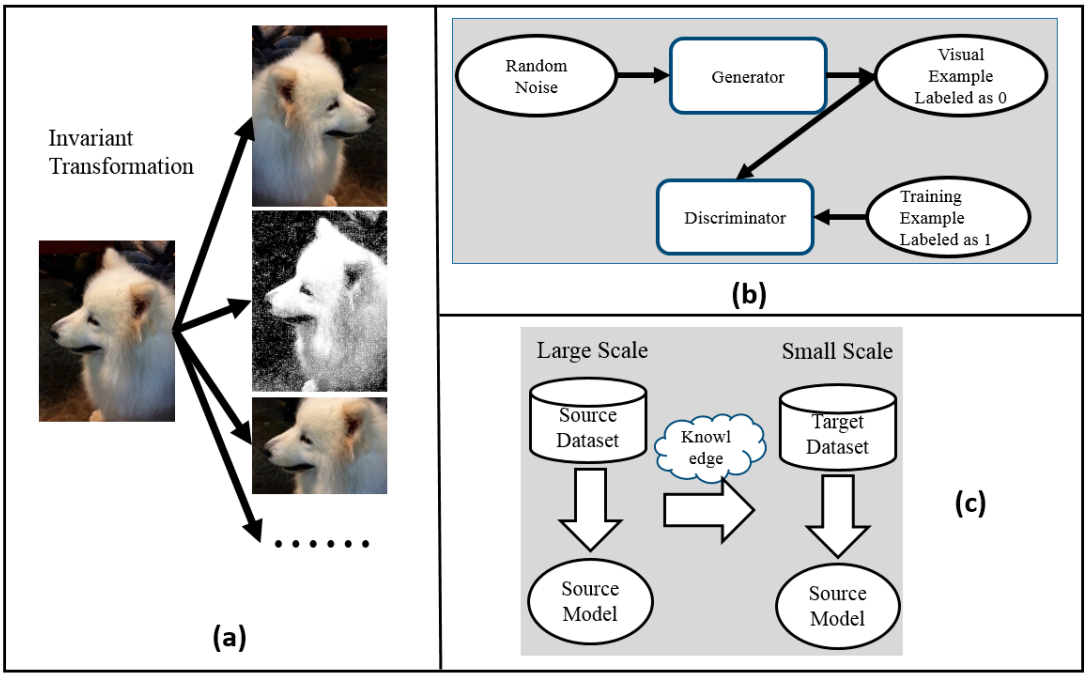


Fig 6. Adaptation to Examples. (a) is the label-invariant transformation on a dog, (b) characterizes the GAN and (c) is the mechanism of knowledge transfer between examples.

- **Virtual Example Generation**

A classical method for virtual example generation is to find invariant transformations[41]. More specifically, with transformation T, an input-output pair is mapped as follows:

$$\big(\mathrm{x}, \mathrm{f}(\mathrm{x})\big) \rightarrow (Tx, y_T(f(x))) \quad (15)$$

$$y_T\big(f(x)\big) = f(x)$$

That is, the invariant transformation performs identity mapping on the label, whereas the transformed example is virtually generated. Some concrete transformations are distortion, flip, scale, rotation or marginal subsampling of the images. These manipulations are detected as invariant transformations, and intra-class invariance can be captured from a merged dataset[41, 42]. Similar methods have been developed for tasks such as speech recognition[41], text classification[43], and scheduling knowledge creation in manufacturing[44]. However, finding such invariant transformations is non-trivial and quite heuristic.

Another technique for visual example generation that has received much attention is a network-based method called generative adversarial networks (GANs). The GAN

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

method is a promising approach proposed by Ian Goodfellow[45] in which two components, a generator G and a discriminator D, constitute the GAN. Both G and D are formulated as artificial neural networks. The generator tries to capture the training data distribution by generating virtual examples from prior noise, whereas the discriminator attempts to distinguish the original examples from the generated ones. The prior noise distribution is updated during the process of back-propagation. The two networks are simultaneously trained in an adversarial manner, essentially via a two-player min-max game. The objective of the game is

$$\min_{G} \max_{D} E_{x\sim p_{data}(x)}\, logD(x) + E_{z\sim p_z(z)} \log(1 - G(z)) \quad (16)$$

When the discriminator cannot distinguish the virtual examples from the training examples, the training process is complete. The ultimate generator corresponds to a probability distribution p_g, which well approximates the training data distribution p_{data} with the existence of global optimal solution in optimizing objective (16) for $p_g = p_{data}$, as proved in[45, 46].

- **Cross-domain Transfer**

In practice, cross-domain transfer techniques do not generate any novel examples; instead, they borrow knowledge from related domains. More precisely, the domain of interest is referred to as the target domain, whereas the auxiliary domain is called the source domain. The relationship between the two domains directly determines whether the transfer technique contributes to our target tasks[47]. Transfer learning is a complicated and systematic framework, and from the perspective of knowledge type to transfer, it includes instances of knowledge-based methods, feature representation knowledge-based methods, parameter knowledge-based methods and relational knowledge-based methods[24]. For target domains with limited example capacities, feature representation seems critical. Ideally, the target domain dataset can well inherit the representational structure transferred from the model trained on the source domain dataset[48, 49].

In some sense, the incorporation of additional examples is another method to exploit domain knowledge. The invariant transformation method partially relies on domain knowledge, and the GAN takes advantage only of the potential to discriminate between source examples and randomly generated ones, while domain transfer utilizes related explainable factors to discover knowledge. Enlarging the capacity of learning examples or the utility of the representation structure by transfer leaves data to speak for itself and imposes more constraints to fit; thus, adaptation of an example is another type of regularization.

However, regularization is not a well-defined concept, and it is impossible to describe it from a unique perspective. In addition to the aforementioned explanations, there are other regularization styles that reduce overfitting, such as minimum-length

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

description[50], pruning criteria[51], and surrogate cost functions[52].

Moreover, there is always confusion regarding whether the optimized solution is in fact the optimal solution. The answer is conditional, and as the complexity of the model increases, the fitting error on the training dataset usually decreases, whereas the risk of overfitting increases. Sometimes, the adoption of an early termination mechanism during the optimization process can improve a model's performance[53]. The consensus among these regularizations is the control of hypothesis complexity from multiple views (See **Fig 7**).

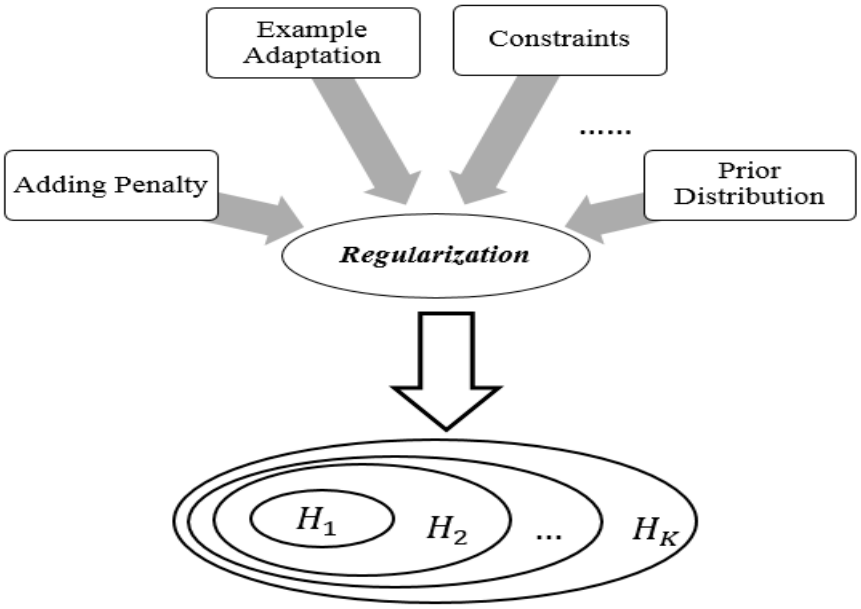


Fig 7. Some representative regularizations. Most of these regularizations can be regarded as constraints imposed in hypothesis space. More constraints mean a narrow but tight hypothesis space

4. Some Research Trends and Challenges

With the enormous influx of attention to the evolution of data and processing techniques, researchers have performed several surveys of the tendencies and crucial issues of machine learning, data mining or artificial intelligence in the era of big data from various perspectives. S. Sagiroglu et al. summarized the preliminary process in addressing big data in terms of three significant properties, variety, volume and velocity, and discussed the importance of security in big data[54]. In [55], machine learning in big data is described as follows: "Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among the data. In this scenario, the mining task is to aggregate heterogeneous information from multi-sources to best attain the aim." With the history and tendencies of artificial intelligence development

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

overviewed, Y. T. Zhuang et al. insisted that explainable, robust and general AI would dominate the future[56]. After reviewing the Bayesian methodologies employed for big data, Zhu et al. demonstrated that a Bayesian framework can well combat uncertainty from environments, advance the conceptual flexibility to embed domain knowledge and easily adapt algorithms to dynamic learning scenarios[57]. A recent survey first referred to the data quality during the learning process[58], with some content overlapping with ours.

