Center for Cognitive Architecture

University of Michigan

2260 Hayward Ave

Ann Arbor, Michigan 48109-2121

TECHNICAL REPORT

CCA-TR-2009-05

Evaluating Evaluations: A Comparative Study of Metrics for Comparing Learning Performances

Investigators

Nicholas A. Gorski
John E. Laird

December 22, 2009

Abstract

One of the challenges in comparing learning performance across multiple conditions is to develop an appropriate evaluation. We identify four candidate metrics and apply them in an empirical example, providing contrast to differences between the metrics. We propose a set of criteria that learning comparison metrics should satisfy, evaluate the four metrics using the identified criteria, and conclude with a discussion of our findings.

Table of Contents

1. Introduction

In empirical studies of learning systems, performance is most commonly evaluated using domain-specific performance metrics measured quantitatively over time. However, when attempting to evaluate more general characteristics of the learning system, such as how well it transfers what it has learned in one situation to another, domain-specific metrics are insufficient. The focus of this paper is to study quantitative metrics that measure the differences between the performances of learning agents under different experimental conditions, with the aim of developing a set of general criteria for such metrics.

This paper covers four related topics. We begin by reviewing and briefly discussing four metrics that have been applied to analyze differences in learning performances across different conditions. We apply the metrics in an empirical example to demonstrate their characteristics and motivate our evaluation. We propose a set of criteria for evaluating these metrics. We proceed to evaluate the learning comparison metrics using our identified criteria, and conclude with a discussion of the results.

All of the four learning comparison metrics that we evaluate share some common traits, but most notably all four were developed specifically to quantitatively measure *transfer learning* performance. In transfer learning, an agent applies knowledge learned from one set of problems (the *source* problems) to a new set of problems (the *target* problems) sampled from a different distribution. In the control condition, the target problems are solved without the preceding source problems, while in the transfer condition (also called the experimental condition below) the target is preceded by the source, which presumably leads to improved performance on the target through the transfer of knowledge learned in the source. The metrics we are studying were developed to measure the improvement from the control to the transfer conditions so that alternative approaches can be compared. Although all these metrics were specifically developed to measure transfer learning, they can be applied to measure differences in learning performance between learning agents with any *a priori* relationship.

Throughout this paper we refer to two classes of metrics. *Performance metrics* measure performance in the domain, for example observed reward per time step. *Learning comparison metrics* (LC metrics) measure differences in performance metrics across conditions, and this is the class of metrics that we evaluate in this work.

2. Learning Comparison Metrics

The four LC metrics we consider were developed specifically to measure differences in transfer learning performance, including measuring progress in the DARPA Transfer Learning program (DARPA 2005). In section 2.6, we also briefly discuss domain-specific metrics and examine why they result in poorer evaluations than the four metrics that we examine in more detail.

Of these four LC metrics: the first three are similar in that they all compare areas over or under learning curves and involve integrating over time; they are presented in (roughly) increasing order of complexity. The fourth differs from the others in that it integrates over the performance dimension instead of time.

2.1. Transfer Ratio

The transfer ratio is a simple method of comparing the relative change in performance from one learning curve to another (Asadi, Papudesi & Huber, 2006; Morrison et al., 2006). Given two learning curves defined by the functions f_C (the control condition) and f_E (the transfer condition), the transfer ratio is defined as

$$\frac{\int_{t_0}^{t_*} f_E(t)dt}{\int_{t_0}^{t_*} f_C(t)dt}$$

which is the ratio of the areas under each respective curve integrated from time t_0 until the end of the comparison.

Note that the transfer ratio additionally assumes that both learning curves consist of uniformly positive values. If any observed data is less than 0, then the data must be transformed by a constant additive value before comparing learning performances with the transfer ratio.

The transfer ratio can be interpreted as the ratio of performance under the control condition to the ratio of the performance under the experimental condition. Transfer ratios are greater than 1 when the experimental condition outperforms the control, are 1 when there are no differences in learning performances between conditions, and less than 1 when the experimental condition hinders performance. If the task is not learnable for the control condition, the transfer ratio may be undefined depending on details of the performance metric.

