Novelty Assessment Report

Paper: DISCO: Diversifying Sample Condensation for Accelerating Model Evaluation
PDF URL: https://openreview.net/pdf?id=SoOgBHa3dZ
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Evaluating modern machine learning models has become prohibitively expensive. Benchmarks such as LMMs-Eval and HELM demand thousands of GPU hours per model. Costly evaluation reduces inclusivity, slows the cycle of innovation, and worsens environmental impact. To address the growing cost of standard evaluation, new methods focused on efficient evaluation have started to appear. The typical approach follows two steps. The typical approach follows two steps. First, select an anchor subset of data. Second, train a mapping from the accuracy on this subset to the final test result. The drawback is that anchor selection depends on clustering, which can be complex and sensitive to design choices. We argue that promoting diversity among samples is not essential; what matters is to select samples that maximise diversity in model responses. Our method, **Diversifying Sample Condensation (DISCO)**, selects the top-k samples with the greatest model disagreements. This uses greedy, sample-wise statistics rather than global clustering. The approach is conceptually simpler. From a theoretical view, inter-model disagreement provides an information-theoretically optimal rule for such greedy selection. **DISCO** shows empirical gains over prior methods, achieving state-of-the-art results in performance prediction across MMLU, Hellaswag, Winogrande, and ARC.

Core Task Landscape

This paper addresses: **Efficient Model Performance Evaluation Through Sample Selection**

A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Strategic Sample Selection for Training Data Efficiency**
- **Efficient Evaluation Through Sample Reduction**
- **Model Selection and Output Ranking**
- **Hyperparameter and Configuration Optimization**
- **Specialized Selection Techniques**
- **Auxiliary Methods and Applications**

Complete Taxonomy Tree

- Efficient Model Performance Evaluation Through Sample Selection Survey Taxonomy
- Strategic Sample Selection for Training Data Efficiency
  - Active Learning and Annotation Budget Optimization
  - Query Strategy Design for Active Learning (5 papers)
    - [13] Active partial label learning based on adaptive sample selection (Yan Li, 2022) View paper
    - [30] Unlabeled data selection for active learning in image classification (Xiongquan Li, 2024) View paper
    - [43] Using active learning methods to strategically select essays for automated scoring (Tahereh Firoozi, 2023) View paper
    - [45] Informative Sample Selection Model for Skeleton-Based Action Recognition With Limited Training Samples (Zhigang Tu, 2025) View paper
    - [47] Active Learning-Based Sample Selection for Label-Efficient Blind Image Quality Assessment (Tianshu Song, 2024) View paper
  - Sample Efficiency Characterization and Prediction (1 papers)
    - [26] To Label or Not to Label: PALM-A Predictive Model for Evaluating Sample Efficiency in Active Learning Models (Nielsen, 2025) View paper
  - Data Curation and Subset Selection for Pretraining
  - Distribution Matching and Importance Resampling (2 papers)
    - [24] Data Selection for Language Models via Importance Resampling (Xie, 2023) View paper
    - [41] Data Mixing Laws: Optimizing Data Mixtures by Predicting Language Modeling Performance (Ye Jiasheng, 2024) View paper
  - Semantic Diversity and Disagreement-Based Selection (2 papers)
    - [10] Efficient Response Generation Strategy Selection for Fine-Tuning Large Language Models Through Self-Aligned Perplexity (Ren Xuan, 2025) View paper
    - [42] Efficient data selection employing Semantic Similarity-based Graph Structures for model training (Maji, 2024) View paper
  - Domain-Specific Sample Optimization (8 papers)
  - [3] Reducing Lithium-Ion Battery Testing Costs Through Strategic Sample Optimization (Z. Almutairi, 2025) View paper
  - [8] Improving the estimation accuracy of alfalfa quality based on UAV hyperspectral imagery by using data enhancement and synergistic band selection strategies (Shuai Fu, 2025) View paper
  - [11] Remote sensing for land cover mapping across Victoria, Australia â a machine learning application (Sabah Sabaghy, 2025) View paper

- [12] Strategic sampling for training a semantic segmentation model in operational mapping: Case studies on cropland parcel extraction (Rui Lu, 2025) View paper
- [14] Impact of Non-Landslide Sample Sampling Strategies and Model Selection on Landslide Susceptibility Mapping (Weijun Jiang, 2025) View paper
- [16] Reducing Annotation Efforts in Electricity Theft Detection Through Optimal Sample Selection (Wenlong Liao, 2024) View paper
- [32] Tools to Predict Unilateral Primary Aldosteronism and Optimise Patient Selection for Adrenal Vein Sampling: A Systematic Review (Elisabeth Ng, 2025) View paper
- [39] The Influence of Non-Landslide Sample Selection Methods on Landslide Susceptibility Prediction (Yu Fu, 2025) View paper