In contrast to the above literature, we seek to explore the trends and challenges in mining process from the following multiple views:
What factors determine data quality, and how do they influence learning systems?
Why does data drive the model in machine learning, and what directs the design of algorithms?
What role can humans play in applying machine learning to big data, and how can the role of humans in advancing the learning system be optimized?
What learning framework in data mining will be most appropriate for big data in the future?
This section is organized around these four topics.

4.1 Complexities from Big Data: Data Quality

Currently, volume is no longer the bottleneck in data science, but tools for knowledge discovery are in great demand. Recalling the 5 V characteristics, the volume of data and the velocity of processing techniques have posed great challenges for data storage and processing; consequently, distributed systems have been developed as a plausible framework to address these challenges[57, 59]. Regarding the variety of big data, it is best matched through multi-modalities, and its inherent heterogeneity makes it easier to gain a global view of phenomena[60]. The property of multiple sources or cross-domains has contributed to the boom in data fusion techniques[60, 61]. However, in recent decades, veracity and value have seemed to outweigh other characteristics. In other words, the challenges posed by veracity and value press us to explore specialized learning algorithms to accommodate the new environment.

Here, we refer to veracity and value as "data quality," which is a broader contextual concept. This topic has been discussed in the contexts of geographical data analysis[62], business data analysis[3] and industrial data analysis[63]. In terms of dimensions in data quality, consumers are clearer about what data quality means. As a domain-specific concept, dimensions of quality are up to consumers and generally specific to a given environment. Examples include information retrieval value properties such as currency, availability, noise ratio, authority, popularity, cohesiveness, and objectivity. Several data quality assessments have been performed[3] [64], but few of them were inspired by machine learning phases.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

In this sub-section, research issues regarding the veracity and value of big data are discussed while considering the learning process. Similar to cognition in [64], we mainly relate the veracity of big data to the core idea of data quality, thus reducing the problem into several issues from the point of view of machine learning. Some of these are classical but still attract interest in this domain.

Curse of Dimensionality

The curse of dimensionality was introduced by Bellman and corresponds to the widely existing phenomenon that the amount of data required to describe a concept grows exponentially with increasing dimensionality in Euclidean space[65]. A direct effect of this phenomenon lies in many analogy-based algorithms, for which using similarity or proximity may not be practical in high-dimensional space[66]. The disaster resulting from high dimensionality is recognized as the second crucial issue after overfitting[17], and high-dimensional datasets are encountered in many domains[2]. Considering the utility of features, two problems arise when involving all features in learning: one is that some features may be only weakly correlated with our feature of interest, which we call non-typicality in feature space. Another is the wide existence of information redundancy between features. To mitigate the impact of high dimensionality on learning problems, feature selection and feature transformation are commonly used techniques.

Literally, feature selection involves identifying a subset of features to optimize the learning performance. Assume a feature set with n dimensions; the scale of possible subsets can reach 2^n, and the scale grows exponentially with the number of required features. Methodologies for feature selection can be mainly categorized into two categories.

The first is filtering-based methods, in which features are screened according to some information or statistical index, such as correlation, distance, or accordance[67]. A typical algorithm is mRMR[68], where the objective is to simultaneously maximize the mutual information between features and labels and alleviate redundancy between features. The second style is called the wrapper method, which directly connects the selection of the feature subset to the performance of some specific classifier, such as an SVM, decision tree, or random forest. The filtering-based approach is more universal and is independent of classifiers, whereas the wrapper-based approach tends to perform better in practice[67, 69]. Both styles rely on efficient optimization algorithms. Some commonly used algorithms are random search and greedy search (such as step backward or step forward) and evolutionary search (such as genetic algorithms or particle algorithms)[70, 71].

Feature transformation is another method of dimension reduction that works by generating latent explainable factors after performing a series of linear or nonlinear

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

transformations. Some traditional methods of feature transformation are PCA, kernel PCA[72] and low-rank approximations[73]. With the rise of representation learning in complex tasks, deep transformations are highly advocated. After a series of linear or nonlinear transformations, deep networks can better represent the original features of images, texts, and voices. Network embedding methods such as deep walk[74] and node2vec[75] have also been proposed to represent relational networks.

Bias in Sampling and Non-stationary Distributions

In real life, the data distribution is generally time-evolving and non-stationary[76, 77]. We cannot accurately predict the range of building prices in a region based on prediction models derived from historical data from ten years ago or on data collected from other regions. Similarly, estimating the numbers of sick and healthy people from hospital data based on some specific disease is biased relative to the distributions across an entire society. When playing StarCraft, the learning strategy must be altered when changes occur in the environment[76]. This is the general scenario in reinforcement learning tasks. Potential reasons include that the prediction dataset does not share the same distribution as the training dataset. Bias in sampling and non-stationary distributions are the two main causes. Using a model trained on very old traces of historical observations, only a few tight relationships between factors or stable conclusions can be maintained for the same learning tasks; thus, the data should be less heavily weighted for future predictions due to this bias. In practice, the performance of offline models tends to decrease significantly with new datasets, even those in the same domain. The drift of distributions in the domain leads to biases between the distributions and violates the core assumption of statistical learning theory—that the training and prediction datasets are sampled independently and identically.

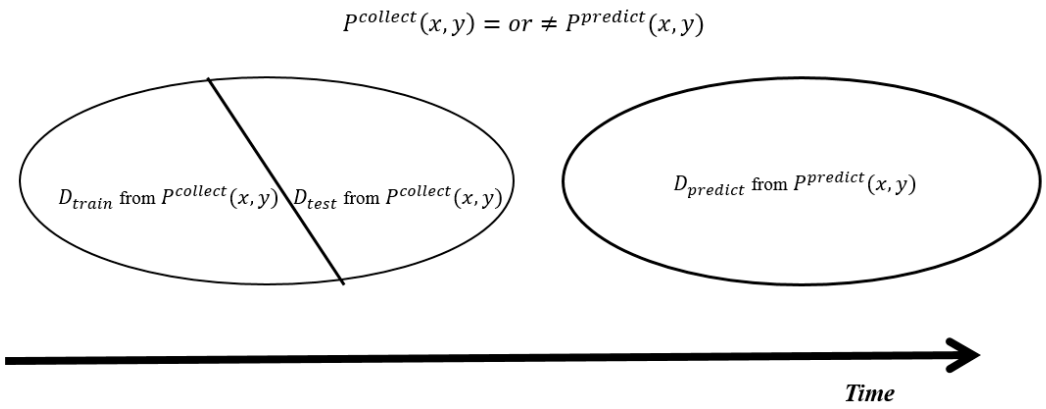


Fig 8. Bias and non-stationary puzzles in the learning process. Datasets for training and testing generally share the same distribution, while the dataset for prediction may not follow such a distribution. The distribution may also change over time

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

Because most big data in domains such as social networks, financial markets, and online commerce are generated sequentially, concept drift is an inevitable phenomenon. Concept drift means that previously learned concepts become invalid as the context changes[66]. Such drift stems from inherent data evolution over time or from external factors such as sensor violations and inexplicit turbulence; more intrinsically, variations in the distribution of evidence space $P_t(x)$, the posterior distribution $P_t(y|x)$, or both lead over time to concept drift, as in **Fig 8**. That is, the dataset is aggregated with multiple sub-datasets $\{D_1, D_2, D_3.., D_T\}$ that are distinguished from each other in terms of their trait distributions. Perfect partitioning of the whole dataset is equivalent to detecting the processes of drift, which is a great challenge.

Due to the possible violation of the traditional identical and independent distribution (i.i.d.) hypothesis, adaptation is being studied to accommodate changing environments. Here, we follow the categorization of learning algorithms for concept drift presented in [78] as two main families: active and passive. The active style explicitly involves the detection of distribution changes as the first step; subsequently, some adaptation is performed on the dataset. Drift detection is employed as the crucial step[79], and the relevant algorithms are based on statistics hypothesis tests[80, 81] and turning point identification[82, 83]. The adaptations include dataset partition methods, such as discarding outdated examples and focusing on recent examples[84] (the sliding window method), example reweighting methods, such as exploiting all available examples but assigning more weight to strongly relevant ones[85, 86], and ensemble methods such as allocating different weights to models trained on different sub-dataset periods[83]. The passive style of learning in non-stationary environments is to directly learn an evolving system that is just-in-time adaptive to the presence of changes in the distribution. These adaptive models include both single models[87-89] and ensemble models[90, 91], but the latter tend to behave more stably and exhibit superior performance.

Imbalanced Data Learning

Imbalanced learning has been recognized as one of 10 grand challenge open problems in data mining[92]. Class imbalance, which means that the class distribution behaves in an extremely imbalanced manner[93], is not a special problem encountered in the era of big data but rather runs throughout the entire history of data analysis. In truth, imbalance appears in essentially any classification task. However, when the instances of various classes behave in an extremely imbalanced manner with regard to their distribution, the learning task tends to be complicated due to the sub-optimal results derived by most traditional algorithms. One of the effects of learning directly from imbalanced data is a lower recall ratio for the minority class(es), which can then fail to satisfy the requirements for specific tasks, such as medical diagnosis[94] and financial fraud detection.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

The phenomenon can be explained in terms of two aspects. One is overutilization of the prior information on the collected dataset. Bayesian classifiers illustrate this aspect sufficiently. Another is more geometrical and intuitive: examples of the minority near the decision boundary tend to be considered as noise, especially when accuracy is employed as the objective. An SVM built from an imbalanced dataset demonstrates the second aspect in **Fig 9**.

