Mathematical Problems of Computer Science 63, 25–41, 2025.
doi:10.51408/1963-0129

UDC 519.72

Information Theory Tools and Techniques to Overcome Machine Learning Challenges

Mariam E. Haroutunian and Gor A. Gharagyozyan

Institute for Informatics and Automation Problems of NAS RA, Yerevan, Armenia
e-mail: armar@sci.am, gor.gharagyozyan@edu.isec.am

Abstract

In this survey, we explore the broad applications of Information Theory in Machine Learning, highlighting how core concepts like *entropy*, *Mutual Information*, and *KL-divergence* are used to enhance learning algorithms. Since its inception by Claude Shannon, Information Theory has provided mathematical tools to quantify uncertainty, optimize decision-making, and manage the trade-off between model flexibility and generalization. These principles have been integrated across various subfields of Machine Learning, including *neural networks*, where the *Information Bottleneck* offers insights into data representation, and *reinforcement learning*, where entropy-based methods improve exploration strategies. Additionally, measures like *Mutual Information* are critical in *feature selection* and *unsupervised learning*. This survey bridges foundational theory with its practical implementations in modern Machine Learning by providing both historical context and a review of contemporary research.. We also discuss open challenges and future directions, such as scalability and interpretability, highlighting the growing importance of these techniques in next-generation models.

Keywords: Information Bottleneck, Neural networks, Entropy-Based regularization, Mutual information, Feature selection, KL-Divergence.

Article info: Received 30 September 2024; sent for review 15 October 2024; received in revised form 1 April 2025, accepted 7 April 2025.

1. Introduction

The intersection of *Information Theory (IT)* and *Machine Learning (ML)* has become increasingly pivotal in advancing the state of the art across a wide range of subfields. *IT*, formalized by Claude

Shannon in his seminal 1948 work [1], introduced foundational concepts like entropy, which measures the uncertainty or disorder of a system, and Mutual Information (MI), which quantifies the amount of information one variable contains about another. These principles have profound implications in ML, particularly in optimizing algorithms, managing uncertainty, and improving decision-making processes.

In the context of ML, models often grapple with the *bias-variance trade-off*, striving to balance flexibility with generalization. Information-theoretic techniques such as *minimum description length* [2] provide an elegant way of navigating this trade-off by minimizing the complexity of models while maintaining accuracy. Similarly, *maximum entropy* models [3] leverage entropy to derive distributions that reflect uncertainty in the absence of prior knowledge, making them useful in many predictive models.

The impact of IT on ML is far-reaching:

- In *neural networks*, the *Information Bottleneck (IB)* method offers a theoretical framework for understanding how deep networks compress and transmit information through their layers [4].
- *Reinforcement learning* employs *entropy-based regularization* to enhance exploration strategies, helping agents avoid local optima and discover better policies [5].
- *Feature selection* relies on *MI* to identify the most relevant variables while discarding redundant or irrelevant data, which is crucial for high-dimensional datasets [6].
- *Unsupervised learning* techniques such as *autoencoders* and *variational autoencoders* rely on information-theoretic measures like *KL-divergence* to ensure that latent representations capture the essential structure of data [7].

As the field of ML continues to evolve, information-theoretic methods remain central to the development of robust and efficient models. Recent advancements have brought renewed attention to these techniques, particularly in addressing the challenges of scalability, interpretability, and privacy in deep learning systems. The *IB* theory, for example, provides insights into how models generalize and perform in real-world tasks by analyzing the flow of information between inputs and outputs [8]. Moreover, information-theoretic approaches have been increasingly employed in cutting-edge fields such as *quantum ML*, where *quantum IT* principles are applied to create more powerful algorithms [9].

This survey aims to provide a comprehensive overview of the recent developments, current applications, and future directions of IT in ML. This investigation will provide future good basis for bridging the gap between foundational theory [10] and cutting-edge research.

The paper is organized as follows: in the next section main concepts of IT are described. Main IT tools applied in ML are discussed in Section 3. Particular emphasis is placed on the IB framework in Section 4. Section 5 discusses the challenges and limitations of IT in ML. The paper is summarized in Section 6.

2. Useful IT Concepts

Entropy: Measuring Uncertainty

Entropy is the cornerstone of IT, introduced by Claude Shannon in 1948 [1], and is a measure of the uncertainty or randomness inherent in a random variable or a probability distribution [11]. In ML, entropy plays critical role in quantifying the amount of unpredictability in data, making it a crucial tool for optimizing algorithms and decision-making processes.

For a discrete random variable X with a probability distribution $P(X)$, where X can take values $\{x_1, x_2, \ldots, x_2\}$ with probabilities $\{p(x_1), p(x_2), \ldots, p(x_n)\}$, the $entropy\ H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log\ p(x_i),$$

where:

- $p(x_i)$ is the probability of accurrence of the outcome x_i,
- log is the logarithm base 2, as entropy is typically measured in *bits.*

The formula represents the *expected number of bits* required to encode the outcomes of X given their probabilities. Entropy achieves its maximum value when all outcomes are equally probable (maximum uncertainty) and its minimum value when one outcome is certain (no uncertainty).