- Efficient Evaluation Through Sample Reduction
  - Benchmark Evaluation Acceleration
  - Greedy Sample Selection for Performance Prediction ★ (1 papers)
    - [0] DISCO: Diversifying Sample Condensation for Accelerating Model Evaluation (Anon et al., 2026) View paper
  - Capability Coverage Maximization (1 papers)
    - [2] EffiEval: Efficient and Generalizable Model Evaluation via Capability Coverage Maximization (Wang Yao-ning, 2025) View paper
  - Lifelong Benchmark Evaluation with Model Reuse (2 papers)
    - [4] Efficient lifelong model evaluation in an era of rapid progress (Samuel Albanie, 2024) View paper
    - [18] Lifelong benchmarks: Efficient model evaluation in an era of rapid progress (Prabhu, 2024) View paper
  - Adaptive Evaluation and Test-Time Selection
  - Test-Time Adaptation with Sample Selection (2 papers)
    - [1] Efficient Test-Time Model Adaptation without Forgetting (Niu, 2022) View paper
    - [5] Continual Test-time Domain Adaptation via Dynamic Sample Selection (Yanshuo Wang, 2023) View paper
  - Uncertainty-Aware Performance Monitoring (1 papers)
    - [15] Incremental Uncertainty-aware Performance Monitoring with Labeling Intervention (A Koebler, 2024) View paper
- Model Selection and Output Ranking
  - Sample-Aware Model Selection for Inference (1 papers)
  - [25] Enhancing the Power of OOD Detection via Sample-Aware Model Selection (Feng Xue, 2024) View paper
  - Output Selection and Ranking (3 papers)
  - [7] Sample-efficient human evaluation of large language models via maximum discrepancy competition (Kehua Feng, 2025) View paper
  - [17] SoTA with Less: MCTS-Guided Sample Selection for Data-Efficient Visual Reasoning Self-Improvement (Wang XiYao, 2025) View paper
  - [20] Scalable Best-of-N Selection for Large Language Models via Self-Certainty (Zhao Xuan-dong, 2025) View paper
  - Ensemble Selection and Aggregation (1 papers)
  - [27] An Efficient Selective Ensemble Learning with Rejection Approach for Classification (Hao Xu, 2023) View paper
- Hyperparameter and Configuration Optimization
  - Prompt and Generation Strategy Selection (2 papers)
  - [22] Hyperband-based Bayesian Optimization for Black-box Prompt Selection (Schneider, 2024) View paper
  - [33] Enhancing llm-as-a-judge through active-sampling-based prompt optimization (Cheng Zhen, 2025) View paper
  - System Configuration and Hyperparameter Tuning (1 papers)
  - [6] Cost-efficient sampling for performance prediction of configurable systems (t) (Atrisha Sarkar, 2015) View paper
- Specialized Selection Techniques
  - Sample Selection for Model Compression and Distillation (2 papers)
  - [29] Efficient and Effective Data Imputation with Influence Functions (Xiaoye Miao, 2021) View paper
  - [49] EnCoDe: Enhancing Compressed Deep Learning Models Through Feature - - - Distillation and Informative Sample Selection (Rebati Gaire, 2023) View paper
  - Sample Selection for Continual and Incremental Learning (1 papers)
  - [48] Curiosity-Driven Class-Incremental Learning via Adaptive Sample Selection (Qinghua Hu, 2022) View paper
  - Cross-Domain and Transfer Learning Sample Selection (2 papers)
  - [35] Hybrid Knowledge Transfer for Improved Cross-Lingual Event Detection via Hierarchical Sample Selection (Luis Guzman Nateras, 2023) View paper
  - [37] Optimization of transfer learning based on source sample selection in Euclidean space for P300-based brain-computer interfaces (Sepideh Kilani, 2024) View paper
  - Sample Selection for Neural Network Training Acceleration (2 papers)
  - [38] EVOS: Efficient Implicit Neural Training via EVOlutionary Selector (Weixiang Zhang, 2024) View paper
  - [40] Flow Distillation Sampling: Regularizing 3D Gaussians with Pre-trained Matching Priors (Lin-Zhuo Chen, 2025) View paper
  - Sample Selection for Model Debugging and Improvement (1 papers)
  - [21] MODE: automated neural network model debugging via state differential analysis and input selection (Shiqing Ma, 2018) View paper
  - Federated and Distributed Learning Sample Selection (1 papers)
  - [50] Time-constrained federated learning (FL) in push-pull IoT wireless access (VÄ□n PhÃºc BÃ¹i, 2025) View paper
- Auxiliary Methods and Applications (9 papers)
  - [9] Enhancing Predictive Model Performance through Comprehensive Pre-processing and Hybrid Feature Selection: A Study using SVM (Bharadwaj Thuraka, 2024) View paper
  - [19] Strategic integration of adaptive sampling and ensemble techniques in federated learning for aircraft engine remaining useful life prediction (Ancha Xu, 2025) View paper
  - [23] Towards interpretable drug interaction prediction via dual-stage attention and Bayesian calibration with active learning (Rongpei Li, 2025) View paper
  - [28] Highâ□resolution flood probability mapping using generative machine learning with largeâ□scale synthetic precipitation and inundation data (Lipai Huang, 2025) View paper
  - [31] A predictive analytics model for strategic business decision-making: A framework for financial risk minimization and resource optimization (FU Ojika, 2023) View paper
  - [34] Simulation approaches of cost optimization in the estimation of population variance with incomplete data imputation under successive sampling (Anup Kumar Sharma, 2025) View paper