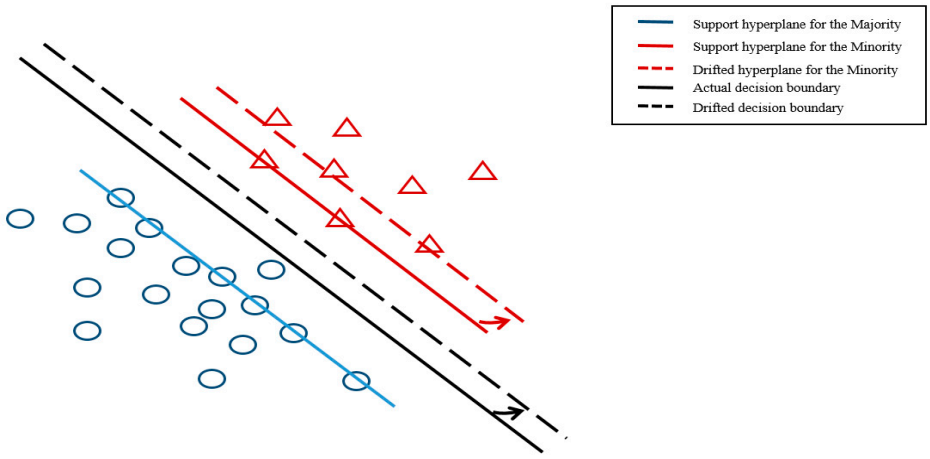


Fig 9. Boundary drift for an SVM on an imbalanced dataset. The support hyperplane for the minority tends to drift towards the dense domain of the minority, and thus, the decision boundary is pushed towards the minority

To characterize the extent of class imbalance, we employ the imbalance measure index proposed in [94]:

$$\mathrm{ImRat} = \frac{\mathrm{N_c}-1}{N_c}\frac{\Sigma_{i=1}^{N_c} I_i}{I_n - I_i} \quad (17)$$

Obviously, the range of the imbalance ratio is $[0, \infty)$.

To address the class imbalance issue, several algorithms have been developed that can be categorized into three families. Sampling methods, as the first family, are quite dominant. Oversampling of the minority or undersampling of the majority at some ratio can relieve the effects of imbalance to some extent[95]. Another important technique called the synthetic minority oversampling technique (SMOTE) creates synthetic examples by interpolating between one minority and its neighbors[96]. Some variants of SMOTE take the boundary information into consideration, including borderline SMOTE[97] and extrapolation SMOTE[93]. Notably, the performance of these algorithms is strongly related to sampling ratios. The second family of algorithms is based on the idea of cost-sensitivity. By adjusting the weights of penalties on different misclassification errors, an adapted classifier can improve upon suboptimal results. Some representative algorithms are Ada-cost[98] and adaptive boosting. The third family comprises hybrid approaches that combine sampling techniques, ensemble tricks, boundary information, clustering, and so forth to enhance performance. Two typical superior algorithms in this family are EasyEnsemble and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

BalanceCascade[99].

In essence, learning from imbalanced data is a problem of multi-objective optimization, where the recall ratio of the minority and global accuracy contradict each other.

Noisy Label Learning

When a dataset is small, examples are generally annotated with high confidence because the workload or cost involved in labeling is acceptable. However, such high labeling confidence is no longer possible when the data volume increases rapidly. The labeling process in the era of big data is fundamental in machine learning—but it is also unimaginably expensive, labor-intensive and time-consuming. To date, only a few datasets with perfect annotations exist that are well managed by science and technological enterprises or research organizations: these include MSCOCO by Microsoft [100], Flicker by Yahoo, ImageNet by Stanford[101], and MNIST by Yann LeCun's research team[102]. As a technique for reducing the cost of labeling, crowdsourcing utilizes spare online resources and accelerates labeling engineering[103, 104]. Crowdsourcing enjoys popularity in multiple domains, such as item ratings in recommendation system, image annotation, and natural language processing. It has been noted that the labeling error rate can be reduced exponentially as the number of workers assigned to each task increases[105]. However, considering the resource restrictions, the nature of humans, and differences in cognition, the labeling process is not error-free even for domain experts[106, 107] [63]. A semi-supervised learning framework is a brilliant idea for assisting label annotation but suffers from the issue of label reliability.

In summary, the noise in labels is ubiquitous and is currently becoming a reality (or will in the future). Compared with noisy features, the reliability of labels appears to be more crucial and more strongly related to the performance[108]. The partition or branch criteria in decision trees, such as mutual information or the Gini index, involve labels. The decision hyperplane is also constructed with the label information in SVM. Obviously, the label is directly or indirectly incorporated into the objective in supervised learning. Importantly, noisy labels may cause more extensive generalization deterioration[109].

To simplify the problem, noise added to labels can generally be reduced into three styles of probabilistic relationships, as depicted in **Fig 10.**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

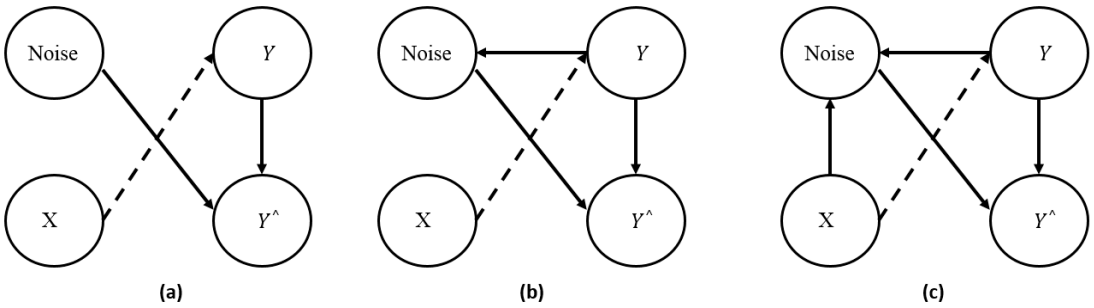


Fig 10. Three dominant noise types on classes. The solid line indicates probability-dependent relationships.

Based on these facts, weakly labeled examples impact our learning tasks and inspire modifications of learning in the presence of label noise. There are two main algorithm types for learning under noisy label conditions. One is to filter the noisy label examples to either remove or correct them[109]. Multiple techniques have been employed to detect noisy labels, such as large margin classifiers[110], nearest neighborhood verification[111], committee voting[112], cross validation[113], and clustering algorithms[114, 115]. Another type learns directly from the weakly labeled dataset by considering the mechanism of label noise. Using a probabilistic model for embedded label noise, algorithms such as robust logistic regression[116] [117] have been shown to be more robust to such noise than the originals.

Among these algorithms, surrogate loss functions are widely studied[118] [119]. The core idea of these algorithms is to revise the bias in the risk function using importance reweighting

$$\mathrm{E}_{(\mathrm{x,y})\sim\mathrm{D}} l(h(x),y) = E_{(x,y^{\wedge})\sim D^{\wedge}} \left(\frac{P(x,y)}{P(x,y^{\wedge})}\right) * l(h(x),y^{\wedge}) \quad (18)$$

The optimality and convergence of reweighting have been demonstrated in [118] [119].

The critical difficulty in this approach lies in estimating the hidden parameters in the probabilistic noise model or the revision coefficients in the objective function.

In summary, research regarding models robust to noise is becoming mainstream, because such models are necessary when the annotation quality cannot be ensured. It should be noted that the class of conditional noise represents only a minority of label noise cases; other types of label noise, such as feature-dependent noise, still leave much room for research.