2.2. Transfer Regret

Transfer regret was introduced in the DARPA transfer learning program as an LC metric to improve upon the transfer ratio. Transfer regret is not *regret* as commonly used in the general machine learning community (the difference between rewards or errors made by a learner and how well it could have performed), but instead transfer regret is defined as

$$\frac{\int_{t_0}^{t_*} f_E(t) - f_C(t)dt}{(p_{\max} - p_{\min})(t_* - t_0)}$$

where $p_{\max}$ and $p_{\min}$ are the best and worst observed performances, respectively.

Transfer regret measures the ratio of difference in areas between the experimental and control conditions to the area defined by maximum and minimal performances over the same time period, and it can be interpreted as the percentage of the area contained within a bounding box that is between the two learning curves. As transfer regret increases it indicates that the experimental condition is increasingly outperforming the control condition (somewhat counter-intuitively, given its name). By the nature of the ratio that it calculates, transfer regret is bounded: it can be at most 1 (when the experimental condition is optimal and the control condition never improves), is 0 when performances are identical, and at the least -1 (in the opposite situation). Transfer regret is always defined.

2.3. Calibrated Transfer Ratio (CTR)

The calibrated transfer ratio attempts to improve on the transfer ratio by calibrating performances with the optimal performance (Gorski & Laird, 2007). The CTR builds upon alternate formulations of the transfer ratio in which the area *above* the learning curves are compared, as in Mehta et al. (2005). The CTR is defined as

$$1 - \frac{\int_{t_0}^{t_*} p_{opt} - f_E(t)dt}{\int_{t_0}^{t_*} p_{opt} - f_C(t)dt}$$

where p_{opt} is the optimal performance possible in the domain.

The CTR can be interpreted as the percentage of possible improvement that was achieved under the experimental condition. As the performance under the experimental condition approaches optimality, the CTR approaches 1; when both conditions have identical performances, the CTR is 0; when the control condition outperforms the experimental condition, the CTR is negative. When the control condition is identical to the optimal performance, then the CTR is undefined.

2.4. Average Relative Reduction (ARR)

Average relative reduction is unique in that it is the only metric that does not compare areas above or under learning curves; rather, it averages the relative reduction integrated across performances (Dietterich, 2007). The average relative reduction is defined as

$$\frac{1}{p_*^E - p_0^C} \int_{p_0^C}^{p_*^E} 1 - \frac{x_E(p)}{x_C(p)} dp$$

where $x_E(p)$ is the trial on which performance level p was first achieved, and the relative reduction is integrated from the level of initial performance for the control condition to the level of asymptotic performance for the experimental condition.

ARR can be interpreted as the reduction in time needed to achieve a particular performance under the experimental condition as compared to the control condition, averaged across performances. As the experimental condition approaches optimal performance, ARR approaches 1; when both conditions perform identically, it is 0; when the experimental condition outperforms the control, it is negative. It is undefined when the asymptotic performance of the experimental condition is the same as the initial performance of the control condition.

2.5. Assumptions

The four LC metrics that we evaluate in this work share several assumptions.

Central to our study is the assumption that there exists a single quantitative performance metric for the performance tasks. This metric can be almost any measure of performance. Examples include quality of solution, time to solution, accumulated reward, and average reward.

Another common assumption is that the performance metric is increasing (i.e. that higher performances are better); the LC metrics can either be trivially reformulated or performance can be subtracted from 0 to accommodate decreasing performance metrics.

The LC metrics that we evaluate further assume that the researcher will identify the domain of performances being examined. Typical usage involves comparing performances from time 0 to the time that both learning curves reach asymptotic performance; in practice it may be difficult to consistently determine the trial on which performance is near enough to the asymptote in order to end the comparison.

Finally, the performances of machine learners are measured on discrete time steps (e.g. per domain step or trial). Thus, our statements suggesting that *integration* is performed over these curves are imprecise; rather, the practical application of these metrics approximates the integrals through summation.

2.6. Other metrics

There are additional LC metrics that have been applied to measure differences in transfer learning performance. Jump-start is the difference in performance on the very first time step and asymptotic advantage is the difference in performance after both learners have reached asymptotic performance (DARPA, 2005; Lee-Urban et al., 2007; Sharma et al., 2007). Both of these LC metrics evaluate learning only at one point of performance in domain-specific units, limiting their utility for comparison across tasks.