Conditional Entropy and *Joint Entropy* are extensions of this concept. *Conditional Entropy* $H(X \mid Y)$ quantifies the uncertainty of X given that Y is known, while *Joint Entropy* $H(X, Y)$ captures the combined uncertainty of two random variables.

$$H(X \mid Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) log\ p(x|y),$$

$$H(X,Y) = -\sum_{y \in Y} \sum_{x \in X} p(x,y) log\ p(x,y).$$

Mutual Information: Quantifying Shared Information

MI measures the amount of information shared between two random variables, quantifying how much knowing the value of one variable reduces uncertainty about the other. Formally, the *MI* between two random variables X and Y is defined as:

$$I(X; Y) = H(X) - H(X \mid Y) = \sum_{y \in Y} \sum_{x \in X} \frac{p(x,y)}{p(x)p(y)}.$$

MI can be thought of as the reduction in uncertainty about X when Y is known. Unlike *correlation*, which captures linear relationships, MI detects any kind of dependency between the variables, making it more robust for applications like *feature selection* [6]. In ML, MI is used to rank features based on their relevance to the target variable, allowing models to focus on the most informative inputs. For example, in *feature selection*, MI helps to identify and remove irrelevant or redundant features, significantly improving model performance by reducing overfitting in high-dimensional spaces.

KL-Divergence: Measuring the Difference Between Distributions

Kullback-Leibler Divergence (KL-Divergence), also known as *relative entropy*, is a measure of

how one probability distribution differs from a second, reference distribution. For two probability distributions P and Q, the KL-Divergence from Q to P is defined as:

$$D_{\mathrm{KL}}(P||Q) = \sum_{x\in X} p(x) log \frac{p(x)}{q(x)}.$$

KL-Divergence is non-negative and equals zero when the distributions are identical. Unlike traditional distance metrics, it is *asymmetric*, meaning $D_{KL}(P||Q) \neq D_{KL}(Q||P)$.

KL-Divergence is particularly useful in tasks where we approximate a complex distribution P with a simpler distribution Q, such as in *variational inference* [7].

In *variational autoencoders,* KL-Divergence is used to measure how close the learned latent variable distribution is to a prior distribution, such as a standard normal distribution. This ensures that the learned representations are regularized and maintain structure during training.

Cross-Entropy: Optimizing Classification Models

Cross-Entropy is closely related to KL-Divergence, but is more commonly used in *classification problems.* While KL-Divergence measures the divergence between two probability distributions, cross-entropy quantifies the total number of bits needed to encode a distribution P using another distribution Q, cross-entropy is given by:

$$H(P,Q) = -\sum_{x\in X} p(x) \log q(x).$$

In ML, cross-entropy loss is widely used as a *loss function* for classification tasks, particularly for models that output probability distributions, like *softmax classifiers.* It measures how well the predicted probabilities (from model Q) align with the true distribution (actual labels, P). Minimizing cross-entropy encourages the model to assign high probabilities to the correct classes.

For binary classification problem, the cross-entropy loss can be written as:

$$L = -[y log\, p + (1-y) log(1-p)],$$

where y is the true label (0 or 1), and p is the predicted probability of the label being 1.

Maximum Entropy Principle

The *Maximum Entropy* principle suggests that, when faced with uncertainty, the best distribution to choose is the one that maximizes entropy, subject to any known constraints. This principle underprints *maximum entropy models*, often used in areas like *natural language processing* [12]. These models choose the distribution that remains as uncertain as possible (i.e., has the highest entropy) while still satisfying the constraints imposed by the available data.

The principle encourages generality and reduces assumptions, making it useful for creating unbiased models when prior knowledge is limited.

3. Overview of IT Tools for ML

The application of IT concepts, such as entropy, MI, KL-divergence, and cross-entropy, has significantly advanced ML methodologies. These tools enable effective feature selection, model optimization, regularization, and performance evaluation. Below, we explore how these principles are utilized in practical ML tasks.

Feature Selection and Dimensionality Reduction

One of the most prominent applications of *MI* is in *feature selection.* In high-dimensional datasets, identifying the most relevant features for the model is crucial to improve performance and reduce overfitting. MI helps in selecting features that share maximum information with the target variable while avoiding redundant or irrelevant features. The *Max-Relevance and Min-Redundancy* algorithm is a widely used feature selection technique, that maximizes MI between features and the target variable while minimizing redundancy among the selected features [6]. This ensures that the selected features are both informative and diverse. In [6], MI was applied to gene selection for cancer detection. This approach identified the genes most relevant for distinguishing between cancerous and non-cancerous cells, reducing the dataset's dimensionality while retaining the most predictive features. This process significantly improved the performance of classification algorithms, such as Support Vector Machines, by focusing on the genes that contained the most meaningful information about the cancer type. Here, MI $I(X; Y)$ is used to quantify the relationship between the input features X and the target label Y, ensuring that the selected features contribute significantly to the predictive power of the model.

Building upon MI-driven feature selection, [13] proposed a fast binary feature selection method using Conditional MI. This approach refines MI-based selection by conditioning on already-selected features, ensuring that each additional feature contributes new, independent information to the model. The efficiency of this method enables rapid selection from datasets with tens of thousands of features, making it highly suitable for large-scale applications in computer vision and pattern recognition. Additionally, [14] explored MI-based feature selection techniques tailored for non-Gaussian data distributions. Their work introduced new feature selection and visualization algorithms that address challenges posed by high-dimensional, non-Gaussian datasets. By leveraging information-theoretic measures, their method improves both interpretability and feature selection performance in complex data environments, making it particularly useful in scientific and industrial applications, where data distributions deviate from Gaussian assumptions. Another approach leveraging MI for feature selection is presented in [15]. The method selects *class-specific informative features*, maximizing MI with the target class to enhance classification performance. This allows even a simple linear classifier to be effective, reducing reliance on complex models. While applied to object recognition, its principles extend to high-dimensional classification tasks, where efficient feature selection is essential.