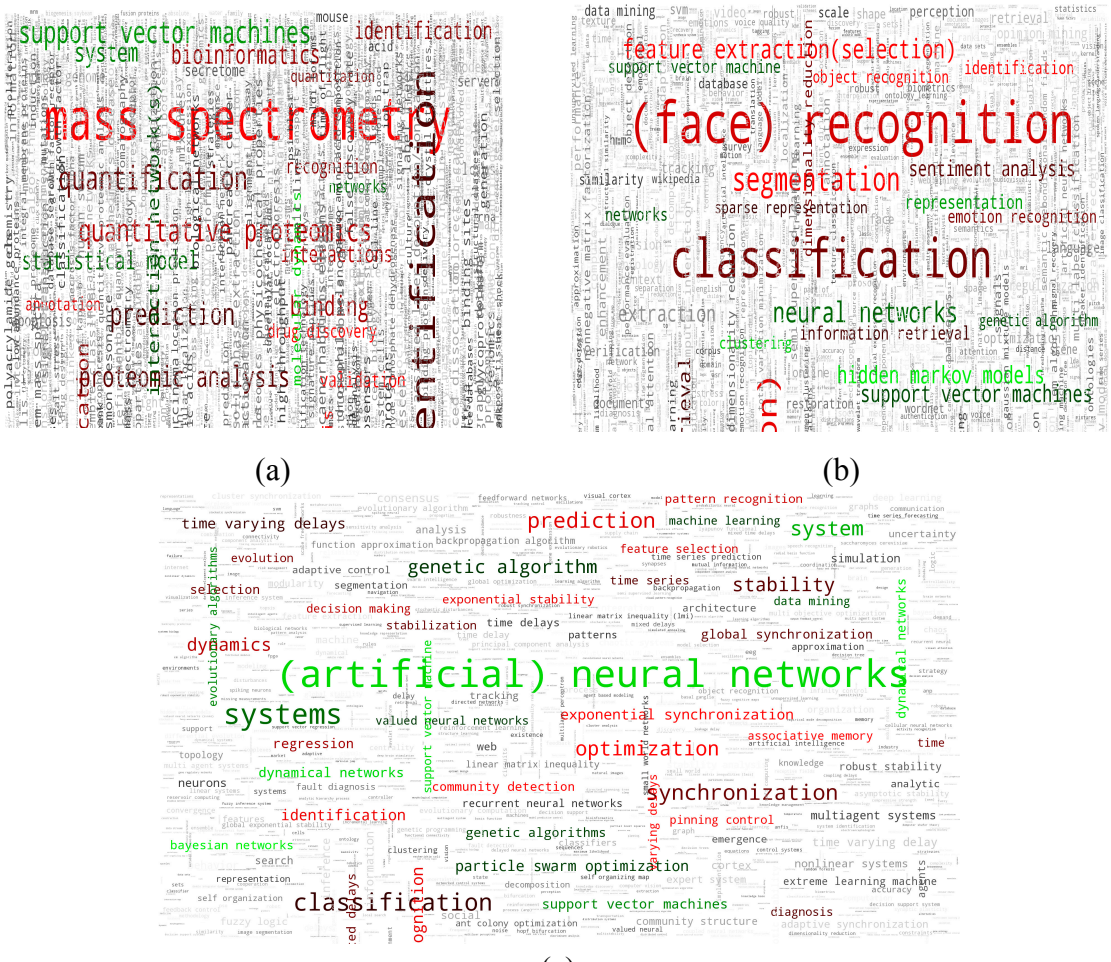
4.2 Data-Driven Machine Learning

The core claim in this sub-section is that scientific programs are driven by the available data[56]. The “no free lunch theorem” posits that any optimal algorithm selected from some hypothesis space and applied to an arbitrary domain would be beaten by some random algorithm[120]. No universal framework or algorithm exists

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

that can address arbitrary learning tasks. Algorithms applied or designed for tasks depend on the characteristics of datasets and specific domains. The superiority of specific algorithms over others is generally narrow and restricted to certain datasets.

Notably, it is impossible to display the development of machine learning covering universal applications. Here, we use the biomedical dataset, multi-media dataset, and social network dataset to illustrate this. We collect key words of papers Science Citation and Social Science Citation Indexed in Web of Science during the period from 2009.01 to 2017.09 and part of top cited papers, which can well illuminate the hotly debated directions, are selected for analysis. The categories in Web of Science are respectively restricted as Computer Science Artificial Intelligence and Mathematical & Computational Biology to match the retrieval words. **Fig 11** illustrates word clouds of the representative retrieval words in three topics. **Fig 11(a)** and **Fig11 (c)** are respectively analyzed from 2000 top cited papers while **Fig 11(b)** are analyzed from 6000 top cited papers.



(a) (b) (c)

Fig11. Word Clouds on three Topics. The red items are crucial related concepts and tasks while the blue ones are techniques. The size of word corresponds frequency. (a) is ‘protein’ topic, (b) is the ‘image’, ‘text’, ‘speech’ aggregated topic and (c) is ‘complex networks’ topic.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

Furthermore, we respectively analyze the **Fig 11** theoretically.

Data-driven in Bioinformatics

Due to the high-throughput techniques widely employed in measuring physical or chemical properties, datasets in bioinformatics research have properties such as high dimensionality, multiple classes, noisy data and missing values[121]. Quantitative methods and machine learning techniques are popular in analysis which are highlighted in **Fig 11(a)**. Consequently, Bayesian inference is the most widely used tool for disentangling relations[122]. Whereas complex network-based methods, such as protein-protein interaction networks, are also widely constructed from the interactions between examples[123]. This style of research is just from the system perspective.

Data-driven in Multi-media

Multi-media datasets are generally easier for humans to recognize; however, discriminative features are difficult for machines to encode. The complexity of datasets, such as in images, voices and texts, causes confusion among researchers regarding how to encode the preliminary forms, such as pixels, signals, and words, into representations at a higher level. The necessity of extracting efficient features highlighted in **Fig 11(b)** has impelled the rise of deep learning. A variety of competitive variants in deep models have been developed and have frequently occurred in competitions on large-scale datasets such as ImageNet. Another highlight in **Fig 11(b)** is the word 'hidden Markov models'. Taking the speech or text learning tasks into considerations, probabilistic models can well encode the context relationships.

Data-driven in Complex Network

Datasets in large-scale social network research involve properties such as complex links between examples, incomplete information for node attributes, heterogeneity in types of nodes or links, and evolution in topology with time, all of which are reflected in **Fig 11(c)**. Specifically, the links between examples make it difficult for direct supervised learning employment. Instead, some indexes concerning nodes or paths are proposed for tasks[124]. A recent heated topic related to complex networks involves network embedding to inexplicitly encode the potential links, paths or other topological properties in new representations for each node. Thus, the rise of deep learning has enlightened the representation of complex networks.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

4.3 Trends in Algorithm Development

This sub-section presents algorithms that are driven by specific datasets and concrete tasks in the domain. Pedro Domingos included a distinguished induction in his book The Master Algorithm[125] as "Knowledge ranging from the former, nowadays to the future can be mined by some unique algorithm called ultimate algorithm". In the book, the existence of such algorithms is supported by evidence from neuroscience, evolution, physics, statistics, and computer science. Nevertheless, the ultimate algorithm may not be represented in an analytical manner. In the following subsections, we discuss traits on some classical or recently developed algorithms, some of which were previously or are now thought to be the ultimate algorithm. These algorithms act as a magic mirror to illuminate directions in algorithm research and reflect trends as well as challenges. They would also highlight the role of humans in learning.

Deep Models, Short Models and Ensemble Models

In practice, most computational effort is expended on feature engineering, which directly determines the final performance. **Table 2** illustrates three hotly debated models in brief with respect to corresponding feature representations and typical variants.

Table 2. Brief Summary of three Models.

Model	Representation Power	Typical Extensions
Deep Neural Networks	Linear or non-linear transformation with activations in multiple layers	Variants such as SAE[126], DSAE[127], DBN[128], CNN[129], LSTM[130], etc.
Support Vector Machine	Kernel tricks in the forms of Gaussian, radial basis function, polynomial, etc.	Variants such as SVDD[131], OC-SVM, twin SVM[132], SVR[133], KRR[134], etc.
Extreme Learning Machine	Random projection and nonlinear activations such as Gaussian, sigmoid, tanh, multi-quadric, etc.	Variants such as deep ELM[135], incremental ELM[136], cost-sensitive ELM[137], etc.

Furthermore, we detail the mechanisms and advantages of related models with regard to challenges in big data.

Deep neural networks are naturally able to perform complex composite transformations and are capable of capturing complex dataset structures hierarchically. Theoretically speaking, deep neural networks have at least two advantages: one is their end-to-end framework, which connects what we collect to what we would like to predict. Thus, their structural design is directly related to their performance, which is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

a challenging open problem. Combining such a framework with reinforcement learning has proved to be successful in some types of control tasks, such as autonomous vehicles and game play. However, a reality contradiction between deep neural networks' wide employment and big data involves noise-induced weak labels, and some researchers have explored robust versions of deep neural networks on images[138-140]. Another issue is the flexibility of these algorithms, which enables them to accommodate more complex datasets or approximate any functions by broadening or deepening the network. Challenges such as local optimality or gradient vanishing also accompany this advantage, and regularizations similar to early stopping, weight decay, dropout[141] and batch normalization[142] remain to be discovered. In addition to networks, deep structures have recently been extended to ensembles of decision trees, such as gcForest[143] and related variants[144, 145]. It can be asserted that deep learning aggregated with other learning frameworks will be promising, and deeper models not limited to networks will emerge in the coming decades to satisfy the need to fit complex datasets.

Undeniably, the risk of overfitting with increasing model complexity introduces another concern regarding deep learning, namely, the restriction of its scope of application. For some problems for which raw features are not allowed, deep learning tends to perform poorly, as exemplified by fraud detection in credit card transactions, which has been represented using high-level features[146].

Compared with deep models, shallow ones are less difficult to optimize, which reduces the computational expense. Here, we analyze the classical SVM and the recently developed extreme learning machine (ELM).

SVMs, which were originally proposed by Vapnik, have long been regarded as the state of the art and still play a crucial role. The objective of an SVM is to construct a decision hyperplane that can separate two classes of examples, as with the perceptron. For a specific dataset, the choice of the proper kernel is challenging, and the kernel directly relates to the final performance in SVM. As a widely used convex model, SVMs can be solved in a dual manner, and they also have the property of non-local optimality. For datasets with sparse representations, swift training methods that operate in linear time have also been formulated[147]. However, when extended to a large-scale dataset, more efficient optimization algorithms are needed and traditional algorithms tend to converge slowly and require more memory.

