Empirical Research in Machine Learning

Orlando Ohashi

Map-i

Abstract. This paper explain and describes empirical research, in the context of machine learning and data mining problems. Doing Empirical Research usually involves doing wrong in the first time, take time to explore the problem before formulate a precise experimental procedure. Because of that is usually mistakenly assumed as a synonymous with the experimental research, but in fact Empirical Research is more like to be a combination of Exploratory and Experimental Research, $empirical = exploratory + experimental$. In this paper we overview the main factors of empirical research, and how is better applied to machine learning and data mining tasks in the project, to benefits of this approach in your research.

1 Machine Learning

Machine learning is the subfield of artificial intelligence that is concerned with the design and development of algorithms that allow computers to learn. The learning process is composed of automatically extract rules and patterns from data. Machine learning is closely related to fields such as data mining, statistics, pattern recognition, etc. One of the fields Data Mining is the process of extracting hidden patterns from data. As more data is gathered, with the exponential quantity of data recorded every day, this field is increasingly gain popularity, is becoming an increasingly important tool to transform this data into information, it is commonly used in a wide range of applications, such as marketing, fraud detection and scientific discovery. [6]

Data mining uses machine learning algorithms, to classify, predict, group, recommend, etc. For example, data mining software can help retail companies find customers with common interests. True data mining software doesn't just change the presentation, but actually discovers previously unknown relationships among the data. Other area closed related is visualization techniques.[4]

2 Empirical Research

by Cohen "Behind every experimental or analytical tool is a scientist who views his or her subject through the lens of a particular collection of beliefs."

Moody said "Empirical research methods are a class of research methods in which empirical observations or data are collected in order to answer particular research questions. While primarily used in academic research, they can also be useful in answering practical questions."

A research usually apply different methodology on different kind of problems. Normally empirical research starts with some a priori theory, which is develop to try to explain and predict what happens in a real scenario. Doing empirical research usually involves doing wrong in the first time, take time to explore the problem before formulate a precise experimental procedure; because of that is usually mistakenly assumed as a synonymous with the experimental research, but in fact empirical research is more like as a combination of exploratory and experimental research, $empirical = exploratory + experimental$. [2]

According to the Oxford English Dictionary [1], empiric is derived from the ancient Greek for experience, which is ultimately derived from trial, experiment. Empirical data is information that is derived from the trials and errors of experience. In this way, the empirical research is similar to the experimental research. However, an essential difference is that in an experiment the different "trials" are strictly manipulated so that an inference can be made as to causation of the observed change that results. Experimental research are based on controlled testing of casual processes. The general procedure is one or more independent variables are manipulated to determine their effect on a dependent variable, this process can be applied in empirical research however with less emphasis.

Based on the work Cohen we can distinguish four classes of empirical studies:

- **Exploratory Studies** yield causal hypotheses that are tested in observations or manipulation experiments. To this end, exploratory studies usually collect lots of data, analyzing it in many ways to find regularities;
- **Assessment Studies** establish baselines and ranges, and other assessments of the behaviors of a system or its environments;
- **Manipulation Experiments** test hypotheses about causal influences of factors by manipulating them and noting effects, if any, on more or more measured variables;
- **Observation Experiments** disclose effects of factors on measured variables by observing associations between levels of the factors and values of the variables. These are also called natural and quasi-experimental experiments.

Most people considered manipulation and observation experiments as a proper experiments, and exploratory and assessment as informal, heuristic and hopefull. However this activities are complementary, a research project will involve them all. These can be considered phases of a research project.

Our main goal is apply empirical methods for studying machine learning programs, methods that involve running programs and recording their behaviors. The study of computer programs that perform tasks in environments, can consider every program as an experiment. The machine learning programs are a little different of traditional programs. In machine learning (depends of the algorithm, regression tree is more interpretable than neural networks) is almost impossible to look inside the algorithm and understand and relate the structure to their behavior, and usually is necessary a lot more than a single experiment to statistically demonstrate that the program behaves as we hope.

The most common way to investigate Machine Learning is to embody our hypothesis in programs, and gather data running this programs [5]. Usually doing

this process more than once, the progress of the experiments depends if our hypothesis is falsify or not. These programs must be capable of behavior not expected by the experiment and the formulation of general theories relating behavior to architecture, tasks, and environments is our goal.