Instead of comparing areas under learning curves using ratios, one might compare the differences in areas by subtracting the area under the control learning curve from that of the transfer condition. The result would be in domain-specific units, but it is otherwise similar to the transfer ratio; thus it has an additional drawback but gains no advantages. Average overall gain is similar:

it averages the differences in performance measured at each time step (Sharma et al., 2007). Again, measurements are made in domain-specific units and it has no benefit over the standard transfer ratio.

3. Empirical Example

In order to illustrate how the LC metrics compare performances we present a simple empirical example. We begin by comparing the performances of two learners in a straightforward example of transfer learning, where the transfer condition involves bootstrapping from Q-values that have been learned in a related task. This is inspired by transfer work such as Taylor, Stone & Liu (2005), although our simple example does not involve any transformation of the value function.

The domain is a simple MDP: a 4x4 episodic grid world where all actions are deterministic, there are four deterministic actions available in each cell of the grid which move the agent in each of the cardinal directions, every action results in a reward of -1 except for an action taken in the goal cell which results in a reward of +6 (which also ends the episode). The agent always starts in the upper-left corner of the grid and the reward cell is adjacent to the lower-right corner of the grid.

In the control condition, the agent begins learning with Q-values initialized to 0. In the transfer condition, the agent is trained on a similar MDP: identical except that the reward cell is in the lower-right corner of the grid. Both learners update Q-values on-policy following a pure greedy policy with a learning rate of 0.5.

Figure 1 plots the mean accumulated reward per episode for both the control and transfer conditions. As expected, the transfer condition performs very well relative to the control condition, as the task that it was trained on prior to the observed task was extremely similar.

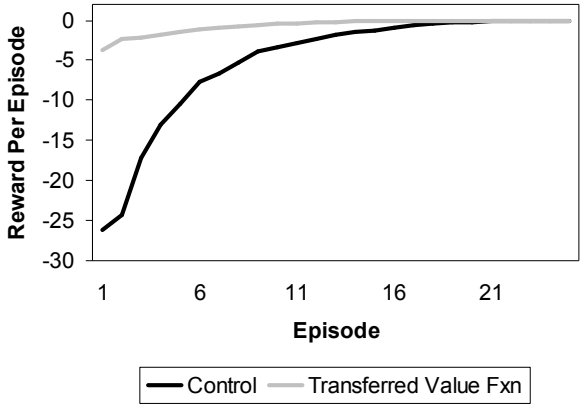

Figure 1: Mean learning curves for control and transfer conditions in the 4x4 grid world task, averaged over 20,000 trials.

Table 1 contains the values of the four LC metrics comparing the performances of the transfer condition to the control. We can interpret these values as follows. The transfer ratio indicates that the area under the learning curve corresponding to the transfer condition is 1.2172 times as large as the area under the control condition learning curve. Transfer regret indicates that the difference in the areas between the two curves was 17.39% of the total area bounding the range and domain of the comparison. CTR indicates that the transfer condition improved performance

by 82.78% of the possible improvement from the control condition to optimal. ARR indicates that the experimental condition required on average 63.39% less training time than the control condition to achieve the same performances.

Table 1: LC metric values comparing performances between the transfer and control conditions in two simple grid-world tasks.

Task	*Transfer Ratio*	*Transfer Regret*	*CTR*	*ARR*
4x4 G.W.	1.2172	0.1739	0.8278	0.6339
3x3 G.W.	1.4205	0.2739	0.7856	0.6236

In order to better understand these values, we modified the domain by decreasing the size of the grid world to be 3x3, identical to the previous domain in other respects except that the terminal reward was changed to +4. The learning agents were identical to those used in the previous task and the methodology remained the same.

While the new task should require less training for the agents to achieve asymptotic optimal performance, we would expect transfer performance to be slightly worse than in the first task: since there are fewer states and the location of the goal cell changes more relative to the size of the grid, it should be slightly more difficult for the agent in the transfer condition to adapt to the new task.