ELM, introduced by Huang et al. [148], is another shallow model that involves a single-layer feed-forward network with universal approximation capability[149] [150]. After random projection on the input, least-squares minimization is employed to perform problem-solving between hidden and output layers, which saves time in engineering. The mechanism of random projection also raises questions regarding this model. One of the most important challenges is in respect to the stability of the model because random projections vary substantially in different turns. Another challenge is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

in regard to the pruning criteria. Though numbers of hidden neurons empower compact distributed representations, the selection of efficient projections is a method to combat overfitting. Some heuristics on pruning are based on measures of an information index[151] or penalizations for the norm of parameters[152].

In addition to these three models, the family of ensemble models has always been of substantial importance throughout the development of machine learning.
A common phenomenon in learning is that the hypothesis space may not be suitable for model selection or that a proper hypothesis is difficult to assume because there are infinite cases for data generation. Ensemble learning is a traditional framework for enhancing the generalization of a learning system by aggregating base learners. The advantages of ensemble models are as follows: they avoid optimization difficulties (weaker learners are easy to train with fewer examples) and they adapt better to approximate the true hypothesis even when the hypothesis space is biased relative to the true hypothesis (the collected examples provide only partial information for the actual data, and there may be no single, consistent model to fit)[153, 154]. The classical strategies of model ensembles include bagging[155], boosting[156] and error-correcting output coding[157]. The essence of the ensemble, which is to create and utilize a diversity of base models. And diversities arise from three main aspects: a diversity in examples, which means performing different sampling rules on the entire dataset or adapting distributions to obtain a subset of data for base model training; a diversity in attributes, which means performing random subspace selections on all attributes to obtain a dataset in different descriptive views for base model learning; and a diversity in model mechanism, which means performing different hypotheses on the dataset for base model learning.

Big data makes ensemble learning more appropriate, because multi-source datasets and abundant attributes are available. Because they fundamentally embody many advantages, ensemble learning techniques have become irresistibly fascinating in big data competitions[158]. For example, XGBoost, a variant of the boosting decision tree, is an increasingly popular ensemble model in Kaggle[159]. It can be predicted that except for feature engineering, a rather large proportion of the workloads in data science will eventually be addressed using ensemble learning.

The Role of Human Beings in Knowledge Discovery

Some well-acknowledged perspectives on machine learning have given priority to automating machine intelligence and have simultaneously attempted to reduce human involvement by shrinking the need for experts to the maximum extent possible[160]. Moreover, some sponsored competitions involve completely automatic learning, such as the ChaLearn Automatic Challenge[161]. A recent state-of-the-art robust automated machine learning system is an extension of Auto-WEKA, called Auto-sklearn[160], in which Bayesian optimization and ensemble learning are included.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

However, some problems that are difficult to address, the style of learning that incorporates manual work promotes efficiency and accuracy, especially for some domain-dependent tasks. In other words, although automatic machine learning toolkits such as WEKA, Java-ML, and SPSS are attempting to search for the best model by including data filtering, model selection, and performance evaluation, the lack of domain knowledge in such approaches inevitably leads to some ridiculous conclusions. In 1999, T. M. Mitchell characterized the drawback of first-generation machine learning techniques for big data as being overly focused on automation and neglecting guidance from users[162]. A question arises regarding how the role of human beings in machine learning can be optimized to plug the gap between machines and humans.

Human beings, and especially experts, have accumulated massive experience when several trials have been made towards definite tasks. The embedding of the domain knowledge or prior information into the intelligence system would further advance the capability of generalization.

Several years ago, one proposed point of view was that computational frameworks inspired by machine learning would model intuitive theories well and improve the theory-building process for humans, in addition to equipping machines with human-like capabilities[163]. Of course, this declaration would lead to the evolution of intelligent techniques together with the learning abilities of human beings. In this paper, we emphasize two learning frameworks, interactive learning and active learning, from two perspectives: human-machine interactions during performance evaluation and dataset reinforcement with manual feedback.

- Interactive machine learning

Often, domain experts lack proficiency in machine learning, whereas data analysts are uncertain of their findings related to unfamiliar domains[164]. To involve human feedback into systems, [164] proposed a method to visualize humans' interactions with machine learning methods, in which updates during iterative learning are human-centered, and domain knowledge is used to assist in execution and evaluation. A more recent focus on interaction between humans and machines is the human-centered visualization technique[165, 166].

- Active learning

Active learning is a typical learning method with human involvement that iteratively learns a model by employing an increasing number of labeled examples. Specifically, the process is made up of two steps: Firstly, the current learning system generates a query recording informative data from the unlabeled data for manual annotation. After new labeled examples are added to the learning system, the model is updated. This mechanism reduces the costs involved in selecting informative examples for training, but the usefulness of the unlabeled data depends on the specific circumstances under which the active learning is incorporated, ranging from semi-supervised cases[167]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

and reinforcement cases[168] to multi-label cases[169].

There are certainly other methods to apply human involvement. The proper integration of the two learning styles with human involvement can help humans accommodate specific tasks and enhance the machines' generalization. The aim of this sub-section is to emphasize the importance of humans in the era of big data.

4.4 Robust Machine Learning Paradigms

A broad consensus is that artificial intelligence (AI) is increasingly impacting numerous aspects of society including economics, culture, law and politics, yet its potential remains to be fully investigated. Taking economic value and safety into considerations, the famous AI expert Russell advocated that robust and beneficial AI should be considered as research priorities to assist humanity in making decisions, planning and making inferences, especially in the expanding context of autonomous systems performing functions such as self-driving, acting as weapons, and trading[170]. This assertion should be adopted for machine learning as well, which is viewed as a vital approach to AI[171].

A measure of the extent of veracity or reliability of the derived model is difficult to develop because machine learning is often a black box, and there will always be some instances that do not seem to follow from the captured rules. Ditterich believed that robust machine learning would tolerate slight violations of the assumptions behind machine learning models and lead to better performance. To accommodate learning systems in big data, we propose a schematic for a paradigm of robust machine learning, as in **Fig 12**.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2017 doi:10.20944/preprints201710.0076.v1

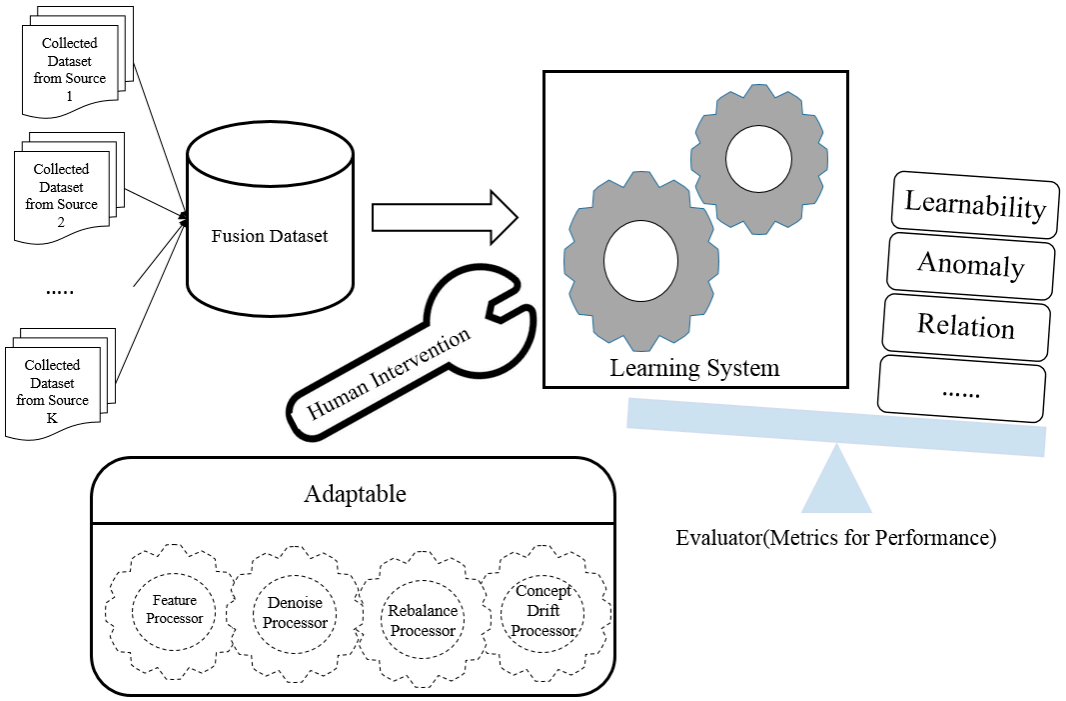


Fig12. Robust Learning Framework.

We insist that the data quality management cannot be isolated from a specific learning task, and embedding the aforementioned issues into the learning phases would result in better generalization capability. The adaptable part includes a feature processor, denoise processor, rebalance processor and concept drift processor; 'adaptable' means that the processor may take part in the preprocessing or training process. When two or more factors simultaneously occur in learning, the problem would be non-trivial. The evaluation process includes learnability analysis, the uncovering of relations, anomaly detection, etc. Additionally, a robust learning system should be hierarchical, interactive and multi-objective[172] [173].

