Predicate Project Outcomes Using Machine Learning

Salwa E. Ibrahim

Prince Sattam Bin Abdulaziz University, College of Science and Humanities Studies, Saudi Arabia

Abstract: ***Due to the project complexity and differences the need of effective planning and management in project is increase. The main objective of each project is to be successful, At least by determined three criteria time, cost and quality. The success criteria vary from project to project since we have different types and different people. Many researchers interested in artificial intelligence to assessment and predicate project outcome they want to see if computers could learn from data. This paper reviews appropriate machine learning models for predicate project outcomes and measure project success.***

Keywords: project management, Machine Learning, Artificial Intelligent

1. Introduction

One of the latest trends in the world of technology and engineering is “machine learning” — in fact, all of the big technology companies today have invested in artificial intelligence and machine learning projects.

The key question of this paper is to use machine learning model to help in predicate projects outcomes and help in making decision using historical data. The term “machine learning” was first defined by Arthur Samuel, way back in 1959. He defined it as “the ability to learn without being explicitly programmed,” which basically means that a machine could learn from its own mistakes and reprogram itself to improve its performance over time.

The idea gained popularity in the 90s when the concept of data mining came into existence. Data mining uses algorithms to look for patterns in a given set of information, which led to data-driven predictions and decision making. This encouraged engineers to develop complex machine learning algorithms by making use of data mining and predictive analytics.[7]

Machine learning is a set of algorithms that train on a data set to make predictions or take actions in order to optimize some systems. This paper outlines an introduction of introduction of machine learning, machine learning algorithms, ML Predictive Modeling, some of the related work and important ML model that can used to predicate project outcomes.

2. Machine Learning

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can change when exposed to new data. The process of machine learning is to search through data to look for patterns. However, instead of extracting data for human comprehension. Machine learning uses that data to detect patterns in data and adjust program actions accordingly. [1]

2.1 Machine learning algorithms

Machine learning algorithms are often categorized to:

- **Supervised machine learning algorithms** can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.
- **Unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn’t figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.
- **Semi-supervised machine learning algorithms** fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabeled data generally doesn’t require additional resources.
- **Reinforcement machine learning algorithms** is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

2.2 ML Predictive Modeling

Predictive modeling leverages statistics to predict outcomes. Most often the event one wants to predict is in the future, but predictive modeling can be applied to any type of unknown event, regardless of when it occurred. There are many different models the most popular models summaries as following: [4]

1) Regression model

Regression is concerned with modeling the relationship between variables that is iteratively refined using a measure of error in the predictions made by the model. Regression methods are a workhorse of statistics and have been co-opted into statistical machine learning. This may be confusing because we can use regression to refer to the class of problem and the class of algorithm. Really, regression is a process.
The most popular regression algorithms are:

- Ordinary Least Squares Regression (OLSR)
- Linear Regression
- Logistic Regression
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)

2) Decision Tree Algorithms

Decision tree methods construct a model of decisions made based on actual values of attributes in the data. Decisions fork in tree structures until a prediction decision is made for a given record. Decision trees are trained on data for classification and regression problems. Decision trees are often fast and accurate and a big favorite in machine learning.
The most popular decision tree algorithms are:

- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0 (different versions of a powerful approach)
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- M5
- Conditional Decision Trees

3) Clustering Algorithms

Clustering, like regression, describes the class of problem and the class of methods. Clustering methods are typically organized by the modeling approaches such as centroid-based and hierarchal. All methods are concerned with using the inherent structures in the data to best organize the data into groups of maximum commonality.
The most popular clustering algorithms are:

- k-Means
- k-Medians
- Expectation Maximisation (EM)
- Hierarchical Clustering

3. Related work

Using Machine Learning To Predict Project Effort

Author in 2 adopted a bottom-up approach for estimating project effort is quite feasible. It is evident that poor product results do not necessarily imply poor project results. Using empirical data gathered from two separate corporations and applying a neural network approach produced average project effort estimates of 13 and 10 percent for each set of experiments. The experiments also produced 90% and 100% accuracy for pred(25). Scaling the data, based on maximum SLOC values, helps to compensate for extrapolation issues. Results improved from an average accuracy of 42 to within 10 percent of actual values. This approach offers very good potential for those life cycle methodologies which incorporate prototyping early in the life cycle.