Figure 2 plots the mean learning curves for the control and transfer conditions in the modified task. Although the general shape of the curves is what we would expect, we require a quantitative comparison of the performances in order to draw any sort of comparison to the change in performances on the first task.

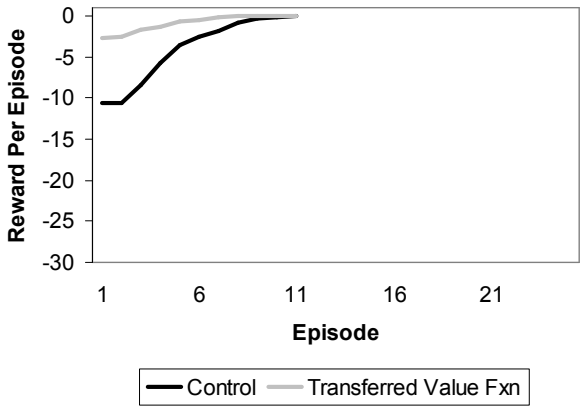

Figure 2: Mean learning curves for control and transfer conditions in the 3x3 grid world task, averaged over 20,000 trials.

As seen in Table 1, the values of the CTR and ARR decrease from the first task to the second, indicating that the transfer condition achieved less improvement in performance on the second task. Note that it is fair to compare the CTR and ARR values across tasks, as they respectively measure the percentage of possible improvement that was achieved and the percentage of average reduction in training time required to achieve the same level of performance.

The transfer ratio and transfer regret, however, cannot be fairly compared across tasks, and indeed their values increase on the second task relative to the first. This does not indicate that the change in performance between conditions was greater on the second task. Rather, they both measure ratios that are affected by characteristics of the domain and reward structure, invalidating such cross-task comparisons, as we will further discuss in the context of our proposed criteria.

4. LC Metric Evaluation Criteria

In developing these criteria, our goal is to identify characteristics of a general LC metric for the evaluation of learning performance across conditions, such that the metric is independent of learning technique, domain, and performance metric. In doing so, we wish to eliminate custom-made evaluation metrics and also to identify an LC metric that supports cross-comparisons of learning performance for learners using different learning techniques, performing in different domains, and being evaluated with different performance metrics.

4.1. Criteria

Below we list the criteria that we have identified to evaluate LC learning metrics. We group them into three categories, as described below. We have named the criteria so as to be succinct, memorable, and accurate.

4.1.1. Value of the Metric

These criteria describe characteristics of the value that is returned by an LC metric.

Defined. An LC metric should return a value for all valid inputs.

Consistent. An LC metric should return a value that is consistent with other returned values. If 0.1 indicates a small amount of positive improvement for the experimental condition and -0.1 indicates a small loss of performance then 0 should indicate no change of performance, and be unambiguous.

Meaningful. An LC metric's value should have a meaningful interpretation. For the LC metrics, meaning is derived from comparison with respect to a standard that is available across tasks and domains.

*Unitless***.** An LC metric's value should not be measured in domain-specific units. Unitless values permit comparisons to be made across tasks and domains.

4.1.2. Characteristics of the Metric

These criteria describe characteristics of how the metric evaluates differences in learning performances.

Distinguishing. An LC metric should distinguish quantitatively between performances of different relative qualities.

*Independent***.** An LC metric should not require domain-specific knowledge in order to measure the differences between observed performances.

Parameter-free. An LC metric should not require parameters that must be set during an experiment and that could influence the calculation of the LC metric.

Similar. An LC metric should not have quantitatively different values when all performances differ only by a scaling factor. If the values of the metric are not similar, then the metric can be

affected by task-dependent ceiling effects; if it is similar, then qualitatively similar performances result in the same quantitative values.

4.1.3. Application of the Metric

These criteria describe characteristics of how the metric may be used.

Degenerate. An LC metric should have a degenerate form that can be applied to measure differences in performance over a single trial (*one-* or *single-shot* performance).

Testable. An LC metric should be statistically testable and permit the construction of confidence intervals.

5. Evaluation

We present here the results of our evaluation of the four selected LC metrics using the proposed criteria. We have grouped the results by criterion to provide maximal contrast between the metrics.