Decision Trees and Information Gain

Entropy plays a central role in the construction of *decision trees*, where it is used to calculate *information gain.* Information gain measures the reduction in uncertainty (or entropy) when a dataset is split based on a particular feature. A decision tree algorithm selects features with the highest information gain to create branches, effectively reducing the overall entropy of the system [16]. In the popular *ID3* and *C4.5* decision tree algorithms, the feature that results in the greatest reduction in entropy after splitting is chosen to create nodes in the tree. This process continues recursively, ensuring that each split reduces uncertainty and leads to the most informative partitions of the data.

$$Information\ Gain\ =\ H(Y)\ -\ H(Y|X).$$

By minimizing entropy at each step, decision trees efficiently organize data and create models that are easy to interpret. However, their usage extends beyond traditional datasets into fields like high-energy physics, where rapid detection of rare phenomena is critical. A recent study [17] demonstrates the application of decision trees in detecting anomalies in proton-proton collision data at nanosecond timescales. This work specifically focuses on identifying rare *Higgs boson decays* in real-time. The decision trees in this application rely on fast, efficient calculations of information gain to classify particle collision data, reducing entropy by isolating potential anomalies that deviate from expected particle behaviors.

Another interesting work is [18]. This study tackles the challenge of securely training and evaluating decision trees in cloud environments without exposing sensitive data. The authors introduce a method based on additive secret sharing and the Paillier cryptosystem to protect both user queries and the cloud-hosted model. Their approach ensures secure computation while supporting offline users, making it suitable for resource-constrained applications like Internet of Thinking. Experimental results confirm its efficiency, particularly for deep but sparse trees, demonstrating reduced computational and communication overhead.

Clustering and Similarity Measurement

In unsupervised learning tasks like clustering, MI is used to measure the similarity between data points or clusters. The goal of clustering is to group similar data points together, and MI can help to determine how much information is shared between the clustering results and the true labels, when available.

One notable application of ML in clustering is *Normalized MI*, which measures the similarity between two clusterings. *Normalized MI* is particularly valuable when evaluating the quality of clustering results, as it quantifies the shared information between the true class labels and the predicted clusters, normalized by the entropy of both distributions. This ensures that the score is independent of the number of clusters and the size of the dataset. *Normalized MI* is widely used in applications such as document clustering, image segmentation and analyzing [19], where it is crucial to assess the quality of unsupervised learning methods.

Fuzzy clustering (a form of clustering in which each data point can belong to more than one cluster) plays a critical role in ML applications. Traditional clustering algorithms, such as k-means, assume hard partitioning of the data, meaning each data point belongs exclusively to one cluster. However, in many real-world scenarios, data points may naturally belong to multiple clusters with varying degrees of membership. Fuzzy clustering, specifically Probabilistic Fuzzy Clustering, allows for such flexibility by assigning each data point a degree of membership across different clusters.

The *Robust Possibilistic Fuzzy Additive Partition Clustering* method, as introduced in a recent study [20], builds upon these principles by incorporating deep local information to optimize the clustering process. This method leverages local data structures to improve clustering accuracy, particularly in noisy and uncertain environments. The algorithm dynamically adjusts the partitioning of data, thus reducing the impact of noise and outliers - a common issue in clustering. A significant extension of MI-based clustering techniques comes from the *Information-Theoretic Co-Clustering* approach introduced in [21]. This method simultaneously clusters both rows and columns of a data matrix, optimizing an MI loss function to uncover latent structures within

datasets. This framework has been particularly influential in *text mining and bioinformatics*, where data is inherently organized in two dimensions, such as documents and words, or genes and experimental conditions. By minimizing information loss in the clustering process, this method provides a more interpretable and structured representation of high-dimensional data.

Further advancing the theoretical foundations of MI in clustering, [22] proposed *Information-Theoretical Clustering via Semidefinite Programming.* Unlike conventional clustering approaches, which often rely on heuristic optimization, this method employs *semidefinite programming* to ensure a *globally optimal partitioning* of data based on MI principles. The approach has shown effectiveness in areas such as *image segmentation and social network analysis*, where precise and stable clustering is crucial.

In the domain of *collaborative filtering*, [23] introduced an *Information-Theoretic Co-Clustering approach* to improve recommendation systems. Traditional collaborative filtering often suffers from sparsity issues, where users have rated only a small fraction of available items. By leveraging MI to extract shared patterns from user-item matrices, this method enhances recommendation accuracy by capturing both cluster-based preferences and rating similarities. This improvement makes it particularly valuable for applications in e-commerce and content recommendation platforms. A more recent contribution by [24] introduces *Co-Clustering via Information-Theoretic Markov Aggregation*. This method constructs a *random walk on a bipartite graph*, optimizing an MI-based cost function to extract meaningful co-clusters. By reducing information loss during clustering, this technique closely aligns with *the IB framework*, demonstrating superior performance in structured datasets like *Newsgroup20* and *MovieLens100k*. Its effectiveness in real-world applications highlights the growing importance of MI-based clustering in data-driven decision-making and knowledge discovery. A new information-theoritical distance measure for evaluating community detection algorithms was introduced in [25].