- [36] Presenting the Management Accounting Model in the Digital Era (Abdolkarim Gholami, 2025) View paper
- [44] Performance enhancementâ□□based active learning sample selection method (Zhonghai He, 2022) View paper
- [46] AI-Powered Forecasting and Optimization of Energy Consumption in the USA: Machine Learning Approaches for Sustainable Urban and Institutional Development (Chowdhury, 2025) View paper

Narrative

Core task: efficient model performance evaluation through sample selection. The field addresses how to reduce computational and annotation costs when assessing model quality by strategically choosing which samples to evaluate or train on. The taxonomy reveals six main branches. Strategic Sample Selection for Training Data Efficiency focuses on choosing informative subsets during training—ranging from active learning methods like PALM Active Learning[26] to curriculum-based approaches. Efficient Evaluation Through Sample Reduction targets accelerating benchmark evaluation itself, employing greedy selection or coverage-based strategies such as EffiEval Capability Coverage[2] to predict performance with fewer test samples. Model Selection and Output Ranking deals with choosing among candidate models or outputs, exemplified by Best-of-N Selection[20]. Hyperparameter and Configuration Optimization explores sample-efficient tuning, including bandit-inspired methods like Hyperband Prompt Selection[22]. Specialized Selection Techniques encompasses domain-specific sampling (e.g., Landslide Sample Sampling[14], Battery Testing Optimization[3]), while Auxiliary Methods and Applications covers supporting techniques like influence functions and variance estimation.

Several contrasting themes emerge across these branches. Training-focused selection often emphasizes diversity and informativeness to improve learning efficiency, whereas evaluation-focused selection prioritizes representativeness and correlation with full-benchmark performance. A handful of works, including Efficient Test-Time Adaptation[1] and Continual Test-Time Adaptation[5], bridge training and evaluation by adapting models during inference with minimal samples. DISCO Sample Condensation[0] sits within the Efficient Evaluation Through Sample Reduction branch, specifically under greedy sample selection for performance prediction. Its emphasis on condensing evaluation sets to predict model rankings aligns closely with EffiEval Capability Coverage[2], which also seeks capability-aware subsets, though DISCO's greedy approach contrasts with coverage-based heuristics. Compared to Efficient Lifelong Evaluation[4], which addresses continual benchmarking over time, DISCO focuses on static benchmark acceleration. This positioning highlights an ongoing tension: balancing sample reduction with faithful performance estimation across diverse model families and tasks.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

All three subtopics address efficient model evaluation by reducing the number of test samples or evaluations needed. The original leaf focuses on greedy selection strategies that maximize disagreement or information gain between models to predict benchmark performance. Siblings differ in their selection criteria: one emphasizes comprehensive capability coverage through clustering/anchoring, while another focuses on reusing evaluation results across models in evolving benchmarks.