5.1. Defined

Transfer regret is the only LC metric that is always defined. The transfer ratio is undefined when integration of the experimental condition is 0 (a rare occurrence); CTR, when the integrated difference between optimal performance and the control condition is 0 (rare); ARR, when the asymptotic performance of the experimental condition is the same as the initial performance of the control condition (rare). In all of these instances, the metrics are undefined because of division by 0.

5.2. Consistent

Both the CTR and transfer regret are consistent: their values are interpretable without expert knowledge of special cases and are consistent with the intended meaning of the metric. ARR is not consistent in one situation identified in Dietterich (2007). When all performances of the experimental condition are higher than all of the performances of the control condition; then the value of the ARR is 1. A more consistent measure in this case would be positive infinity.

The transfer ratio is not consistent when the integration over the experimental condition is 0 but the integration over the control condition is non-zero: it results in a transfer ratio of 0. In this case, the experimental condition is performing worse than the control condition and therefore a negative value would be more consistent.

5.3. Meaningful

Both the CTR and ARR are meaningful. The CTR is explicitly calibrated in relation to the optimal performance on a task. The notion of optimal performance has meaning across tasks, domains and performance metrics, and by performing this comparison it endows values of the CTR with meaning.

Similarly, ARR is computed by implicitly comparing values to the optimal trial on which a performance could have been achieved. These values are trials on which a performance was achieved, rather than the value of the performance itself. This concept of the optimal trial also has meaning across tasks and domains, which imbues the values of ARR with meaning.

Neither transfer regret nor transfer ratio compare task performances with any standard value. Instead, they compare the integrated performances of the two conditions with each other directly, which certainly has a meaning, but not one that is grounded in terms that apply across tasks, domains or performance metrics. While the two metrics themselves are meaningful comparisons (i.e. one can certainly describe the comparisons that they perform), their values lack standardization

5.4. Unitless

All four LC metrics are unitless. Although this does not distinguish between them, it is an important characteristic of other LC metrics that could potentially be applied to compare performances across performance metrics.

5.5. Distinguishing

Transfer regret is the only LC metric that fails to satisfy the distinguishing criterion. Consider the two sets of hypothetical learning curves in Figure 3. In both cases, the optimal performance is 1 and the worst possible performance is 0. Given these two sets of learning curves, the experimental condition in figure 3b performs better relative to the control condition than the experimental condition in figure 3a. In 3b, the experimental condition approaches the optimal performance almost immediately, and much more quickly than in figure 3a. Whereas the transfer ratio, CTR and ARR would quantitatively differentiate between these two sets of curves, the transfer regret values would be identical in both cases. This failure to distinguish identical between qualitatively different performances creates a confounding effect and creates potential ambiguity in the metrics' values.

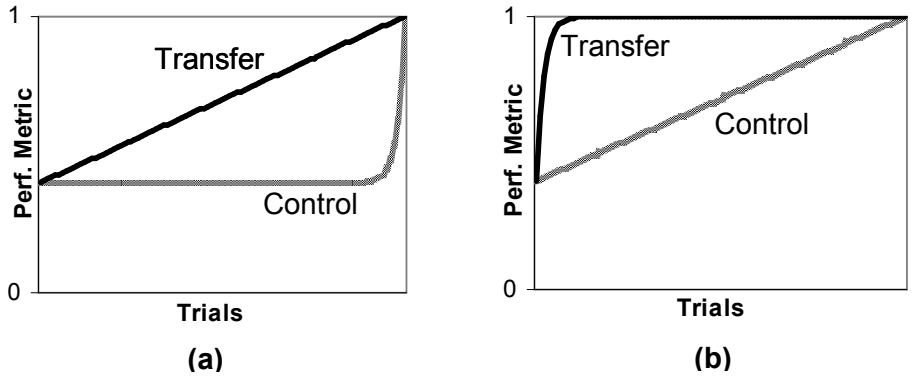

Figure 3: Two sets of hypothetical learning curves illustrating that transfer regret does not satisfy the distinguishing criterion.