These contributions collectively reinforce the role of MI in clustering, from optimizing objective functions to handling complex, structured datasets. As research continues, integrating MI-based clustering with deep learning and representation learning frameworks remains a promising direction for uncovering intricate patterns in high-dimensional data.

Regularization and Neural Networks

KL-Divergence plays a central role in *generative models* such as *Variational Autoencoders*, which are used to generate new data samples by learning the latent structure of the data. In this context, KL-divergence is used to regularize the latent space by ensuring that the learned distribution (the approximate posterior) is close to the prior distribution. The KL-divergence regularization term encourages the latent variable distribution to resemble a standard Gaussian distribution, promoting generalization and preventing overfitting [7]. By minimizing KL-divergence, the model ensures that the learned latent representations are smooth and continuous, allowing for better generation of new data samples and improved model robustness. Beyond generative models, MI and IB principles have also been explored as regularization techniques for deep learning. [8] introduced an information-theoretic analysis of Deep Neural Networks, showing that training consists of two key phases: an initial *empirical risk minimization* phase, followed by a *compression phase*, where MI between the input and the hidden layers is gradually reduced. This

compression process aligns with the (IB) principle, acting as a form of implicit regularization. Their findings provide theoretical support for why deep networks generalize well despite overparameterization, suggesting that MI-based constraints naturally shape the learning dynamics. Expanding on this, [26] proposed a framework for *learning deep representations by maximizing MI* between input data and learned representations. Their method, *Deep InfoMax (DMI)*, uses contrastive learning objectives to estimate MI and enforce high-information content in learned representations. Unlike traditional supervised learning, which relies on external labels, DMI ensures that learned features are task-relevant while filtering out noise. This MI maximization strategy has proven effectiveness in improving self-supervised learning, domain adaptation, and robust feature extraction, reinforcing the growing role of *information-theoretic constraints in deep learning regularization*. Cross-entropy remains the standard loss function for optimizing classification tasks, ensuring that models align their predicted probability distributions with true labels to achieve accurate predictions [27]. Together, these information-theoretic measures (KL-Divergence**,** MI and Cross-Entropy) serve as fundamental tools in deep learning regularization, helping models generalize, reduce overfitting, and learn meaningful representations.

The applications of IT in ML are both diverse and fundamental. Core concepts, such as entropy, MI, KL-divergence and Cross-Entropy, underpin a variety of crucial tasks in ML, from feature selection and decision-making to unsupervised learning and generative modeling.

Metric and Deep Learning

MI and other information-theoretic measures play a fundamental role in *Metric Learning and Deep Learning*, guiding how models learn structured and generalizable representations. By leveraging entropy, divergence measures, and the IB principle, researchers have developed techniques, that enhance similarity learning, privacy-aware learning, and transfer learning.

A foundational contribution in *metric learning* comes from [28], where *Information-Theoretic Metric Learning (ITML)* was introduced. Their method optimizes a Mahalanobis distance metric by minimizing *differential entropy*, ensuring that similar points are pulled closer while maintaining constraints on dissimilarity. Unlike traditional distance-learning approaches, ITML leverages relative entropy constraints, making it more robust in high-dimensional feature spaces. This approach has influenced a range of applications, from face verification to text similarity measurement. Privacy concerns in deep learning have led to the development of information-theoretic frameworks that balance data utility and confidentiality.

[29] proposed a privacy-aware time-series data-sharing framework using *Deep Reinforcement Learning*. Their approach formulates data sharing as an optimization problem, where the agent learns an optimal information disclosure policy under privacy constraints. By integrating MI constraints, the model selectively reveals useful data while minimizing privacy risks, demonstrating its effectiveness in financial and healthcare applications.

The theoretical foundations of *Information-Theoretic Learning (ITL)* were established in [30], introducing a framework for learning based on entropy and divergence measures rather than traditional statistical learning methods. ITL provides a more general approach to feature selection, clustering, and kernel methods, making it a precursor to modern information-based deep learning models. The use of *Renyi entropy* and *Cauchy-Schwarz divergence* in ITL offers an alternative to classical probability-based learning techniques, leading to more flexible and adaptive models.

Beyond individual learning paradigms, information-theoretic generalization bounds provide insights into the transferability of learned representations. [31] explored the role of MI in Transfer Learning, analyzing how information retained from the source domain affects generalization in the target domain. Results of this work highlight the importance of controlling information flow between layers in deep networks to prevent overfitting while maximizing knowledge transfer. This work establishes upper bounds on transfer learning generalization errors, making it highly relevant for domain adaptation and self-supervised learning.

Together, these studies illustrate the growing intersection between *IT* and *Deep Learning*, demonstrating how MI, entropy, and divergence measures drive advancements in metric learning, privacy-aware learning, and transfer learning. As deep learning models continue to evolve, information-theoretic regularization techniques are expected to play an even greater role in improving model robustness and interpretability.