Similarities: - All aim to reduce computational cost of model evaluation on benchmarks - All involve strategic sample/test selection rather than exhaustive evaluation - All seek to predict or approximate full benchmark performance from partial evaluations

Differences: - Greedy Sample Selection uses inter-model disagreement as the selection criterion, while Capability Coverage uses task dimension coverage and clustering - Greedy methods operate on model-comparative signals (disagreement, information gain), while Capability Coverage focuses on test sample diversity independent of model comparisons - Lifelong Benchmark Evaluation uniquely addresses temporal/dynamic scenarios with continuously expanding benchmarks and leverages cross-model result reuse, while the other two focus on static benchmark acceleration - The original leaf's exclusion of clustering-based methods directly distinguishes it from Capability Coverage's anchor selection approach

Suggested Search Directions: - Hybrid approaches combining greedy disagreement-based selection with capability coverage constraints - Active learning frameworks for benchmark construction that balance information gain and coverage - Transfer of evaluation results across model families or architectures in static benchmarks

Sibling Subtopics

- **Capability Coverage Maximization** (leaves: 1, papers: 1)
- Scope: Approaches that select test samples to maximize coverage of model capabilities or task dimensions while minimizing redundancy.
- Exclude: Excludes greedy disagreement-based methods; those belong to greedy sample selection categories.
- **Lifelong Benchmark Evaluation with Model Reuse** (leaves: 1, papers: 2)
- Scope: Frameworks that leverage previously evaluated models to reduce evaluation cost in continuously expanding benchmarks through dynamic programming or sorting.
- Exclude: Excludes static benchmark acceleration without model reuse; those belong to capability coverage categories.

Contributions Analysis

Overall novelty summary. The paper proposes DISCO, a greedy sample selection method that identifies test samples maximizing inter-model disagreement to predict full benchmark performance with reduced evaluation cost. Within the taxonomy, DISCO resides in the 'Greedy Sample Selection for Performance Prediction' leaf under 'Benchmark Evaluation Acceleration'. Notably, this leaf contains no sibling papers in the current taxonomy, suggesting a relatively sparse research direction. The broader 'Benchmark Evaluation Acceleration' category includes only three leaves, indicating that greedy disagreement-based approaches represent a less crowded niche compared to training-focused sample selection methods.

The taxonomy reveals that DISCO's closest neighbors lie in adjacent leaves: 'Capability Coverage Maximization' (containing EffiEval) and 'Lifelong Benchmark Evaluation with Model Reuse'. While EffiEval emphasizes clustering-based coverage of task dimensions, DISCO adopts a simpler greedy strategy targeting model response diversity. The broader 'Efficient Evaluation Through Sample Reduction' branch contrasts with 'Strategic Sample Selection for Training Data Efficiency', which dominates the taxonomy with active learning and data curation methods. DISCO's focus on test-time efficiency without model modification distinguishes it from adaptive evaluation techniques like test-time adaptation, which belong to a separate subtopic.

Among the three contributions analyzed, the literature search examined 26 candidates total. The core DISCO method (9 candidates examined, 0 refutable) and the model signature framework (7 candidates, 0 refutable) appear relatively novel within the limited search scope. However, the information-theoretic justification for disagreement-based selection (10 candidates examined, 1 refutable) shows overlap with prior work. The analysis indicates that while the algorithmic approach may be distinctive, the theoretical grounding has some precedent among the examined candidates. The absence of sibling papers in DISCO's taxonomy leaf suggests limited direct competition, though the small search scale (26 papers) leaves open the possibility of unexamined related work.

Based on the top-26 semantic matches and taxonomy structure, DISCO appears to occupy a relatively underexplored niche within benchmark evaluation acceleration. The greedy disagreement-based approach contrasts with clustering-heavy methods in neighboring leaves, and the lack of sibling papers suggests limited direct prior work in this specific formulation. However, the analysis covers a

narrow slice of the literature, and the refutable theoretical contribution indicates that some conceptual elements have precedent. A more exhaustive search might reveal additional related efforts in efficient evaluation or active testing domains.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: DISCO method for efficient model evaluation via sample selection

Description: The authors propose DISCO, a method that selects evaluation samples based on inter-model disagreement rather than clustering-based representativeness. This greedy, sample-wise approach simplifies subset selection by focusing on samples that maximize diversity in model responses.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Assessing generalization of SGD via disagreement

URL: View paper

Brief Assessment

SGD Disagreement Generalization[51] focuses on estimating test error via inter-model disagreement without requiring sample selection or subset construction. DISCO addresses a different problem: selecting informative evaluation subsets to reduce computational cost, not estimating generalization from disagreement patterns.

---

2. Querying Easily Flip-flopped Samples for Deep Active Learning

URL: View paper

Brief Assessment

Flip-Flopped Samples[59] focuses on active learning for training data selection based on predictive uncertainty and disagreement metrics, not on efficient model evaluation or benchmark compression for already-trained models.

---

3. A Note on" Assessing Generalization of SGD via Disagreement"

URL: View paper

Brief Assessment

Disagreement Generalization Note[57] examines theoretical properties of inter-model disagreement for estimating test error in a Bayesian framework, but does not propose a practical sample selection method for efficient evaluation like DISCO. The candidate focuses on calibration theory rather than subset selection algorithms.

