Organizational Policy Learning and Evaluation Using Monte Carlo Methods

2

Clement Leung
Chinese University of Hong Kong, China

Nikki Lijing Kuang
University of California, San Diego, USA

Vienne W. K. Sung
Hong Kong Baptist University, Hong Kong

INTRODUCTION

Organizations continually need to develop new policies and undertake policy evaluations in a variety of contexts with a view of determining their feasibility and the possibility of adopting them. Such policies may be internal ones relating to the procedural structures or workflows within the company, or ones that relate to its management or operational strategies, e.g. should the company use a particular advertising channel. A key consideration in the course of policy evaluation is the underlying environment within which the policy learning and evaluation is performed. It is often the case that the dynamics of the environment is of such a level of complexity that the result of each evaluation tends to exhibit significant variability which is governed by a variety of factors of a non-deterministic nature.

In order to systematically determine the feasibility or optimality of a new policy or strategy, it is necessary to observe its performance and monitor the outcomes of repeated trials. Repetitive sequential trials are important and necessary in order to ensure reliability and ongoing effectiveness particularly in an environment which is subject to chance influence and where complete information on all the underlying factors are not available. Such trials need to be undertaken in a methodic manner and they need not be limited to:

1. Evaluations carried out by a single organization; i.e. such evaluations can be undertaken by different stakeholders representing the interests of different organizations
2. Evaluations carried out prior to adoption or acceptance; i.e. such evaluations can continue well beyond the acceptance or delivery phase.

In general, such evaluations may be undertaken with varying degrees of stringency – more stringent criteria should be adopted for safety critical operations, while less stringent ones may be used for non-mission critical policies.

We shall make use of an example particularly relevant to current concerns to illustrate some of these points. Recently there have been heightened global concerns regarding aviation safety, as multiple fatal crashes have occurred across different parts of the world relating to a particular model of aircraft, which took place within the space of only a few months. To be concrete, we shall use Model X to denote the particular model of the aircraft, where the policy in question is to adopt Model X for service. At the

DOI: 10.4018/978-1-7998-3473-1.ch016

very beginning of the process, it starts with a learning stage – one needs to learn whether Model X is sufficiently safe to be in service. Such learning is done by carrying out repeated flights (trials) of the aircraft, with each trial labeled a success or a failure. In the context of reinforcement learning, a success corresponds to a reward, while a failure corresponds to a punishment. Subsequent to the initial stage, in a strict sense learning still continues and never stops as ongoing data are gathered about the safety of Model X. There are three types of organizations involved in the evaluation process, and these include:

1. The manufacturer of Model X
2. The aviation safety certification authority
3. The many client companies that use Model X for its daily operations.

Each flight of Model X may be regarded as an evaluation trial, and hopefully each flight would be uneventful and be labeled as success. Unfortunately, such were not the case for Model X. While the first type of organization concluded from their own pre-delivery trials resulted in overall success and acceptability, such success did not continue to hold subsequent to aircraft delivery. After delivery took place, trials did not stop but continued to be carried out by the client organizations. While these trials were still labeled as success at the very beginning, the policy was put in doubt due to some disastrous outcomes, which resulted in the termination of the policy (i.e. ending the adoption of the Model X for service) and the global grounding of the aircraft.

In most practical situations, the cost of carrying out such learning trials can be significant. We thus see in our example that through repeated trials resulting in outcomes of either reward or punishment, one establishes the feasibility of the new policy and completes the learning phase. Sometimes subsequent to the learning phase, the new policy, if learned successfully (i.e. when the rewards to punishments ratio is sufficiently high), may be adopted from that point onwards without it being questioned or evaluated afterwards. In some situations, one tends to assume that the learning is primarily done during the pre-adoption phase. In most situations, however, even after the policy is adopted, ongoing validation and monitoring is still carried out and this is especially necessary for safety-critical and mission-critical operations, as illustrated in this aviation example. If in the course of ongoing monitoring, there is an overwhelming number of punishments observed, then the adoption of the policy may be called into question, and termination of the policy may be necessary.

In this Chapter, we shall explain how reinforcement learning can be leveraged for organizational policy evaluation, and describe how to deploy a Monte Carlo approach to learn and estimate the underlying structure of the environment by making use of the observations gathered in the course of the evaluation process. We shall provide practical rules for such learning reinforcement scenarios where rewards and punishments for both the pre-adoption phase as well as the post-adoption phase are received. To be concrete, we shall continue to use a scenario of aviation safety, as we believe this scenario is sufficiently general and of particular relevance and currency to present day concerns. Despite this, we wish to point out that many other everyday learning situations are similar to this; examples include trialing a new machine translation algorithm, learning the effectiveness of a new advertising channel, and route discovery in self-drive vehicles.