4. IB Framework Applications in ML

The IB framework, first introduced in [32], has become a fundamental tool in ML by providing a principled approach to optimizing information flow in learning systems. IB offers a way to balance compression and relevance, formalizing the principle as an information-theoretic tradeoff between MI with the input and relevance to the target, ensuring that models retain the most essential information while discarding irrelevant noise. Over the years, IB has been applied across various ML domains, including representation learning, clustering, deep learning, privacy-aware learning, and image processing. The follow-up work [33] further refined the mathematical foundations of IB, emphasizing how different *distortion measures* impact information retention in learning systems. [34] expanded IB's role in representation learning, showcasing IB's effectiveness in enhancing generalization for multi-agent systems. In the context of *deep learning*, in [35], the authors introduced *Deep Variational Information Bottleneck*, which extends IB by incorporating variational inference. This approach has been widely adopted in training robust and generalizable neural networks by enforcing a structured latent space that reduces overfitting and improves generalization. Similarly, in [36] information flow in Deep Neural Networks is explored, demonstrating how IB principles guide the learning process by distinguishing between representation compression and task-relevant information. In [37], IB is further analyzed for application in Convolutional Neural Networks, optimizing feature extraction and regularization. In [38], the authors explored IB for splitting composite neural networks, improving model modularity and efficiency.

The IB framework has also found extensive applications in *image processing*. In [39], IB is applied to *image segmentation*, optimizing feature selection for improved segmentation accuracy. In [40], the authors introduced the *Residual Bottleneck Dense Network* for image super-resolution, demonstrating how IB-based architectures enhance high-resolution image synthesis. In [41], IB is explored for compressed sensing image reconstruction, leveraging IB principles to enhance the quality of reconstructed images in resource-constrained environments. IB's role in *5G-LDPC decoding with coarse quantization* is examined in [42], improving information retention in error-correcting code applications. Additionally, in [43], Exponential IB Theory is applied to pedestrian attribute recognition, optimizing robustness against intra-attribute variations.

Beyond vision-related tasks, the IB principle has been successfully applied to a range of other domains, including *clustering and feature selection* [44],[45],[46],[47], *geospatial learning* [48], and *multimodal natural language processing* [49]. The IB framework has also been utilized in *speech and audio processing* [50],[51],[52], as well as in *environmental monitoring* and *time-series analysis* [53]**,** while continuing to play a central role in self-supervised visual representation learning [54].

In privacy-aware ML, IB has been utilized to balance data utility and confidentiality. A *Privacy-Aware Joint Source-Channel Coding* method based on *Disentangled IB* is introduced in [55], optimizing secure data transmission. Similarly, in [56], the authors proposed *FIBNet*, demonstrating how IB can prevent leakage of sensitive attributes while retaining necessary identification information. In [57], *Robust IB feature extraction* is explored, enhancing adversarial robustness in ML models.

Several additional contributions have extended the application of the IB framework across diverse ML domains. In *reinforcement learning* and *decision-making*, *Collaborative* [58] and *Two-Way Cooperative* [59] IB frameworks were introduced to optimize multi-agent systems under information-theoretic constraints. In the context of *scheduling* and *optimization*, an IB-based heuristic for job-shop scheduling is proposed in [60], demonstrating IB's utility in large-scale combinatorial problems. In [61], the authors applied *tunable IB with Rényi measures* to improve fairness and interpretability in classification tasks.

As IB research continues to evolve, its applications across deep learning, clustering, privacy, and reinforcement learning highlight its broad impact in ML. Future directions include integrating IB with large-scale self-supervised learning and enhancing IB-based optimization techniques for more efficient model training. The increasing adoption of IB principles underscores its importance as a fundamental tool for structured and efficient learning in ML. For more details on this topic, we refer to a comprehensive survey [62].

5. Challenges and Limitations of IT in ML

While IT has significantly contributed to the advancement of ML, its practical application is not without challenges. Techniques using entropy, MI, and KL-divergence offer powerful tools for managing uncertainty, optimizing models, and guiding decision-making. However, as ML models scale to handle ever-increasing amounts of data and complexity, several challenges emerge.

One key limitation is the *scalability* of information-theoretic measures, particularly when applied to high-dimensional datasets. Computing metrics like MI or entropy often becomes computationally expensive as the dimensionality of the data increases. For example, in [63] authors introduced *MINE (Mutual Information Neural Estimation),* a scalable method for estimating MI by using gradient descent over neural networks. While *MINE* improves scalability, it still faces computational challenges when applied to extremely large datasets or high-dimensional input spaces, requiring efficient optimization techniques to ensure the model doesn’t become prohibitively slow.

Another challenge is *Approximation errors*, as noted in [64], estimating MI accurately is difficult in practice, especially for continuous variables. MI is sensitive to the quality of the

probability distribution estimates, and small errors in density estimation can lead to significant misestimation of MI values.

Despite these challenges, efforts to address the limitations of IT in ML are ongoing. Researchers are continuously exploring ways to improve the scalability and accuracy of information-theoretic measures, particularly in high-dimensional spaces. For instance, advancements in approximation techniques, such as neural estimation methods like *MINE*, provide a promising foundation for mitigating computational constraints. Additionally, adaptive models that can handle noisy and imbalanced data more effectively, such as the *IB* framework, continue to evolve.