The spanner in **Fig 12** depicts interactions between experts and machines, and improvements stem not only from expert experience or knowledge but also from new findings by the machine system. Such a view differs considerably from traditional view because it takes machine intelligence into consideration. Machines tend to capture tiny but crucial factors in concrete tasks. Some successful examples are in computer vision and strategy games. Agents currently tend to learn from self-played games when exploring more complex ideas for games. This is exactly what is most difficult for humans to understand or extract. However, human beings would be cleverer or would increase their experience by monitoring or playing with those systems. Ideally, the expert experience would co-evolve with the intelligent systems.

In establishing a robust learning system, other factors should be taken into consideration. The parallelization of an algorithm, which is important in efficiency, is related to factors such as the style of data generation, data structure in the database

 doi:10.20944/preprints201710.0076.v1

and algorithm adaptation ability. This is challenging in learning system design but essential for big data. In the layer optimization objective design, machine learning is generally regarded as a multi-objective optimization problem[172]; however, most algorithms optimize a scalarized objective in a manner similar to a regularized risk function[173]. Thus, a multi-objective optimization framework can also be involved in our system with Pareto frontiers provided.

5. Conclusions

In this paper, we reviewed issues regarding developments in machine learning relevant to big data analysis. Some learning principles were revisited theoretically to establish relationships between generalization capability, model complexity and example capacity. Analysis of regularization was pervasive throughout the principles. Any dataset may suffer some of the four dimensions of quality problems mentioned in Section 4. We also established a novel robust data-driven learning framework that was believed to be robust for big data.

One avenue for future research is to assess quality based on well-designed indicators that correspond to the curse of dimensionality, biased and nonstationary distributions, imbalanced data, and label noise as well as other issues. Moreover, consensus regarding different evaluation metrics should be investigated.

Acknowledgements

The work described in this paper is supported by the National Natural Science Foundation of China under Grant no. 71701205. Thanks to Dr. Sihang Qiu for encouragement and inspiration during this research. The author(s) declare(s) that there is no conflict of interest regarding the publication of this article.

Reference

[1] W. Pan, Q. Yang, C. Aggarwal, and C. Koch, “Big Data,” *IEEE Intelligent Systems,* vol. 32, no. 2, pp. 7-8, 2017.

[2] J. Li, and H. Liu, “Challenges of Feature Selection for Big Data Analytics,” *IEEE Intelligent Systems,* vol. 32, no. 2, pp. 9-15, 2017.

[3] L. Cai, and Y. Zhu, “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era,”

 doi:10.20944/preprints201710.0076.v1

Data Science Journal, vol. 14, no. 1, pp. 21-3, 2015.

[4] I. Taleb, R. Dssouli, and M. A. Serhani, "Big Data Pre-processing: A Quality Framework." pp. 191-198.

[5] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big Data: The Next Frontier For Innovation, Competition, And Productivity," *Analytics*, 2011.

[6] M. Beyer, "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data," *Unn | United News Network Gmbh*, 2011.

[7] J. S. Ward, and A. Barker, "Undefined By Data: A Survey of Big Data Definitions," *Computer Science*, 2013.

[8] M. Batty, "Big data, smart cities and city planning," *Dialogues in Human Geography,* vol. 3, no. 3, pp. 274-279, 2013.

[9] S. Yin, and O. Kaynak, "Big Data for Modern Industry: Challenges and Trends [Point of View]," *Proceedings of the IEEE,* vol. 103, no. 2, pp. 143-146, 2015.

[10] F. Provost, and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Big Data,* vol. 1, no. 1, pp. 51, 2013.

[11] S. Ma, and C. Ji, "Performance and efficiency: recent advances in supervised learning," *Proceedings of the IEEE,* vol. 87, no. 9, pp. 1519-1535, 2002.

[12] X. Chen, S. Peng, J. Z. Huang, F. Nie, and Y. Ming, "Local PurTree Spectral Clustering for Massive Customer Transaction Data," *IEEE Intelligent Systems,* vol. 32, no. 2, pp. 37-44, 2017.

[13] Kusiak, and Andrew, "Smart manufacturing must embrace big data," *Nature*, pp. 23-25, 2017.

[14] S. M. Krishnan, "Application of Analytics to Big Data in Healthcare." pp. 156-157.

[15] M. I. Jordan, and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science,* vol. 349, no. 6245, pp. 255-60, 2015.

[16] S. Yu, "Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data," *IEEE Access,* vol. 4, pp. 1-1, 2016.

[17] P. Domingos, "The Master Algorithm," 2015.

[18] Z. Xu, R. Jin, J. Zhu, I. King, and M. R. Lyu, "Efficient Convex Relaxation for Transductive Support Vector Machine." pp. 1641-1648.

[19] A. Blum, and T. Mitchell, "Combining labeled and unlabeled data with co-training." pp. 92-100.

[20] F. Wang, and C. Zhang, "Label propagation through linear neighborhoods." pp. 985-992.

[21] A. Talwalkar, S. Kumar, and H. Rowley, "Large-scale manifold learning." pp. 1-8.

[22] I. Menache, S. Mannor, and N. Shimkin, "Basis Function Adaptation in Temporal Difference Reinforcement Learning," *Annals of Operations Research,* vol. 134, no. 1, pp. 215-238, 2005.

[23] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with Deep Reinforcement Learning," *Computer Science*, 2013.

[24] S. J. Pan, and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge & Data Engineering,* vol. 22, no. 10, pp. 1345-1359, 2010.

[25] A. J. Smola, "Learning with Kernels," *IEEE Transactions on Signal Processing,* vol. 52, no. 8, pp. 2165-2176, 1998.

[26] X. Xu, T. M. Hospedales, and S. Gong, *Multi-Task Zero-Shot Action Recognition with Prioritised Data Augmentation*: Springer International Publishing, 2016.

[27] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active Learning with Statistical Models," vol. 4, no. 1, pp. 705-712, 1996.

[28] R. Herbrich, "Machine Learning at Amazon." pp. 535-535.

[29] S. Qiu, B. Chen, R. Wang, Z. Zhu, Y. Wang, and X. Qiu, "Estimating contaminant source in chemical

industry park using UAV-based monitoring platform, artificial neural network and atmospheric dispersion simulation," *RSC Advances,* vol. 7, no. 63, pp. 39726-39738, 2017.

[30] V. D. M. Laurens, G. Hinton, and V. D. M. Hinton, Geoffrey, "Visualizing Data using t-SNE," *Journal of Machine Learning Research,* vol. 9, no. 2605, pp. 2579-2605, 2008.

[31] P. Domingos, "A few useful things to know about machine learning," *Communications of the Acm,* vol. 55, no. 10, pp. 78–87, 2012.

[32] P. Domingos, "The Role of Occam's Razor in Knowledge Discovery," *Data Mining & Knowledge Discovery,* vol. 3, no. 4, pp. 409-425, 1999.

[33] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive Auto-Encoders: Explicit Invariance During Feature Extraction."

[34] V. N. Vapnik, *The Nature of Statistical Learning Theory*: Springer, 1995.

[35] C. Aldrich, and L. Auret, "Statistical Learning Theory and Kernel-Based Methods," *Advances in Computer Vision & Pattern Recognition*, pp. 117-181, 2013.

[36] S. Floyd, "Space-bounded learning and the Vapnik-Chervonenkis dimension," *Proceedings of the Second Annual Workshop on Computational Learning Theory*, pp. 349-364, 1989.

[37] Z. Chen, and S. Haykin, "On different facets of regularization theory," *Neural Computation,* vol. 14, no. 12, pp. 2791-2846, 2002.

[38] A. Tikhonov, "Solution of Incorrectly Formulated Problems and the Regularization Method," *Soviet Math Dokl,* vol. 5, 1963.

[39] P. L. Bartlett, and S. Mendelson, "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results," *Journal of Machine Learning Research,* vol. 3, pp. 463-482, 2003.

[40] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning from data*: AMLBook New York, NY, USA:, 2012.

[41] P. Niyogi, F. Girosi, and T. Poggio, "Incorporating prior information in machine learning by creating virtual examples," *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2196-2209, 2009.

[42] M. Paulin, J. Revaud, Z. Harchaoui, F. Perronnin, and C. Schmid, "Transformation Pursuit for Image Classification." pp. 3646-3653.

[43] M. Sassano, "Virtual examples for text classification with support vector machines," vol. 13, no. 3, pp. 21-35, 2003.

[44] D. C. Li, and Y. S. Lin, "Using virtual sample generation to build up management knowledge in the early manufacturing stages," *European Journal of Operational Research,* vol. 175, no. 1, pp. 413-434, 2006.

[45] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets." pp. 2672-2680.

[46] J. Zhao, M. Mathieu, and Y. Lecun, "Energy-based Generative Adversarial Network," 2016.

[47] M. T. Rosenstein, Z. Marx, L. P. Kaelbling, and T. G. Dietterich, "To Transfer or Not To Transfer." p. S20.

[48] H. Zhang, H. Ji, and X. Wang, "Transfer Learning from Unlabeled Data via Neural Networks," *Neural Processing Letters,* vol. 36, no. 2, pp. 173-187, 2012.

[49] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning," 2007.

[50] V. Cherkassky, and F. Mulier, "Learning from Data: Concepts, Theory, and Methods (Adaptive and Learning Systems for Signal Processing, Communications and Control Series)," 1998.

[51] R. Reed, "Pruning algorithms-a survey," *IEEE Transactions on Neural Networks,* vol. 4, no. 5, pp. 740-7, 1993.

