Testing Methodologies for Unsupervised Machine Learning

Protyasha Roy
Advisor: Dr. Sajjan Shiva
Department of Computer Science
University of Memphis

Abstract

Unsupervised learning (UL) is primarily used to discover hidden patterns in datasets, making accurate evaluation crucial. However, studies on testing UL are limited, with only seven papers identified. This work reviews testing techniques for UL compared to supervised (SL) and reinforcement learning (RL). More importantly, we aim to address three key questions:
(i) Are there any testing methodologies in SL and RL that can be used to test UL systems?
(ii) What are the challenges included in testing UL systems?
(iii) How effective are different testing techniques in detecting faults of UL systems?

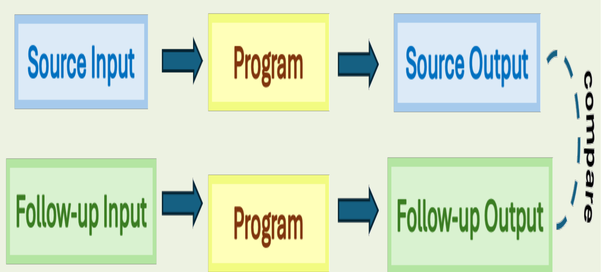
Introduction

- ❑ Lot of applications, such as anomaly detection, natural language processing applications, genetic research, etc. use UL.
- ❑ UL models must be validated to ensure output accuracy and prevent severe consequences.
- ❑ Testing verifies UL models by identifying flaws, enabling corrections and improved performance.
- ❑ Limited research on UL testing. Majority of current research is concerned with various SL testing techniques with few works on RL.
- ❑ Three main testing techniques have been found - symbolic analysis with statistical hypothesis testing, metamorphic testing, and mutation testing

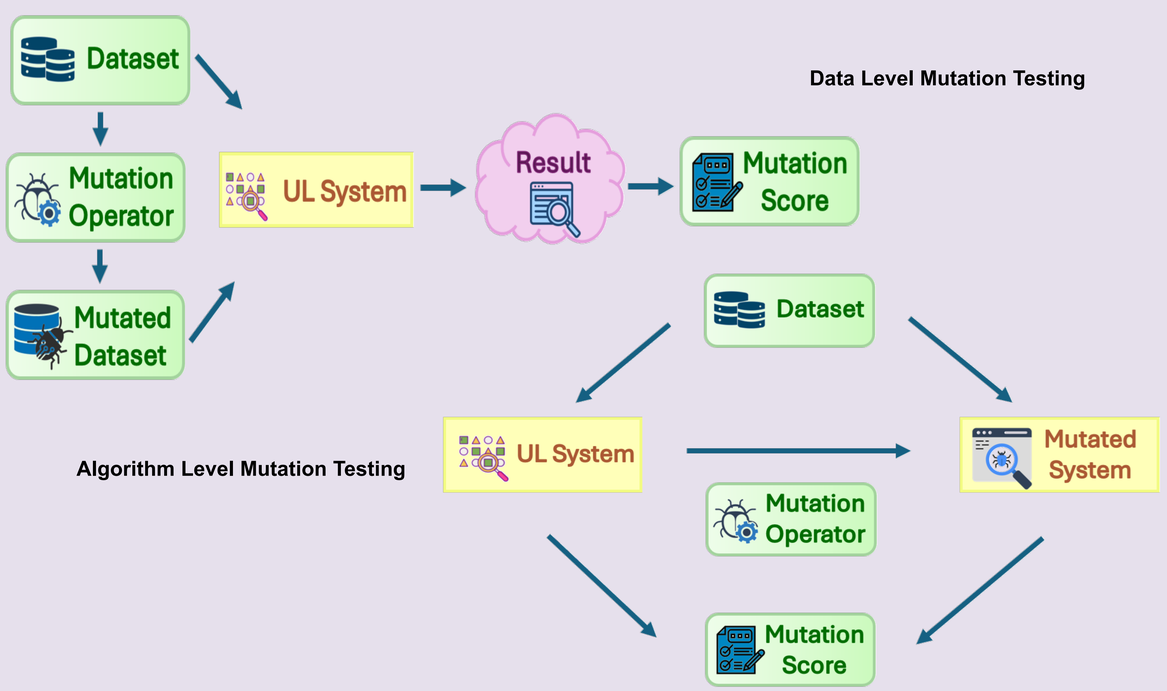
Symbolic Analysis with Statistical Hypothesis Testing