Moving forward, future work will likely focus on refining these methods to better suit real-world datasets, particularly those characterized by non-stationarity and high dimensionality. By developing more robust estimation techniques and improving the adaptability of models in dynamic environments, researchers can further harness the power of IT to unlock its full potential in ML.

6. Conclusion

This survey has highlighted the critical role that IT plays in ML, providing a framework for managing uncertainty, optimizing models, and improving decision-making. Through the use of concepts like entropy, MI, and KL-divergence, information-theoretic approaches have enhanced various ML tasks. However, challenges such as scalability, approximation errors, and dependency on accurate data modeling remain key obstacles.

Addressing these issues through ongoing research and improved techniques will help unlock the full potential of IT in ML, driving future innovations and making models more robust and adaptable to complex, real-world problems.

References

[1] C. E. Shannon, “A Mathematical Theory of Communication”, *Bell System Technical Journal*, vol. 27, no 3, pp. 379-423, 1948.

[2] J. Rissanen, “Modeling by shortest data description”, *Automatica*, vol. 14, no. 5, pp. 465-471, 1978.

[3] E. T. Jaynes, “IT and statistical mechanics”, *Physical Review*, vol. 106, no. 4, pp. 620-630, 1957.

[4] N. Tishby and N. Zaslavsky, “Deep learning and the information bottleneck principle”, *IEEE IT Workshop*, Jeju Island, Korea, pp. 1-5, 2015.

[5] V. Mnih, K. Kavukcuoglu and D. Silver, “Human-level control through deep reinforcement learning”, *Nature*, vol. 518, pp. 529-533, 2015.

[6] H. Peng, F. Long and C. Ding, “Feature selection based on MI: Criteria of max-dependency, max-relevance, and min-redundancy”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.

[7] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes”, *International Conference on Learning Representations*, vol. 1, 2013.

[8] R. Shwartz-Ziv and N. Tishby, “Opening the black box of Deep Neural Networks via Information”, *arXiv abs/1703.00810*, 2017.

[9] J. Biamonte, “Quantum Machine Learning”, *Nature*, vol. 549, pp. 195-202, 2017.

[10] E. Haroutunian, M. Haroutunian and A. Harutyunyan, “Reliability criteria in information theory and in statistical hypothesis testing”. *Foundations and Trends in Communications and Information Theory*, vol 4(2-3), pp. 97-263, 2007.

[11] T. M. Cover and J. A . Thomas, *Elements of information theory*, Second Edition, Wiley, New York, 2006.

[12] A. L. Berger, V. J. D. Pietra and S. A. D. Pietra, “A maximum entropy approach to natural language processing”, *Computational Linguistics*, vol. 22, no. 1, pp. 39-71, 1996.

[13] F. Fleuret, “Fast binary feature selection with conditional mutual information”, *Journal of Machine Learning Research*, vol. 5, pp. 1531-1555, 2004.

[14] H. H. Yang and J. Moody, “Data visualization and feature selection: new algorithms for nongaussian data”, *Advances in Neural Information Processing Systems*, vol. 12, 2000.

[15] M. V. Naquet and S. Ullman, “Object recognition with informative features and linear classification”, *Ninth IEEE International Conference on Computer Vision (ICCV)*, pp 281-288, 2003.

[16] J. R. Quinlan, “Induction of decision trees”, *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.

[17] T. M. Hong, S. T. Roche and B. Carlson, “Nanosecond anomaly detection with decision trees and real-time application to exotic Higgs decays”, *Nature Communications*, vol. 15, 2024.

[18] L. Liu, R. Chen, X. Liu, J. Su and L. Qiao, “Towards practical privacy-preserving decision tree training and evaluation in the cloud”, *IEEE Transactions on information forensics and security*, vol. 15, pp. 2914-2929, 2020.

[19] M. Haroutunian, D. Asatryan and K. Mastoyan, “Analyzing the quality of distorted images by the normalized mutual information measure”, *Mathematical Problems of Computer Science*, vol. 61, pp. 7-14, 2024.

[20] R. G. Congalton, et al., “Robust possibilistic fuzzy additive partition clustering motivated by deep local information”, *Circuits, Systems, and Signal Processing*, 2024.

[21] I. S. Dhillon, S. Mallela and D. S. Modha “Information-theoretic co-clustering”, *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 89-98, 2003.

[22] M. Wang and F. Sha, “Information theoretical clustering via semidefinite programming”, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 761-769, 2011.

[23] C. Liang and Y. Leng, “Collaborative filtering based on information-theoretic co-clustering”, *International Journal of Systems Science*, vol. 45, no. 3, pp. 589-597, 2012.

[24] C. Bloechl, R. A. Amjad and B. C. Geiger, “Co-clustering via information-theoretic Markov aggregation”, arXiv:1801.00584, 2018.

[25] M. Haroutunian, K. Mkhitaryan and J. Mothe, “A new information-theoritical distance measure for evaluating community detection algorithms”, *Journal of Universal computer science*, vol. 25, no. 8, pp. 887-903, 2019.

[26] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler and Y. Bengio, “Learning deep representations by mutual information estimation and maximization”, *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Massachusetts, USA, 2016.

[28] J. V. Davis, B. Kulis, P. Jain, S. Sra and I. S. Dhillon, “Information-theoretic metric learning”, *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pp. 209–216, 2007.