[52] D. C. Li, and C. W. Liu, "Extending Attribute Information for Small Data Set Classification," *IEEE Transactions on Knowledge & Data Engineering,* vol. 24, no. 3, pp. 452-464, 2010.

 doi:10.20944/preprints201710.0076.v1

[53] L. Prechelt, “Automatic early stopping using cross validation: quantifying the criteria,” *Neural Networks the Official Journal of the International Neural Network Society,* vol. 11, no. 4, pp. 761, 1998.

[54] S. Sagiroglu, and D. Sinanc, "Big data: A review." pp. 42-47.

[55] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, “Data Mining with Big Data,” *IEEE Transactions on Knowledge & Data Engineering,* vol. 26, no. 1, pp. 97-107, 2014.

[56] Y. T. Zhuang, F. Wu, C. Chen, and Y. H. Pan, “Challenges and opportunities: from big data to knowledge in AI 2.0,” *Frontiers of Information Technology & Electronic Engineering,* vol. 18, no. 1, pp. 3-14, 2017.

[57] J. Zhu, J. Chen, W. Hu, and B. Zhang, “Big Learning with Bayesian Methods,” *Computer Science*, 2014.

[58] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, “Machine learning on big data: Opportunities and challenges,” *Neurocomputing,* vol. 237, pp. 350-361, 2017.

[59] T.-Y. Liu, W. Chen, and T. Wang, "Distributed Machine Learning: Foundations, Trends, and Practices." pp. 913-915.

[60] D. Lahat, T. Adali, and C. Jutten, “Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects,” *Proceedings of the IEEE,* vol. 103, no. 9, pp. 1449-1477, 2015.

[61] Y. Zheng, “Methodologies for Cross-Domain Data Fusion: An Overview,” *IEEE Transactions on Big Data,* vol. 1, no. 1, pp. 16-34, 2015.

[62] A. J. C. Walters, and G. M. Freeman, “The quality of big (geo)data,” *Dialogues in Human Geography,* vol. 3, no. 3, pp. 280-284, 2013.

[63] R. Y. Wang, and D. M. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” *Journal of Management Information Systems,* vol. 12, no. 4, pp. 5-33, 1996.

[64] B. Saha, and D. Srivastava, "Data quality: The other face of Big Data." pp. 1294-1297.

[65] R. Bellman, “Dynamic programming,” *Science,* vol. 153, no. 3731, pp. 34-37, 1966.

[66] C. Sammut, and G. I. Webb, *Encyclopedia of Machine Learning*: Springer US, 2010.

[67] M. Dash, and H. Liu, *Feature Selection for Classification*: IOS Press, 1997.

[68] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Transactions on Pattern Analysis & Machine Intelligence,* vol. 27, no. 8, pp. 1226, 2005.

[69] H. Liu, and Z. Zhao, “Manipulating Data and Dimension Reduction Methods: Feature Selection,” *Nature Methods,* vol. 6, no. 11 Suppl, pp. 109-114, 2009.

[70] B. Xue, M. Zhang, W. N. Browne, and X. Yao, “A Survey on Evolutionary Computation Approaches to Feature Selection,” *IEEE Transactions on Evolutionary Computation,* vol. 20, no. 4, pp. 1-1, 2015.

[71] B. Xue, M. Zhang, W. N. Browne, and X. Yao, “A Survey on Evolutionary Computation Approaches to Feature Selection,” *IEEE Transactions on Evolutionary Computation,* vol. 20, no. 4, pp. 606-626, 2016.

[72] R. Lab, and P. D. Gunnar Rätsch, “Kernel PCA and De-Noising in Feature Spaces,” *Advances in Neural Information Processing Systems,* vol. 11, pp. 536--542, 1999.

[73] C. Eckart, and G. Young, “The approximation of one matrix by another of lower rank,” *Psychometrika,* vol. 1, no. 3, pp. 211-218, 1936.

[74] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: online learning of social representations." pp. 701-710.

[75] A. Grover, and J. Leskovec, “node2vec: Scalable Feature Learning for Networks,” vol. 2016, pp. 855, 2016.

[76] S. B. Ho, "Deep thinking and quick learning for viable AI."

[77] Z. Obradovic, and S. Vucetic, “Challenges in Scientific Data Mining: Heterogeneous, Biased, and Large Samples,” 2004.

[78] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, “Learning in Nonstationary Environments: A Survey,”

IEEE Computational Intelligence Magazine, vol. 10, no. 4, pp. 12-25, 2015.

[79] C. Alippi, and M. Roveri, "An adaptive CUSUM-based test for signal change detection," 2006.

[80] K. Nishida, and K. Yamauchi, "Detecting concept drift using statistical testing." pp. 264-269.

[81] C. Alippi, G. Boracchi, and M. Roveri, "A hierarchical, nonparametric, sequential change-detection test." pp. 2889-2896.

[82] D. M. Hawkins, P. Qqui, and C. W. Kang, "The changepoint model for statistical process control," *Journal of Quality Technology,* vol. 35, no. 4, pp. 355-366, 2003.

[83] C. Alippi, "Ensembles of change-point methods to estimate the change point in residual sequences," *Soft Computing,* vol. 17, no. 11, pp. 1971-1981, 2013.

[84] A. Bifet, and R. Gavaldà, "Learning from Time-Changing Data with Adaptive Windowing."

[85] I. Koychev, "Gradual Forgetting for Adaptation to Concept Drift." pp. 101--106.

[86] R. Klinkenberg, "Learning drifting concepts: Example selection vs. example weighting," *Intelligent Data Analysis,* vol. 8, no. 3, pp. 281-300, 2004.

[87] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams." pp. 97-106.

[88] L. Cohen, G. Avrahami-Bakish, M. Last, A. Kandel, and O. Kipersztok, "Real-time data mining of non-stationary data streams from sensor networks ☆," *Information Fusion,* vol. 9, no. 3, pp. 344-353, 2008.

[89] C. Alippi, D. Liu, D. Zhao, and L. Bu, "Detecting and Reacting to Changes in Sensing Units: The Active Classifier Case," *IEEE Transactions on Systems Man & Cybernetics Systems,* vol. 44, no. 3, pp. 353-362, 2017.

[90] C. Alippi, G. Boracchi, and M. Roveri, "Just-in-time classifiers for recurrent concepts," *IEEE Transactions on Neural Networks & Learning Systems,* vol. 24, no. 4, pp. 620, 2013.

[91] R. Elwell, and R. Polikar, "Incremental learning of concept drift in nonstationary environments," *IEEE Transactions on Neural Networks,* vol. 22, no. 10, pp. 1517, 2011.

[92] Q. YANG, and X. WU, "10 CHALLENGING PROBLEMS IN DATA MINING RESEARCH," *International Journal of Information Technology & Decision Making,* vol. 5, no. 04, pp. -, 2006.

[93] Q. Wang, Z. Luo, J. Huang, Y. Feng, and Z. Liu, "A Novel Ensemble Method for Imbalanced Data Learning: Bagging of Extrapolation-SMOTE SVM," *Computational Intelligence & Neuroscience,* vol. 2017, pp. 1827016, 2017.

[94] A. K. Tanwani, and M. Farooq, "The Role of Biomedical Dataset in Classification." pp. 370-374.

[95] R. Barandela, R. M. Valdovinos, J. S. Sánchez, and F. J. Ferri, "The Imbalanced Training Sample Problem: Under or over Sampling?." p. 806.

[96] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research,* vol. 16, no. 1, pp. 321-357, 2002.

[97] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning." pp. 878-887.

[98] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "AdaCost: Misclassification Cost-Sensitive Boosting." pp. 97--105.

[99] X. Y. Liu, J. Wu, and Z. H. Zhou, "Exploratory undersampling for class-imbalance learning." pp. 965-969.

[100] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* vol. 39, no. 4, pp. 652-663, 2017.

[101] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer*

Vision, vol. 115, no. 3, pp. 211-252, 2015.

[102] Y. Lecun, and C. Cortes, "The mnist database of handwritten digits," 2010.

[103] J. Bootkrajang, and A. Kaban, "Boosting in the presence of label noise," *Computer Science*, 2013.

[104] J. Bootkrajang, and N. Ata, "Label-Noise robust logistic regression and its applications." pp. 143-158.

[105] W. Wei, and Z. H. Zhou, "Crowdsourcing label quality: a theoretical analysis," *Science China Information Sciences,* vol. 58, no. 11, pp. 1-12, 2015.

[106] C. Thiel, "Classification on Soft Labels Is Robust against Label Noise." pp. 65-73.

[107] J. Zhang, V. S. Sheng, J. Wu, X. Fu, and X. Wu, "Improving Label Quality in Crowdsourcing Using Noise Correction." pp. 1931-1934.

[108] X. Zhu, and X. Wu, "Class noise vs. attribute noise: A quantitative study," *Artificial Intelligence Review,* vol. 22, no. 3, pp. 177-210, 2004.

[109] D. Garcíagil, J. Luengo, S. García, and F. Herrera, "Enabling Smart Data: Noise filtering in Big Data classification," 2017.

[110] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Support Vector Machine for Outlier Detection in Breast Cancer Survivability Prediction." pp. 99-109.