---

4. Agree to Disagree: Robust Anomaly Detection with Noisy Labels

URL: View paper

Brief Assessment

Robust Anomaly Detection[52] addresses anomaly detection with noisy labels using sample selection and label refurbishment strategies. This is fundamentally different from DISCO's focus on efficient model evaluation through inter-model disagreement-based sample selection for benchmark compression.

---

5. Reward Uncertainty for Exploration in Preference-based Reinforcement Learning

URL: View paper

Brief Assessment

Reward Uncertainty Exploration[54] focuses on exploration in preference-based RL using reward uncertainty, not on efficient model evaluation through sample selection based on inter-model disagreement for benchmark compression.

---

6. TriagedMSA: Triaging Sentimental Disagreement in Multimodal Sentiment Analysis

URL: View paper

Brief Assessment

TriagedMSA Disagreement[53] focuses on multimodal sentiment analysis with sentiment disagreement detection between modalities (text, audio, visual), not on efficient model evaluation through sample selection based on inter-model disagreement for benchmark compression.

---

7. Handling disagreement in hate speech modelling

URL: View paper

Brief Assessment

Hate Speech Disagreement[55] focuses on handling annotator disagreement in hate speech classification tasks, not on efficient model evaluation through sample selection based on inter-model disagreement. The candidate addresses a fundamentally different problem domain (annotation quality and disagreement) rather than evaluation efficiency.

---

8. How does disagreement help generalization against label corruption?

URL: View paper

Brief Assessment

Disagreement Label Corruption[56] focuses on training robust models under label noise using inter-model disagreement to filter corrupted training data, not on efficient evaluation through sample selection for benchmarking purposes.

---

9. Agree to disagree: Diversity through disagreement for better transferability

URL: View paper

Brief Assessment

Diversity Through Disagreement[58] focuses on training diverse model ensembles for OOD generalization and uncertainty estimation, not on efficient benchmark evaluation through sample selection. The disagreement mechanism serves a fundamentally different purpose—promoting feature diversity during training rather than selecting informative evaluation samples.

Contribution 2: Information-theoretic justification for disagreement-based selection

Description: The authors establish that inter-model disagreement, measured via Jensen-Shannon Divergence or Predictive Diversity Score, is information-theoretically optimal for selecting samples that best differentiate and rank models when estimating benchmark performance.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Theory of disagreement-based active learning

URL: View paper

Brief Assessment

Disagreement-Based Active Learning[60] focuses on active learning for label acquisition in supervised learning, where disagreement is used to select which unlabeled samples to query for labels. The original paper uses disagreement for efficient model evaluation by selecting samples that differentiate models' performance predictions, which is a fundamentally different application domain and problem setting.

---

2. A disagreement-based active matrix completion approach with provable guarantee

URL: View paper

Brief Assessment

Disagreement Matrix Completion[68] addresses matrix completion with active learning for missing entries, not model evaluation or benchmark performance estimation through inter-model disagreement.

---

3. DISCO: Diversifying Sample Condensation for Efficient Model Evaluation

URL: View paper

Prior Art Analysis

DISCO Efficient Evaluation[63] demonstrates that inter-model disagreement measured via Jensen-Shannon Divergence or Predictive Diversity Score was already established as information-theoretically optimal for sample selection. The candidate paper presents Proposition 1 proving that mutual information between model index and predictions equals JSD, establishing that samples with greatest JSD convey maximum information for predicting model performance. This theoretical framework, along with the PDS metric and its relationship to JSD through enveloping inequalities (Proposition 2), was already published in DISCO[63], indicating the original paper's claim of being first to establish this information-theoretic optimality is refuted.

Evidence

Evidence 1 - **Rationale**: Both papers claim the same theoretical result about information-theoretic optimality of inter-model disagreement. The identical phrasing indicates DISCO[63] already established this theoretical justification. - **Original**: from a theoretical view, inter-model disagreement provides an information-theoretically optimal rule for such greedy selection - **Candidate**: from a theoretical view, inter-model disagreement provides an information-theoretically optimal rule for such greedy selection

Evidence 2 - **Rationale**: DISCO[63] explicitly proves (Proposition 1) that inter-model disagreement is most informative for ranking models, establishing the information-theoretic foundation that the original paper claims as novel. - **Original**: we prove that inter-model disagreement is the most informative signal for estimating benchmark performance when the goal is to differentiate and rank models (proposition 1) - **Candidate**: we argue that promoting diversity among samples is not essential; what matters is to select samples that $ \textit{maximise diversity in model responses}$