[29] E. Erdemir, P. L. Dragotti, D Gunduz, “Privacy-aware time-series data sharing with deep reinforcement learning”, *EEE Transactions on Information Forensics and Security*, vol. 16, pp. 389-401, 2021.

[30] J. C. Principe and D. Xu, “An introduction to information theoretic learning”, *Proceedings of the IJCNN'99 International Joint Conference on Neural Networks*, vol. 3, pp. 1783-1787, 1999.

[31] X. Wu, J. H. Manton, U. Aickelin and J. Zhu, “On the generalization for transfer learning: an information-theoretic analysis”, arXiv:2207.05377, 2022.

[32] N. Tishby, F. C. Pereira and W. Bialek, “The information bottleneck method”, *arXiv:physics/0004057*, 1999.

[33] P. Harremoes and N. Tishby, “The information bottleneck revisited or how to choose a good distortion measure”, *2007 IEEE International Symposium on Information Theory*, Nice, France, pp. 566-570, 2007.

[34] M. Vera, P. Piantanida and L. R. Vega, “The role of the information bottleneck in representation learning”, *IEEE International Symposium on Information Theory (ISIT)*, Vail, CO, USA, pp. 1580-1584, 2018.

[35] A. A. Alemi, I. Fischer, J. V. Dillon and K. Murphy, “Deep variational information bottleneck”, *arXiv:1612.00410*, 2019.

[36] R. Shwartz-Ziv, “Information flow in deep neural networks”, *arXiv:2202.06749*, 2022.

[37] J. Li and D. Liu, “Information bottleneck theory on convolutional neural networks”, *Neural Processing Letters*, vol. 53, no. 2, pp. 1385-1400, 2021.

[38] S. Han, K. Nakamura and B. Hong, “Splitting of composite neural networks via proximal operator with information bottleneck”, *IEEE Access*, vol. 12, pp. 157-167, 2024.

[39] A. Bardera, J. Rigau, I. Boada, M. Feixas and M. Sbert, “Image segmentation using information bottleneck method”, *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1601-1612, 2009.

[40] Z. An, J. Zhang, Z. Sheng, X. Er and J. Lv, “RBDN: Residual bottleneck dense network for image super-resolution”, *IEEE Access*, vol. 9, pp. 103440-103451, 2021.

[41] B. Lee, K. Ko, J. Hong, B. Ku and H. Ko, "Information bottleneck measurement for compressed sensing image reconstruction", *IEEE Signal Processing Letters*, vol. 29, pp. 1943-1947, 2022.

[42] M. Stark, L. Wang, G. Bauch and R. D. Wesel, "Decoding rate-compatible 5G-LDPC codes with coarse quantization using the information bottleneck method", *IEEE Open Journal of the Communications Society*, vol. 1, pp. 646-660, 2020.

[43] J. Wu, Y. Huang, M. Gao, Z. Gao, J. Zhao, J. Shi and A. Zhang, "Exponential information bottleneck theory against intra-attribute variations for pedestrian attribute recognition", *IEEE Transactions on Information Forensics and Security*, vol. 1, pp. 5623-5635, 2023.

[44] S. Wang, C. Li, Y. Li, Y. Yuan and G. Wang, "Self-supervised information bottleneck for deep multi-view subspace clustering", *IEEE Transactions on Image Processing*, vol. 32, pp. 1555-1567, 2023.

[45] X. Yan, Y. Ye, Y. Mao and H. Yu, "Shared-private information bottleneck method for cross-modal clustering", *IEEE Access*, vol. 7, pp. 36045-36056, 2019.

[46] Z. Liu, X. Wang, X. Huang, G. Li, K. Sun and Z. Chen, "Incomplete multi-view representation learning through anchor graph-based GCN and information bottleneck", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pp. 71130-7134, 2024.

[47] X. Yan, Y. Mao, Y. Ye and H. Yu, "Cross-modal clustering with deep correlated information bottleneck method", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 10, pp. 13508-13522, 2024.

[48] K. Yang, W. Tai, Z. Li, T. Zhong, G. Yin, Y. Wang and F. Zhou, "Exploring self-explainable street-level IP geolocation with graph information bottleneck", *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, pp. 7270-7274, 2024.

[49] S. Cui, J. Cao, X. Cong, J. Sheng, Q. Li, T. Liu and J. Shi, "Enhancing multimodal entity and relation extraction with variational information bottleneck", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1274-1285, 2024.

[50] T. Gu, G. Xu and J. Luo, "Sentiment analysis via deep multichannel neural networks with variational information bottleneck", *IEEE Access*, vol. 8, pp. 121014-121021, 2020.

[51] Z. Wu and S. King, "Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training", *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, 2016.

[52] S. H. Lee, H. R. Noh, W. J. Nam and S. -W. Lee, "Duration controllable voice conversion via phoneme-based information bottleneck", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1173-1183, 2022.

[53] C. Wang, S. Du, W. Sun and D. Fan, "Self-supervised learning for high-resolution remote sensing images change detection with variational information bottleneck", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 5849-5866, 2023.

[54] X. Liu, Y. l. Li and S. Wang, "Learning generalizable visual representations via self-supervised information bottleneck", *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, pp. 5385-5389, 2024.

[55] L. Sun, C. Guo, M. Chen and Y. Yang, "Privacy-aware joint source-channel coding for image transmission based on disentangled information bottleneck", *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, pp. 9016-9020, 2024.