5.6. Independent

The CTR fails to satisfy the independent criterion. The CTR requires domain-specific knowledge in the form of the measure of optimal performance. There are two implications of the CTR's failure to be unambiguous. First, it requires additional effort in order to successfully apply it. Second, if the optimal performance is estimated inaccurately, then its measure is inaccurate and cannot be fairly compared across tasks and domains.

As discussed in Gorski & Laird (2007), the CTR requires a measure of optimal performance in the domain. For simple domains (e.g. many hand-coded Markov Decision Processes, or MDPs), and some complex domains (e.g. some single player General Game Player games as in Genesereth, Love & Pell, 2005), it is trivial to analytically determine the optimal performance that is possible; however, in many complex domains it is more difficult.

For tasks where it is non-trivial to determine the optimal performance, there are several options. First, convergence to the optimal policy can be guaranteed in some domains for particular learning agents (e.g. Q-learners in MDPs); in such situations, a trained agent's performance can be measured and used as optimal. Second, if the domain is too complex for convergence to be guaranteed human experts can achieve mastery in it (e.g. the Urban Combat Testbed in Cook, Holder & Youngblood, 2007), then a human expert's performances can be measured and used as an estimate of optimal performance. Third, it may be possible to derive a theoretical upper limit on performance, which can serve as optimal even though it may be an overestimate.

When the optimal performance cannot be determined for a particular task, then the CTR can still be applied, but would no longer satisfy the meaningful criterion. Instead of a measure of optimal performance, the best performance observed under either condition can be used for calibration.

While a domain-specific value is certainly a parameter, we differentiate between independent and parameter-free. A metric is independent when it does not require a particular domain-specific value which is used as a standard for comparison. Although parameters used for measuring performance (discussed below) are grounded in the domain, they are not used as standards for the purpose of comparison.

5.7. Parameter-free

ARR satisfies the parameter-free criterion, while the CTR, transfer regret and the transfer ratio do not. These three metrics integrate over time, usually from the initial trial until learners reach asymptotic performance. The determination of asymptotic performance requires a parameter whose value is important because under- or over-estimating when asymptotic performance is achieved influences the calculation of their values by a small amount.

The CTR satisfies this criterion in situations where both learners approach the optimal possible performance asymptotically. As the CTR is calibrated by optimal performance, misestimating the end of the window of comparison has negligible effect on its value.

5.8. Similar

The transfer ratio is the only LC metric that fails to satisfy the similar criterion. Figure 4 shows a pair of hypothetical learning curves, one pair on a task showing a considerable ceiling effect (figure 4a), and the other pair on a task without a considerable ceiling effect (figure 4b). Applying one of the other three metrics in each of these cases would result in identical values for the comparisons in both figures. However, applying the transfer ratio to these curves will result in two quantitatively different values. This is undesirable as the performances are qualitatively similar given the ceiling effect in the first task. In this case, the transfer ratio would obscure the underlying performances.

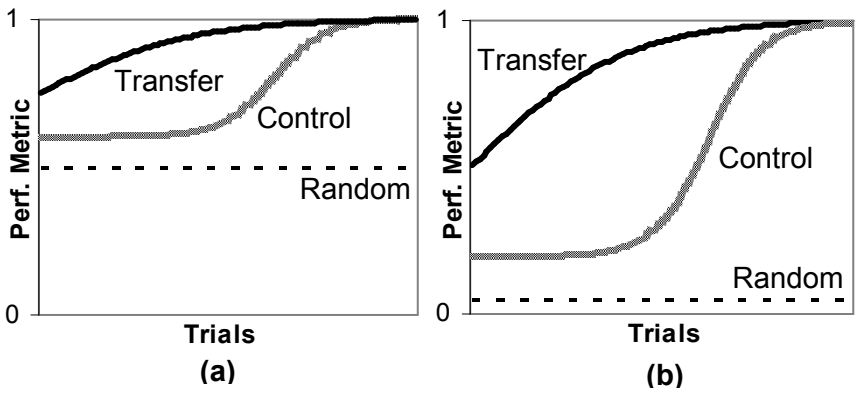

Figure 4: Two sets of hypothetical learning curves illustrating that the transfer ratio does not satisfy the similar criterion.