Evidence 3 - **Rationale**: DISCO[63] provides the complete formal proof (Proposition 1) showing mutual information equals JSD, establishing the information-theoretic optimality. This mathematical framework was already published in DISCO[63]. - **Original**: proposition 1.let d="tpx i, yiqun i be a test set and m„unift1, . . . , mu(a1) be the index of a uniformly chosen model. let fm c pxiqpr0,1s be the predictive probability for class c of model fm on input xi. we write pym i for the categorical random variable following catpfm 1 pxiq, . . . , fm c pxi... - **Candidate**: from a theoretical view, inter-model disagreement provides an information-theoretically optimal rule for such greedy selection

Evidence 4 - **Rationale**: DISCO[63] establishes the mathematical relationship between PDS and JSD through enveloping inequalities (Proposition 2), providing the theoretical foundation for using either metric. This demonstrates the information-theoretic framework was already established. - **Original**: proposition 2.denoting pdsi :"pds `py1 i , . . . ,pym i ˇ ,jsd i :"jsd` py1 i , . . . ,pym i ˇ for each samplei, we have m2 ln 2 ppdsi ´1q2 djsd i d' m m´1 logm¨ppds i ´1q - **Candidate**: from a theoretical view, inter-model disagreement provides an information-theoretically optimal rule for such greedy selection

---

4. Committee-Based Sample Selection for Probabilistic Classifiers

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

---

5. Active learning for estimating reachable sets for systems with unknown dynamics

URL: View paper

Brief Assessment

Reachable Sets Learning[62] focuses on active learning for reachable set estimation in control systems using disagreement between oracles (strong vs. weak), not on inter-model disagreement for benchmark performance evaluation. The technical contexts are fundamentally different.

---

6. A Spatial-Spectral Disagreement-Based Sample Selection With an Application to Hyperspectral Data Classification

URL: View paper

Brief Assessment

Spatial-Spectral Disagreement[66] uses disagreement between classifiers for sample selection in hyperspectral data classification, but does not provide information-theoretic optimality proofs or establish connections to Jensen-Shannon Divergence for model evaluation benchmarking.

---

7. Ensemble multiple kernel active learning for classification of multisource remote sensing data

URL: View paper

Brief Assessment

Ensemble Kernel Active[67] uses maximum disagreement-based active learning for sample selection in remote sensing classification, not for model evaluation or benchmark performance estimation. The paper does not provide information-theoretic justification for disagreement measures.

---

8. Self-Supervised Exploration via Disagreement

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

---

9. Training Robust Deep Neural Networks on Noisy Labels Using Adaptive Sample Selection with Disagreement

URL: View paper

Brief Assessment

Adaptive Sample Selection[61] uses disagreement between two networks for robust training with noisy labels, not for information-theoretic optimality in model evaluation or benchmark performance estimation.

---

10. Hybrid Disagreement-Diversity Active Learning for Bioacoustic Sound Event Detection

URL: View paper

Brief Assessment

Hybrid Disagreement-Diversity[65] focuses on bioacoustic sound event detection using committee voting disagreement (between MLP and NN classifiers) combined with diversity sampling. The paper does not establish information-theoretic optimality of inter-model disagreement for sample selection in evaluation contexts, nor does it address benchmark performance estimation or model ranking.

Contribution 3: Model signature-based performance prediction framework

Description: The authors introduce a direct prediction approach using model signatures (concatenated outputs on selected samples) as input to simple metamodels, bypassing the complexity of estimating hidden model parameters required by prior methods like IRT-based approaches.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Performance Modeling and Estimation of a Configurable Output Stationary Neural Network Accelerator

URL: View paper

Brief Assessment

Neural Accelerator Modeling[72] focuses on hardware performance estimation (latency, power, chip area) for neural network accelerators, not on predicting benchmark performance using model output signatures as metamodel inputs.

---

2. Including stochastics in metamodel-based DEM model calibration

URL: View paper

Brief Assessment

Stochastic DEM Calibration[75] focuses on metamodel-based calibration of discrete element method parameters for granular processes, not on predicting benchmark performance using model output signatures as inputs to metamodels for machine learning model evaluation.

---

3. Scaling Laws for Downstream Task Performance in Machine Translation

URL: View paper

Brief Assessment

Translation Scaling Laws[73] focuses on predicting downstream translation performance from pretraining data size using log-laws and power-laws, not on using model output signatures as metamodel inputs. The candidate's approach involves scaling laws based on dataset size, whereas the original contribution uses concatenated model outputs (signatures) as direct inputs to predictors.