[56] Z. Chen, Z. Yao, B. Jin, M. Lin and J. Ning, "FIBNet: Privacy-enhancing approach for face biometrics based on the information bottleneck principle", *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 8786-8801, 2024.

[57] A. Pensia, V. Jog and P. L. Loh, "Extracting robust and accurate features via a robust information bottleneck", *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 131-144, 2020.

[58] M. Vera, L. R. Vega and P. Piantanida, "Collaborative information bottleneck", *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 787-815, 2019.

[59] M. Vera, L. Rey Vega and P. Piantanida, "The two-way cooperative information bottleneck", *2015 IEEE International Symposium on Information Theory (ISIT)*, Hong Kong, China, pp. 2131-2135, 2015.

[60] Z. Yan, G. Hanyu and X. Yugeng, "Modified bottleneck-based heuristic for large-scale job-shop scheduling problems with a single bottleneck", *Journal of Systems Engineering and Electronics*, vol. 18, no. 3, pp. 556-565, 2007.

[61] A. Gronowski, W. Paul, F. Alajaji, B. Gharesifard and P. Burlina, "Classification utility, fairness, and compactness via tunable information bottleneck and Rényi measures", *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 1630-1645, 2024.

[62] Z. Goldfeld, Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning", *IEEE Journal On Selected Areas In Information Theory*, vol. 1, no. 1, pp. 19-38, 2020.

[63] M. I. Belghazi, "Mutual information neural estimation (MINE)", *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, pp. 1391-1400, 2018.

[64] D. McAllester and K. Stratos, "Formal limitations on the measurement of mutual information",*The Proceedings of AISTATS 2020*, pp. 875-884, 2020.

Ինֆորմացիայի տեսության գործիքներն ու տեխնիկաները մեքենայական ուսուցման մարտահրավերների հաղթահարման համար

Մարիամ Ե. Հարությունյան և Գոռ Ա. Ղարագյոզյան

ՀՀ ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտ, Երևան, Հայաստան
e-mail: armar@sci.am, gor.gharagyozyan@edu.isec.am

Ամփոփում

Այս հոդվածում ուսումնասիրվում է Ինֆորմացիայի տեսության լայն կիրառությունները մեքենայական ուսուցման մեջ՝ ընդգծելով, թե ինչպես են հիմնական հասկացությունները օգտագործվում ուսուցման ալգորիթմները բարելավելու համար: Ինֆորմացիայի տեսության գործիքները ինտեգրվել են մեքենայական ուսուցման տարբեր ճյուղերում, այդ թվում՝ նեյրոնային ցանցերում: Մասնավորապես, Ինֆորմացիոն խցանի մեթոդը առաջարկում է պատկերացումներ տվյալների ներկայացման և «Ամրապնդող ուսուցման» վերաբերյալ, որտեղ էնտրոպիայի վրա հիմնված մեթոդները բարելավում են հետազոտության ռազմավարությունները: Ավելին, փոխադարձ ինֆորմացիան կենտրոնական դեր է խաղում չկառավարվող ուսուցման և հատկանիշների ընտրության խնդիրներում: Տրամադրելով և՛ վերջին տարիների արդյունքները, և՛ ժամանակակից հետազոտությունների միտումները, այս հոդվածը կապում է հիմնարար տեսությունը ժամանակակից մեքենայական ուսուցման մեջ իր գործնական իրականացման հետ: Քննարկվում են նաև բաց հարցերը և ապագա ուղղությունները, ինչպիսիք են՝ մասշտաբայնությունը, մեկնաբանելիությունը՝ ընդգծելով այս մեթոդների աճող կարևորությունը նոր սերնդի մոդելներում:

Բանալի բառեր՝ ինֆորմացիոն խցան, նեյրոնային ցանցեր, էնտրոպիայի հիմքով կարգավորում, փոխադարձ ինֆորմացիա, հատկանիշների ընտրություն, Կուլբակ-Լեյբլերի տարամիտություն:

Инструменты и техники теории информации для преодоления вызовов машинного обучения

Мариам Е. Арутюнян и Гор А. Карагёзян

Институт проблем информатики и автоматизации НАН РА, Ереван, Армения
e-mail: armar@sci.am, gor.gharagyozyan@edu.isec.am

Аннотация

В данной статье рассматривается широкое применение Теории Информации в Машинном Обучении, подчеркивается, как основные понятия используются для улучшения алгоритмов обучения. Техники Теории информации были интегрированы в различные подполя Машинного Обучения, включая нейронные сети. В частности метод Информационной пробки дает представление о данных, и обучение с подкреплением, где методы, основанные на энтропии, улучшают стратегии поиска. Кроме того, такие величины, как взаимная информация, имеют решающее значение для отбора признаков и обучения без контроля. Предоставляя последние достижения и обзор современных тенденций, эта статья связывает фундаментальную теорию с ее практической реализацией в современном машинном обучении. Мы также обсуждаем открытые проблемы и будущие направления, такие как масштабируемость, интерпретируемость, подчеркивая растущую важность этих методов в моделях нового поколения: масштабируемость, интерпретируемость, подчеркивая растущую важность этих методов в моделях нового поколения.

Ключевые слова: информационная пробка, нейронные сети, регуляризация на основе энтропии, взаимная информация, выбор характеристик, КЛ-дивергенция .