5.9. Degenerate

The CTR, transfer regret and the transfer ratio all have degenerate forms; ARR does not. For the metrics that have degenerate forms, one simply substitutes instantaneous performance measures in place of integrations within their equations, which then results in a measure of one-shot performance improvement (i.e. performance improvement as observed over a single trial).

Although the ARR can measure instantaneous reduction, it requires two time series from which to extract the time steps on which a particular performance was achieved. Thus, it is only applicable over multiple trials exhibiting a range of different performances. If a researcher is investigating different learning approaches, some which perform over a single trial and others which perform over multiple trials, it can be useful to compare performances using both methodologies.

5.10. Testable

All four metrics are testable. As the LC metrics are complex statistics and the samples are drawn from unknown distributions, calculating measures of statistical confidence requires computer-intensive bootstrap methods to calculate an empirical sampling distribution, which can then be used to test the null hypothesis or construct confidence intervals (Cohen, 1995).

6. Discussion

Table 2 summarizes the results of our evaluation. Our evaluation makes clear that no single LC metric dominates the others according to our criteria. Each is unique with its own advantages and disadvantages, so that different LC metrics are applicable in some situations while others are not. Transfer ratio and transfer regret are applicable to all domains, but satisfy fewer criteria than the other two metrics. The CTR is applicable in domains for which a measure of optimal performance can be analytically determined, accurately measured, or estimated with confidence. ARR, due to what it measures, is applicable to experiments measuring performances over a significant number of time steps and where there is a significant range of performances observed. Note that although a metric is *applicable*, that does not imply that it is *appropriate*.

Table 2: A summarized evaluation of the four LC metrics using the proposed criteria.

Criterion	*Transfer Ratio*	*Transfer Regret*	*Calibrated Transfer Ratio (CTR)*	*Average Relative Reduction (ARR)*
Defined	X	✓	X	X
Consistent	X	✓	✓	X
Meaningful	X	X	✓	✓
Unitless	✓	✓	✓	✓
Distinguishing	✓	X	✓	✓
Independent	✓	✓	X	✓
Parameter-free	X	X	X	✓
Similar	X	✓	✓	✓
Degenerate	✓	✓	✓	X
Testable	✓	✓	✓	✓

6.1. Relative Importance of Criteria

The relative importance of one criterion to another depends on the task at hand and a researcher's goals. In an exploratory investigation involving a single task but multiple learners, the researcher is primarily interested in teasing out differences in performances. The criteria of central concern are thus consistent, distinguishing, and similar – criteria that describe how accurately a metric portrays differences in performance.

In studies comparing performances across tasks, domains or performance metrics, then the meaningful and unitless criteria gain importance – they are criteria that describe necessary characteristics of metrics that allow for such cross-comparisons.

Other criteria may gain relative importance given specific needs. If a study involves measuring performances on single trials, then satisfying the degenerate criterion is necessary. In a formal study requiring a test of statistical significance, the testable criterion is essential.

6.2. LC Metric Comparisons

When comparing the three ratio metrics to each other, the CTR has significant advantages over transfer regret and the transfer ratio. The CTR is meaningful and satisfies both the distinguishing and similar criteria, which make the CTR more comprehensively informative than the others.

The CTR fails to satisfy three of the criteria: defined, independent and parameter-free. The rare circumstances where the CTR is undefined can be prevented by avoiding control conditions that are identical to optimal performances. Although the CTR can be influenced by the setting of parameters, this is only the case when asymptotic performance is significantly different from optimal performance and can be controlled by adopting a single convention for determining when a curve is near-asymptotic. Failing to satisfy the independent criterion is the most serious of the CTR's faults; thus, the CTR is most appropriate in domains for which the optimal performance is easily measured (or estimated with certainty).

ARR fails to satisfy three criteria: defined, consistent, and degenerate. The defined criterion is least significant, as ARR's value will be undefined only in a rare corner case. It has no degenerate form, which precludes its use over very small time series or single trials. It also fails to satisfy the consistent criterion, which requires some knowledge of the metric to interpret ARR values of 1.

ARR is the only metric to satisfy both the independent and meaningful criteria. In applications involving cross-task comparisons in which the optimal performance cannot be measured or estimated, ARR has significant advantages over the other metrics.