---

4. Test selection for deep neural networks using meta-models with uncertainty metrics

URL: View paper

Brief Assessment

Meta-Models Uncertainty Metrics[71] focuses on test data prioritization for identifying DNN weaknesses using uncertainty metrics from base models, not on predicting benchmark performance from model output signatures. The candidate's meta-model predicts whether individual test inputs will cause errors, while the original paper's framework predicts overall benchmark performance from concatenated model outputs.

---

5. Performance predictive metamodel for dynamic facade shading

URL: View paper

Brief Assessment

Facade Shading Metamodel[74] focuses on predicting building illuminance performance from facade shading parameters using machine learning metamodels. The candidate does not address predicting benchmark performance from model output signatures, which is the core novelty of the original paper's contribution.

---

6. Meta-learning and Data Augmentation for Stress Testing Forecasting Models

URL: View paper

Brief Assessment

Stress Testing Forecasting[70] focuses on predicting large forecasting errors using time series features (e.g., linearity, lumpiness) as metamodel inputs, not model output signatures. The candidate addresses stress testing in forecasting models, while the original paper addresses efficient benchmark evaluation using model signatures (concatenated outputs on selected samples).

---

7. Model selection via meta-learning: a comparative study

URL: View paper

Brief Assessment

Meta-Learning Model Selection[76] focuses on selecting appropriate inducers for classification tasks using dataset characteristics as meta-features, not on using model output signatures (concatenated predictions on selected samples) as inputs to metamodels for performance prediction.

Appendix: Text Similarity Detection

Textual similarity detection checked 27 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. DISCO: Diversifying Sample Condensation for Efficient Model Evaluation

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] DISCO: Diversifying Sample Condensation for Accelerating Model Evaluation View paper
- [1] Efficient Test-Time Model Adaptation without Forgetting View paper
- [2] EffiEval: Efficient and Generalizable Model Evaluation via Capability Coverage Maximization View paper
- [3] Reducing Lithium-Ion Battery Testing Costs Through Strategic Sample Optimization View paper
- [4] Efficient lifelong model evaluation in an era of rapid progress View paper
- [5] Continual Test-time Domain Adaptation via Dynamic Sample Selection View paper
- [6] Cost-efficient sampling for performance prediction of configurable systems (t) View paper
- [7] Sample-efficient human evaluation of large language models via maximum discrepancy competition View paper
- [8] Improving the estimation accuracy of alfalfa quality based on UAV hyperspectral imagery by using data enhancement and synergistic band selection strategies View paper
- [9] Enhancing Predictive Model Performance through Comprehensive Pre-processing and Hybrid Feature Selection: A Study using SVM View paper
- [10] Efficient Response Generation Strategy Selection for Fine-Tuning Large Language Models Through Self-Aligned Perplexity View paper
- [11] Remote sensing for land cover mapping across Victoria, Australia â□□ a machine learning application View paper
- [12] Strategic sampling for training a semantic segmentation model in operational mapping: Case studies on cropland parcel extraction View paper
- [13] Active partial label learning based on adaptive sample selection View paper
- [14] Impact of Non-Landslide Sample Sampling Strategies and Model Selection on Landslide Susceptibility Mapping View paper
- [15] Incremental Uncertainty-aware Performance Monitoring with Labeling Intervention View paper
- [16] Reducing Annotation Efforts in Electricity Theft Detection Through Optimal Sample Selection View paper
- [17] SoTA with Less: MCTS-Guided Sample Selection for Data-Efficient Visual Reasoning Self-Improvement View paper
- [18] Lifelong benchmarks: Efficient model evaluation in an era of rapid progress View paper
- [19] Strategic integration of adaptive sampling and ensemble techniques in federated learning for aircraft engine remaining useful life prediction View paper
- [20] Scalable Best-of-N Selection for Large Language Models via Self-Certainty View paper
- [21] MODE: automated neural network model debugging via state differential analysis and input selection View paper
- [22] Hyperband-based Bayesian Optimization for Black-box Prompt Selection View paper
- [23] Towards interpretable drug interaction prediction via dual-stage attention and Bayesian calibration with active learning View paper
- [24] Data Selection for Language Models via Importance Resampling View paper
- [25] Enhancing the Power of OOD Detection via Sample-Aware Model Selection View paper
- [26] To Label or Not to Label: PALM-A Predictive Model for Evaluating Sample Efficiency in Active Learning Models View paper
- [27] An Efficient Selective Ensemble Learning with Rejection Approach for Classification View paper
- [28] Highâ□□resolution flood probability mapping using generative machine learning with largeâ□□scale synthetic precipitation and inundation data View paper
- [29] Efficient and Effective Data Imputation with Influence Functions View paper
- [30] Unlabeled data selection for active learning in image classification View paper
- [31] A predictive analytics model for strategic business decision-making: A framework for financial risk minimization and resource optimization View paper
- [32] Tools to Predict Unilateral Primary Aldosteronism and Optimise Patient Selection for Adrenal Vein Sampling: A Systematic Review View paper
- [33] Enhancing llm-as-a-judge through active-sampling-based prompt optimization View paper
- [34] Simulation approaches of cost optimization in the estimation of population variance with incomplete data imputation under successive sampling View paper
- [35] Hybrid Knowledge Transfer for Improved Cross-Lingual Event Detection via Hierarchical Sample Selection View paper
- [36] Presenting the Management Accounting Model in the Digital Era View paper
- [37] Optimization of transfer learning based on source sample selection in Euclidean space for P300-based brain-computer interfaces View paper
- [38] EVOS: Efficient Implicit Neural Training via EVOlutionary Selector View paper
- [39] The Influence of Non-Landslide Sample Selection Methods on Landslide Susceptibility Prediction View paper
- [40] Flow Distillation Sampling: Regularizing 3D Gaussians with Pre-trained Matching Priors View paper
- [41] Data Mixing Laws: Optimizing Data Mixtures by Predicting Language Modeling Performance View paper
- [42] Efficient data selection employing Semantic Similarity-based Graph Structures for model training View paper
- [43] Using active learning methods to strategically select essays for automated scoring View paper
- [44] Performance enhancementâ□□based active learning sample selection method View paper
- [45] Informative Sample Selection Model for Skeleton-Based Action Recognition With Limited Training Samples View paper
- [46] AI-Powered Forecasting and Optimization of Energy Consumption in the USA: Machine Learning Approaches for Sustainable Urban and Institutional Development View paper
- [47] Active Learning-Based Sample Selection for Label-Efficient Blind Image Quality Assessment View paper
- [48] Curiosity-Driven Class-Incremental Learning via Adaptive Sample Selection View paper
- [49] EnCoDe: Enhancing Compressed Deep Learning Models Through Feature - - - Distillation and Informative Sample Selection View paper
- [50] Time-constrained federated learning (FL) in push-pull IoT wireless access View paper
- [51] Assessing generalization of SGD via disagreement View paper
- [52] Agree to Disagree: Robust Anomaly Detection with Noisy Labels View paper
- [53] TriagedMSA: Triaging Sentimental Disagreement in Multimodal Sentiment Analysis View paper