The importance of adopt some methodology in the research project it's clear, however not every research do. Many research's in artificial intelligence and computer science speak casually of experiments, as if any activity that involves building and running a program is experimental. For example, when Cohen surveyed 150 papers in the Proceedings of the Eighth National Conference on Artificial Intelligence (1990), The author discovered that:

- only 42% of the papers suggested a program had run on more than one example;
- just 30% demonstrated performance in some way;
- 21% framed hypothesis or made predictions;
- Almost nobody embodies hypothesis in programs, or gathers data by running the programs;
- Very few papers reported negative or unexpected results.

3 Empirical Method

by Douglas Lenat "...we suffer poverty of uncover of imagination; it is thus much easier for us to uncover than to invent..."

The empirical method is generally characterized by the collection of a large amount of data before much speculation as to their significance, or without much idea of what to expect, and is to be contrasted with more theoretical methods in which the collection of empirical data is guided largely by preliminary theoretical exploration of what to expect. Empirical methods, help us find general features by studying specific programs

Three basic research questions

- How will a change in the agent's structure affect its behaviors given a task and an environment ?
- How will a change in an agent's task affect its behavior in a particular environment ?
- How will a change in an agent's environment affect its behavior on a particular task ?

How will a change in the agent structure affect its behavior given a task and an environment ? suppose that training a neural network, how will be the impact of change when some parameters, like the range of random weight initiation ? A typical approach is run the two version of the network (old configuration and with new range weight) and compare the results. How a change in the task affect the result ? Imagine a smart robot (simulation or not) change the task of find the gold for cleaning could have different results in the behavior on the environment.

As well as change the environment can have impact in the result. The common approach is run test with the differences.

General theories in Machine Learning arise from feature characterizations of programs, their environments, tasks and behaviors.

- Build a program that exhibits a behavior of interest while performing particular tasks in particular environments;
- Identify specific features of the program, its tasks and environments that influence the target behavior;
- Develop and test a casual model of how these features influence the target behavior;
- Once the model makes accurate predictions, generalize the features so that other programs, tasks, and environments are encompassed by the causal model;
- Test whether the general model predicts accurately the behavior of this larger set of programs, tasks, and environments.

Empirical research methods can be divided into two categories:

Quantitative research methods: such methods collect numerical data (data in the form of numbers) and analyze it using statistical methods.

Qualitative research methods: such methods collect qualitative data (data in the form of text, images, sounds) drawn from observations, interviews and documentary evidence, and analyze it using qualitative data analysis methods.

Qualitative methods tend to be more appropriate in the early stages of research (exploratory research) and for theory building. Quantitative methods tend to be more appropriate when theory is well developed, and for purposes of theory testing and refinement. In practice, no research method is entirely qualitative or quantitative (Yin, 1994). Mixing qualitative and quantitative research methods is called triangulation of method.

Empirical methods for applying machine learning programs involve running theses programs and recording their behaviors, computer programs that perform tasks in environments. It shouldn't be difficult, compare with other systems, like biological. Machine learning systems are simple, compared with human cognition, their tasks are rudimentary; and compared with everyday physical environments, those in which our programs operate are extremely reduced. Haven so we can compare, we don't know how they work, we cannot say how long they will take to run, when they will fail, how much knowledge is required to attain a particular error rate, how many nodes of a search tree must be examined, Like other systems.

Cohen said that study a machine learning system is no different that a lab experiment. One obliges the agent (monkey, rat or program) to perform a task according to a experimental protocol, observing and analyzing the behavior. Six components are common to these scenarios: agent, task, environment, protocol data collection and analysis. The first three are the domain theories of behavior and the last three are empirical methods. [3]

4 Conclusion

In this paper we overview the main factors of empirical research. The difference about empirical, experimental and exploratory research, and how empirical research can be applied to machine learning and data mining project tasks to achieve the best benefits of use this approach in a research project.

References

1. *Oxford English Dictionary A*.
2. *The Future of Empirical Methods in Software Engineering Research*. 2007.
3. Paul R. Cohen. Empirical methods for artificial intelligence. 1995.
4. Usama Fayyad, Georges G. Grinstein, and Andreas Wierse. Information visualization in data mining and knowledge discovery. 2001.
5. Douglas B. Lenat and Edward A. Feigenbaum. On the thresholds of knowledge. *Artif. Intell.*, 47(1-3):185–250, 1991.
6. Tom M. Mitchell. Machine learning and data mining. *Communications of the ACM*, 42:30–36, 1999.