[111] D. L. Wilson, "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data," *IEEE Transactions on Systems Man & Cybernetics,* vol. 2, no. 3, pp. 408-421, 1972.

[112] B. Frénay, "Uncertainty and label noise in machine learning," *Verleysen Michel*, 2013.

[113] B. Sluban, D. Gamberger, and N. Lavra, "Advances in Class Noise Detection." pp. 1105-1106.

[114] U. Rebbapragada, and C. E. Brodley, "Class Noise Mitigation Through Instance Weighting." pp. 708-715.

[115] C. Bouveyron, and S. Girard, "Robust supervised classification with mixture models: Learning from data with uncertain labels," *Pattern Recognition,* vol. 42, no. 11, pp. 2649-2658, 2009.

[116] J. Bootkrajang, and A. Kabán, "Label-Noise Robust Logistic Regression and Its Applications." pp. 143-158.

[117] J. Bootkrajang, and A. Kaban, "Multi-class Classification in the Presence of Labelling Errors."

[118] T. Liu, and D. Tao, "Classification with Noisy Labels by Importance Reweighting," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* vol. 38, no. 3, pp. 447-461, 2016.

[119] R. Wang, T. Liu, and D. Tao, "Multiclass Learning With Partially Corrupted Labels," *IEEE Transactions on Neural Networks & Learning Systems,* vol. PP, no. 99, pp. 1-13, 2017.

[120] H. David, and G. William, "No Free Lunch Theorems for Search," *Working Papers,* vol. 122, no. 1431, pp. 431-434, 1995.

[121] A. K. Tanwani, J. Afridi, M. Z. Shafiq, and M. Farooq, "Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets." pp. 128-139.

[122] D. J. Wilkinson, "Bayesian methods in bioinformatics and computational systems biology," *Briefings in Bioinformatics,* vol. 8, no. 2, pp. 109-116, 2007.

[123] Q. Wang, Y. Feng, and T. Wang, "Drug Target Protein-Protein Interaction Networks: A Systematic Perspective," *BioMed Research International,2017,(2017-6-11),* vol. 2017, no. 7, pp. 1-13, 2017.

[124] D. Liben-Nowell, and J. Kleinberg, "The link prediction problem for social networks," *Journal of the American Society for Information Science & Technology,* vol. 58, no. 7, pp. 1019–1031, 2007.

[125] Domingos, and Pedro, *The master algorithm : how the quest for the ultimate learning machine will remake our world*: Basic Books, a member of the Perseus Books Group, 2015.

[126] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-444, 2015.

[127] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *Journal of*

Machine Learning Research, vol. 11, no. 12, pp. 3371-3408, 2010.

[128] H. Lee, L. Yan, P. Pham, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks." pp. 1096-1104.

[129] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks." pp. 1097-1105.

[130] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," *Computer Science*, pp. 338-342, 2014.

[131] D. M. J. Tax, and R. P. W. Duin, "Support vector domain description," *Pattern Recognition Letters,* vol. 20, no. 11–13, pp. 1191-1199, 1999.

[132] Jayadeva, R. Khemchandani, and S. Chandra, "Twin Support Vector Machines for pattern classification," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* vol. 29, no. 5, pp. 905, 2007.

[133] M. Awad, and R. Khanna, "Support Vector Regression," *Neural Information Processing Letters & Reviews,* vol. 11, no. 10, pp. 203--224, 2007.

[134] V. Vovk, *Kernel Ridge Regression*: Springer Berlin Heidelberg, 2013.

[135] S. Ding, N. Zhang, X. Xu, L. Guo, and J. Zhang, "Deep Extreme Learning Machine and Its Application in EEG Classification," *Mathematical Problems in Engineering,2015,(2015-5-27),* vol. 2015, no. 1, pp. 1-11, 2015.

[136] S. Scardapane, D. Comminiello, M. Scarpiniti, and A. Uncini, "Online Sequential Extreme Learning Machine With Kernels," *IEEE Transactions on Neural Networks & Learning Systems,* vol. 26, no. 9, pp. 2214, 2015.

[137] B. Mirza, Z. Lin, and K. A. Toh, *Weighted Online Sequential Extreme Learning Machine for Class Imbalance Learning*: Kluwer Academic Publishers, 2013.

[138] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training Convolutional Networks with Noisy Labels," *Computer Science*, 2014.

[139] G. Patrini, A. Rozza, A. Menon, R. Nock, and L. Qu, "Making Deep Neural Networks Robust to Label Noise: a Loss Correction Approach," 2016.

[140] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training Deep Neural Networks on Noisy Labels with Bootstrapping," *Computer Science*, 2014.

[141] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research,* vol. 15, no. 1, pp. 1929-1958, 2014.

[142] S. Ioffe, and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," pp. 448-456, 2015.

[143] Z. H. Zhou, and J. Feng, "Deep Forest: Towards An Alternative to Deep Neural Networks," 2017.

[144] L. V. Utkin, and M. A. Ryabinin, "A Siamese Deep Forest," 2017.

[145] K. Miller, C. Hettinger, J. Humpherys, T. Jarvis, and D. Kartchner, "Forward Thinking: Building Deep Random Forests," 2017.

[146] T. Dietterich, "Machine Learning Challenges in Ecological Science and Ecosystem Management."

[147] T. Joachims, "Training linear SVMs in linear time." pp. 217-226.

[148] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks." pp. 985-990 vol.2.

[149] G. B. Huang, L. Chen, and C. K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Transactions on Neural Networks,* vol. 17, no. 4, pp. 879, 2006.

[150] G. B. Huang, and L. Chen, “Letters: Convex incremental extreme learning machine,” *Neurocomputing,* vol. 70, no. 16–18, pp. 3056-3062, 2007.

[151] H. J. Rong, Y. S. Ong, A. H. Tan, and Z. Zhu, “A fast pruned-extreme learning machine for classification problem. Neurocomputing,” *Neurocomputing,* vol. 72, no. 1, pp. 359-366, 2008.

[152] T. Guo, L. Zhang, and X. Tan, “Neuron Pruning-Based Discriminative Extreme Learning Machine for Pattern Classification,” *Cognitive Computation*, no. 1, pp. 1-15, 2017.

[153] T. G. Dietterich, “Machine Learning Research: Four Current Directions,” *Ai Magazine,* vol. 18, no. 4, pp. 97-136, 1997.

[154] Z. H. Zhou, *Ensemble Methods: Foundations and Algorithms*: Taylor & Francis, 2012.

[155] L. Breiman, “Bagging Predictors,” *Machine Learning,* vol. 24, no. 2, pp. 123-140, 1996.

[156] D. D. Margineantu, and T. G. Dietterich, "Pruning Adaptive Boosting." pp. 211-218.

[157] E. B. Kong, and T. G. Dietterich, "Error-Correcting Output Coding Corrects Bias and Variance." pp. 313–321.

[158] G. Casalicchio, “Being successful on Kaggle,” 2017.

[159] T. Chen, T. He, and M. Benesty, “xgboost: Extreme Gradient Boosting,” 2017.

[160] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, “Efficient and robust automated machine learning,” *Ecological Informatics,* vol. 30, pp. 49-59, 2015.

[161] I. Guyon, K. Bennett, G. Cawley, and H. J. Escalante, "Design of the 2015 ChaLearn AutoML challenge." pp. 1-8.

[162] T. M. Mitchell, “Machine Learning and Data Mining: Machine learning algorithms enable discovery of important “regularities” in large data sets,” 1999.

[163] J. B. Tenenbaum, *Building Theories of the World: Human and Machine Learning Perspectives*: Springer Berlin Heidelberg, 2008.

[164] D. Sacha, M. Sedlmair, L. Zhang, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim, “What you see is what you can change: human-centred machine learning by interactive visualization,” 2017.

[165] D. Sacha, M. Sedlmair, L. Zhang, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim, “What You See Is What You Can Change: Human-Centered Machine Learning By Interactive Visualization,” *Neurocomputing*, 2017.

[166] E. Bertini, and D. Lalanne, “Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery,” *Acm Sigkdd Explorations Newsletter,* vol. 11, no. 2, pp. 9-18, 2010.

[167] G. Tur, and D. Hakkani, “Combining active and semi-supervised learning for spoken language understanding,” *Speech Communication,* vol. 45, no. 2, pp. 171-186, 2004.

[168] L. Mihalkova, and R. J. Mooney, "Using Active Relocation to Aid Reinforcement Learning." pp. 580-585.

[169] X. Zhang, J. Cheng, C. Xu, H. Lu, and S. Ma, "Multi-view multi-label active learning for image classification." pp. 258-261.

[170] S. Russell, D. Dewey, and M. Tegmark, “Research Priorities for Robust and Beneficial Artificial Intelligence,” *Ai Magazine*, 2015.

[171] Z.-H. Zhou, “Machine learning challenges and impact: an interview with Thomas Dietterich,” *National Science Review*.

[172] Y. Jin, “Multi-Objective Machine Learning,” vol. 38, no. 3, pp. 2, 2006.

[173] Y. Jin, and B. Sendhoff, “Pareto-Based Multiobjective Machine Learning: An Overview and Case Studies,” *IEEE Transactions on Systems Man & Cybernetics Part C,* vol. 38, no. 3, pp. 397-415, 2008.