- [54] Reward Uncertainty for Exploration in Preference-based Reinforcement Learning View paper
- [55] Handling disagreement in hate speech modelling View paper
- [56] How does disagreement help generalization against label corruption? View paper
- [57] A Note on" Assessing Generalization of SGD via Disagreement" View paper
- [58] Agree to disagree: Diversity through disagreement for better transferability View paper
- [59] Querying Easily Flip-flopped Samples for Deep Active Learning View paper
- [60] Theory of disagreement-based active learning View paper
- [61] Training Robust Deep Neural Networks on Noisy Labels Using Adaptive Sample Selection with Disagreement View paper
- [62] Active learning for estimating reachable sets for systems with unknown dynamics View paper
- [63] DISCO: Diversifying Sample Condensation for Efficient Model Evaluation View paper
- [64] Committee-Based Sample Selection for Probabilistic Classifiers View paper
- [65] Hybrid Disagreement-Diversity Active Learning for Bioacoustic Sound Event Detection View paper
- [66] A Spatial-Spectral Disagreement-Based Sample Selection With an Application to Hyperspectral Data Classification View paper
- [67] Ensemble multiple kernel active learning for classification of multisource remote sensing data View paper
- [68] A disagreement-based active matrix completion approach with provable guarantee View paper
- [69] Self-Supervised Exploration via Disagreement View paper
- [70] Meta-learning and Data Augmentation for Stress Testing Forecasting Models View paper
- [71] Test selection for deep neural networks using meta-models with uncertainty metrics View paper
- [72] Performance Modeling and Estimation of a Configurable Output Stationary Neural Network Accelerator View paper
- [73] Scaling Laws for Downstream Task Performance in Machine Translation View paper
- [74] Performance predictive metamodel for dynamic facade shading View paper
- [75] Including stochastics in metamodel-based DEM model calibration View paper
- [76] Model selection via meta-learning: a comparative study View paper