Machine Learning Approaches To Estimating Software Development Effort

Author in 3 describes two methods of machine learning, which we use to build estimators of software development effort from historical data. Our experiments indicate that these techniques are competitive with traditional estimators on one dataset, but also illustrate that these methods are sensitive to the data on which they are trained. This cautionary note applies to any model-construction strategy that relies on historical data. All such models for software effort estimation should be evaluated by exploring model sensitivity on a variety of historical data.

Estimating Software Project Effort Using Analogies

Author in 6 describe an approach to estimation based upon the use of analogies. The underlying principle is to characterize projects in terms of features (for example, the number of interfaces, the development method or the size of the functional requirements document). Completed projects are stored and then the problem becomes one of finding the most similar projects to the one for which a prediction is required. Similarity is defined as Euclidean distance in n-dimensional space where n is the number of project features. Each dimension is standardized so all dimensions have equal weight. The known effort values of the nearest neighbors to the new project are then used as the basis for the prediction. The process is automated using a PC-based tool known as ANGEL. The method is validated on nine different industrial datasets (a total of 275 projects) and in all cases analogy outperforms algorithmic models based upon stepwise regression. From this work we argue that estimation by analogy is a viable technique that, at the very least, can be used by project managers to complement current estimation techniques.

Software Cost Estimation

Author in 8 describes 29 software cost models that have been created since 1966. Due to the lack of data early in the software life cycle, most of these models apply to the latter stages of the software life cycle. However, there is one approach that identifies requirements measures and uses those to predict development effort

Success Evaluation Model For Project Management

Author in 9 presents an expert decision-making fuzzy model for evaluating project success. The proposed model (including sub-models) is implemented in the MATLAB software environment with the use of the Fuzzy Logic Toolbox application where it is also verified and further specified. MATLAB software was chosen for the construction of models in view of the fact that it is not necessary to perform a detailed examination of the essence of the principle of fuzzy sets (with which fuzzy logic works) which is an indisputable advantage in view of the varying standard of mathematics in the intercultural. The proposed model provides project managers and others with a tool for the "measurement" of selected project processes (Assessment of the state of the project, assessment of project risks, assessment of project quality, and assessment of project success). The fuzzy approach including knowledge base of expert rules and its ability to systematically, hierarchical and comprehensively evaluate three key criteria of project success is the main asset and simultaneously differ from current Models. A significant general advantage of the application of the technique of modeling in project management is the possibility of subsequent experimentation with the model, in the form of simulation for example. This makes further information about the possible variant development of projects available and can provide warning signals to support future Decision-making.

An Approach To Predict Software Project Success By Data Mining Clustering

Author in 10, developed an empirical analysis of several projects at various software industries brought out. In this analysis, defect count is considered to be one of the influencing parameters to predict the success of the projects. Various clustering algorithms are applied on the empirical projects to evaluate the above said objective. Observational inferences indicate that K-means is more efficient than other clustering techniques in terms of processing time, efficiency and reasonably scalable. The experimental results of K-means with rest of the clustering algorithms have ensured its continued relevance and progressively increased its effectiveness as well.

4. ML model to predicate project outcomes

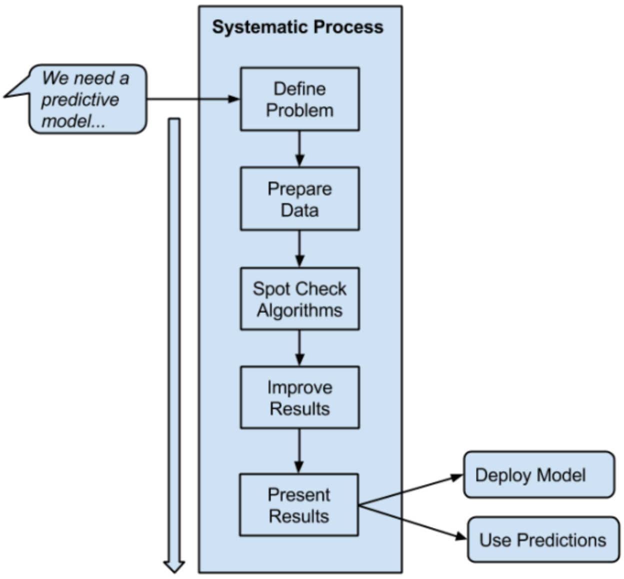
Project management concerned with cost, quality and on-time delivery. So ML application might be to learn from a training set of projects to learn what indicates the likelihood of delays. That could be good feedback for PMs.

Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look. Most companies working with large amounts of data have recognized the value of machine learning technology. By gleaning insights from this data – often in real time – organizations are able to work more efficiently or gain an advantage over competitors. [1]

The knowledge contained in historical data can be used to develop a predictive model with Machine Learning (ML) methods, such as neural networks, decision trees, or SVM (Support Vector Machine). Some Project Management domains with greatest ML application are Quality Management, Cost Management and Time Management. For example, a predictive approach to project times and costs requires the development of a "predictive function or model" through the choice of a specific ML method which, according to historical data, can provide the prediction of times and costs as output, for both the entire project and for specific activities. [5]

The following steps are used to predicate project outcomes from historical data:

- Importing data
- Cleaning data
- Splitting it into train/test or cross-validation sets
- Pre-processing
- Transformations
- Feature-engineering



this paper focus in Supervised machine learning algorithms to predicate project outcomes from historical data to know how its work, supervised learning Consist of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data.

The most important and helpful model is Logistic Regression model and Decision Tree model [4]

Logistic Regression model is a classification algorithm from supervised machine learning algorithms. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variable(s). In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as logit regression. Since, it predicts the probability, its output values lies between 0 and 1 (as expected).

Decision Tree model it is a type of supervised learning algorithm that is mostly used for classification problems.it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

Using these models we can build an artificial intelligence application which, simplifying, collects and analyzes data, highlighting patterns and imitating – with the aim of improving them and helping to make better decisions based on an accurate as possible analysis of data.

Machine learning algorithms can use to understand how you work, how you analyze problems and how you take decisions.

5. Conclusion

This paper explore how ML models can help in improve project outcomes or project performance by using learned data to make better decision.

The process of learning begins with observations or data in order to look for patterns in and make better decisions in the future. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results.

References

[1] Can Machines Deep Learn Project Management? , February 17, 2016, Quick Base Blog
[2] Using Machine Learning to Predict Project Effort: Empirical Case Studies in Data-Starved Domains (2001), by Gary D. Boetticher
[3] Machine learning approaches to estimating software development effort, IEEE Transactions on Software Engineering (Volume: 21, Issue: 2, Feb 1995) K. Srinivasan , D. Fisher
[4] A Tour of Machine Learning Algorithms, by Jason Brownlee on November 25, 2013 in Machine Learning Algorithms
[5] Project Management and Artificial Intelligence by MARCO CARESSA 17 JAN Ingenium
[6] Estimating software project effort using analogies, IEEE Transactions on Software Engineering (Volume: 23, Issue: 11, Nov 1997), M. Shepperd, C. Schofield
[7] How to use machine learning in today's enterprise environment, November 9,2016 , CONNECTED DEVICES, FINTECH, HEALTH, INDUSTRIAL, SMA RT CITIES,TRANSPORT , SHUVRO SARKAR
[8] Heemstra, F. "Software Cost Estimation," Information and Software Technology, October 1992, Pp. 627-639.
[9] Radek Doskočil, Stanislav Škapa, Petra Olšová, SUCCESS EVALUATION MODEL FOR PROJECT MANAGEMENT, Information management(Volume: 19, Issue: 4, 2016)
[10] International Conference on Data Mining and Computer Engineering (ICDMCE'2012) December 21-22, 2012 Bangkok (Thailand), An Approach to Predict Software Project Success by Data Mining Clustering, Suma.V, Pushpavathi T.P, and Ramaswamy.V

Author Profile

Salwa E. Ibrahim received the B.S. in computer science and statistic from University of Khartoum- Sudan in 2001-2006 and M.S. degrees in computer science and information from Gezira University in 2009-2011.