There are interesting similarities between the CTR and ARR. They are the only two of the LC metrics to satisfy the meaningful criterion, which is derived from their comparisons with standard values. We noted earlier that the CTR compared against the optimal performance as a standard, while the ARR compared against the optimal time as a standard. Thus, both metrics compare against standards involving optimal, but across different dimensions.

7. Conclusions

The CTR and ARR metrics evaluate most favorably with the identified criteria, while our evaluation suggests that the transfer ratio and transfer regret should be avoided.

The CTR measures improvement in the performance dimension, while the ARR measures reduction in training time. As machine learning evaluations typically measure changes in performance, researchers may prefer this convention; if so, the CTR is preferable to ARR. When the optimal performance in a task is inaccessible, then ARR should be used. Past transfer learning evaluations have involved learning agents applying very different approaches with some capable of performing over a single trial; in such cases, the CTR should be used. The CTR and ARR are very similar, as they both measure differences in learning performance by comparing to a measure of optimal; they are different in that this comparison is across different dimensions.

Although the primary contribution of this paper is the comparative evaluation of the four LC metrics, the criteria themselves are also a contribution. These criteria were constructed to evaluate metrics used to measure differences in learning performance. Although created for this specific purpose, many of the criteria are not specific to differences in learning performance and could apply to evaluations of other types of metrics used within the machine learning community. Evaluations of metrics, specifically comparative evaluations, are uncommon but yet are critical to an informed choice of metric.

References

Asadi, M., Papudesi, V., & Huber, M. (2006). Learning Skill and Representation Hierarchies for Effective Control Knowledge Transfer. *Proceedings of the ICML-06 Workshop on Structural Knowledge Transfer for Machine Learning*. Pittsburgh, PA.

Cohen, P. R. (1995). *Empirical Methods for Artificial Intelligence*. Cambridge, MA: MIT Press.

Cook, D., Holder, L, & Youngblood, G. M. (2007). Graph-based Analysis of Human Level Transfer Using a Game Testbed. *IEEE Transactions on Knowledge and Data Engineering*, *19*(11), 1465-1478.

DARPA (2005). *Transfer Learning* (BAA 05-29).

Dietterich, T. G. (2007). *Proposed Metrics for Transfer Learning.* (Technical Report No. 2007-5). Corvallis, Oregon: School of Electrical Engineering and Computer Science, Oregon State University.

Genesereth, M., Love, N. & Pell, B. (2005). General Game Playing: Overview of the AAAI Competition. *AI Magazine*, *26*(2), 62-72.

Gorski, N. A. & Laird, J. E. (2007). *Investigating Transfer Learning in the Urban Combat Testbed.* (Technical Report No. CCA-TR-2007-02). Ann Arbor, MI: Center for Cognitive Architecture, University of Michigan.

Lee-Urban, S., Muñoz-Avila, H., Parker, A., Kuter, U. & Nau, D. (2007). Transfer Learning of Hierarchical Task-Network Planning Methods in a Real-Time Strategy Game. *Proceedings of the ICAPS-07 Workshop on AI Planning and Learning*. Providence, RI.

Mehta, N., Natarajan, S., Tadepalli, P. & Fern, A. (2005). Transfer in Variable-Reward Hierarchical Reinforcement Learning. *Proceedings of the NIPS 2005 Workshop on Inductive Transfer: 10 Years Later*. Whistler, B.C., Canada.

Morrison, C. T., Chang, Y., Cohen, P. R. & Moody, J. (2006). Experimental State Splitting for Transfer Learning. *Proceedings of the ICML-06 Workshop on Structural Knowledge Transfer for Machine Learning*. Pittsburgh, PA.

Sharma, M., Holmes, M., Santamaria, J., Irani, A., Isbell, C. & Ram, A. (2007). Transfer Learning in Real-Time Strategy Games Using Hybrid CBR/RL. *Proceedings of the 20th International Joint Conference on Artificial Intelligence.* Hyderabad, India.

Taylor, M. E., Stone, P. & Liu, Y. (2005). Value Functions for RL-Based Behavior Transfer: A Comparative Study. *Proceedings of the 20th National Conference on Artificial Intelligence*. Pittsburgh, PA.