- ➢ A program analysis technique where inputs are treated as symbolic variables instead of actual values.
- ➢ Explores all possible execution paths in a program.
- ➢ Creates path constraints that define conditions for each execution path.
- ➢ Uses a constraint solver to generate test inputs that satisfy the constraints.
- ➢ Hypothesis Testing: a statistical method to determine if a given hypothesis is likely true or due to random chance.

Metamorphic Testing



Mutation Testing



Comparison among SL, UL & RL systems

Topic	Type of Testing	Criteria	SL	UL	RL
Test Input Generation	Symbolic	Design	Combination of the execution of concrete inputs with a symbolic analysis technique	Combination of symbolic analysis and statistical hypothesis technique	
		Implementation	Analyze different behaviors of a DNN	Detect minor bugs in the KM algorithm	
Test Oracle	Metamorphic	Metamorphic Properties	Additive, multiplicative, permutative, invertive, inclusive, and exclusive	Additive, multiplicative, permutative, invertive, inclusive, and exclusive	
		Example of MRs	Adding 10% of new images into the training, validation, and test data set	Add or removing attributes; Duplicating and replacing data instance	
		Validated Algorithms	Classification Algorithms	Clustering Algorithms	
Test Adequacy Evaluation	Mutation	Example of Mutation Operators	Duplicate training data, Add noise to training data, Falsify data findings; Modify labels, Remove training data	Add or delete input data, Reduce randomness, Inject noise, Change gradient of model	Modify the rewards, Create wrong state-action connection; Add noise to true reward
		Implementation	DL systems that cover both classification and regression problems	Testing clustering system and GAN	Q-Learning, Deep Q-Network (DQN), Advantage Actor-Critics (A2C) and Proximal Policy Optimization (PPO)

Analysis

RQ1: Are there any testing methodologies in SL and RL that can be used to test UL systems?

- ❑ Symbolic execution used in both SL & UL.
- ❑ In metamorphic testing, there are six common properties between SL & UL.
- ❑ In mutation testing, most mutation operators in SL & RL are not applicable for UL

RQ2: What are the challenges included in testing UL systems?

Symbolic Analysis with Statistical Hypothesis Testing
- ❑ Does not fully address high-dimensional feature spaces, relies on approximations, and may overlook certain errors
- ❑ Accepts certain errors at a given confidence level due to ML's stochastic nature, causing output variability.

Metamorphic Testing
- ❑ Do not fully resolve the oracle problem.
- ❑ Effectiveness relies on the quality of the identified MRs and the tests derived from them.
- ❑ Designing effective MRs which capture all possible fault scenarios is a challenge.

Mutation Testing
- ❑ Effectiveness depends on the quality of the test dataset.
- ❑ A high or low mutation score doesn't always reflect test suite quality or fault-detection effectiveness.
- ❑ Challenging to design mutation operators that reflect real-world faults.

RQ3: How effective are different testing techniques in detecting faults of UL systems?

Symbolic Analysis with Statistical Hypothesis Testing
- ❑ Detected even minor bugs in the KM algorithm despite slight input variations.

Metamorphic Testing
- ❑ Address the oracle problem, verify and validate clustering systems; KM shows the most violations.
- ❑ For all the algorithms, rotating the coordinate system causes reassignment of boundary data points and the reason for this is min-max normalization.
- ❑ Other type of violations - Reclustering occurs for noisy objects, Different centroids, etc.

Mutation Testing
- ❑ Data-level mutation testing improves test suites by creating and filtering datasets that significantly impact clustering,
- ❑ Data-level mutation testing detects test data that cause abnormal behavior in GANs.
- ❑ Algorithm level mutation testing evaluates UL system resilience against structural and parameter mutations.

Future Work

- ➢ Applying testing on various UL algorithms, such as Apriori, Singular Value Decomposition (SVD), etc.
- ➢ Developing more effective mutation operators and MRs to identify error occurrence in real-life high-dimensional data samples.