# A Neural Network Based algorithms for Project Duration Prediction

WanJiang HAN[1, a], HeYang JIANG [2, b], Xiaoyan ZHANG[1, c],Weijian LI[2]
[1]*School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*

[2]*International School, Beijing University of Post and Telecommunication Beijing, 100876, China*

[a]*hanwanjiang@bupt.edu.cn,* [b]*2014213480 @bupt.edu.cn*, [c]*xiaoyan@bupt.edu.cn*

***Abstract***

*Prediction of software development effort and duration is the key task for the effective Software Project Management (SPM). The accuracy and reliability of prediction mechanisms is also important. In this paper, we used several machine learning algorithms to predict the software duration. Experiment results show this kinds of algorithms is effective.*

## 1. Introduction

Before starting a project, software project planning is one of the most critical activities in modern software development. Without a realistic and objective plan, a software development project cannot be managed in an effective way. The knowledge stored in the historical datasets can be used to develop predictive models, by either statistical methods such as linear regression and correlation analysis or machine learning methods such as ANN (Artificial Neural Network) and SVM (Support Vector Machine), to predict the duration of projects.

Prediction-based approaches require a prediction function that according to the current/past data of the project, will predict the future project effort and duration. However, due to the vast number of machine learning algorithms, there are still different machine learning algorithms not analyzed.

In this paper, we present an experimental comparison of several neural network algorithms. We evaluate the algorithms according to their correlation coefficient.

## 2. Related work

Artificial neural network is composed by neurons. These neurons in the neural network are dealt with the message, and connected through weights. In this section, we will briefly describe the main characteristics of the machine learning algorithms compared.

### 2.1 Linear regression

Linear regression (LR) is a method to model the linear relationship between a scalar dependent variable y and one or more independent variables x.

### 2.2 Least Median Square

978-1-4799-8206-6/14 $31.00 © 2014 IEEE
DOI 10.1109/CA.2014.21

Least Median Square (LMS) method is one of the statistical methods for solving the equations which are more than unknown. This method almost used in analytical regression. In fact, Least Median Square is a method for fitting the dataset. The least Median Square must yield the smallest value for the median of squared residuals computed for the entire data set. It means the residuals, the difference between real data and predicted data.

### 2.3 Gaussian process

In probability theory, the Gaussian process (GP) is realization consist of random values associated with every point in a range of times. The random variable has a normal distribution. In it has been mentioned that "a Gaussian process is a stochastic process whose generalization of a Gaussian distribution over a finite vector space to a function space of infinite dimension".

### 2.4 Sequential minimal optimization

Sequential minimal optimization (SMO) was proposed as an iterative algorithm. This algorithm uses SVM for solving regression problem and SVM classifier design. The SMO algorithm has two worthy aspects: implementation is easy and computational speed is fast.

### 2.5 Multilayer Perceptron

The multilayer perceptron (MLP) is a very simple model of biological neural networks. The network is structured in a hierarchical way. Multilayer Perceptron includes some nodes located in different layers where the information flows only from one layer to the next layer. The first and the last layers are the input and output. Layers between the input and output layer are called hidden layers. This network is trained based on back propagation error in that the real outputs are compared with network outputs, and the weights are set using supervised back propagation to achieve the suitable model.

## 3. Experimental Results

### 3.1 Project information

We have collected lots data of software projects. Table 1 describes the attributes of projects, which are used to build the prediction model by a machine learning technique.

Table 1 Information of the projects collected.

| Number of projects | Number of Workers | Time of projects(Month) | Effort of projects(PM) | KLOC |
|---|---|---|---|---|
| 38 | 5-26 | 2-25 | 2-200 | 1-200 |

We use 70 percent of collected data to build the model and 30 percent of them to predict and evaluate. The data instances used to train and test the models was chosen randomly, once. Then the same training data and testing data was used for all the models.

### 3.2 Experiment Results

In this Section, we compared the machine learning algorithms. Each project dataset is used to compare the algorithms, including linear regression, SMO, multilayer Perceptron, M5P, Leastmedsqr, REPtree and Gaussian process. For comparing machine learning algorithms, we use the correlation coefficient to evaluate the results. Based on this value, we rank the algorithms according to each metric of interest. 错误!未找到引用源。 presents the correlation coefficient of each algorithm. We can state that predicting values by Gaussian Process are directly and closely related to training dataset values, comparing other algorithms.

Table 2 The evaluation for per algorithm

| | LR | MLP | M5P | SMO | GP | LMS | RT |
|---|---|---|---|---|---|---|---|
| **Correlation Coefficient** | 0.910 | 0.921 | 0.897 | 0.956 | 0.976 | 0.898 | 0.921 |

## 4. Conclusions

In this paper, we use different machine learning algorithms to build software predict model, based on lots of project data. We comprehensively compared the results. Our experimental results show that Gaussian Process obtains the better overall performance among the studied algorithms.

## 5. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61170273).

## 10. References

[1] Ching-seh Wu , Wei-chun Chang, Ishwar K. Sethi,A Metric-Based Multi-Agent System for Software Project Management,2009 Eigth IEEE/ACIS International Conference on Computer and Information Science

[2] Ian.H. Witten, and Eibe.Frank, “Data Mining Practical Machine Learning Tools and Techniques” , Morgan Kaufmann Publishers Is an Imprint of Elsevier, pp. 187-427, 2005.

[3] K. Nigam, A. Mccallum, and T. Mitchell. Text classi_cation from labeled and unlabeled documents using em. Machine Learning, 39(2):103{134, 2000.

[4] Zhang, Ye Yang, Qing Wang, Handling missing data in software effort prediction with naive Bayes and EM algorithm, PROMISE ’11, September 20-21, 2011, Banff, Canada.

[5] A.Andrzejak, L.Silva, “Using Machine Learning for Non-Intrusive Modeling and Prediction of Software Aging” , Network Operations and Management Symposium, IEEE, pp.25-32 , April 7-11 ,2008.

[6] J.Alonso Lopez, J.L.Berral, R.Gavalda and J.Torres, “Adaptive on-line software aging prediction based on machine learning” , in Procs. 40th IEEE/IFIP Intl. Conf. on Dependable Systems and Networks, pp 497-506, June 28 ,July 1, 2010.

[7] Y.Wang, and I.H.Witten, “Inducing Model Trees for Continuous Classes” , thE European Conf. on Machine Learning Poster Papers, 1997.

[8] Weka 3.6.8 http://www.cs.waikato.ac.nzlmllwekal.

[9] A. J.Smola, and B.Schölkopf, “A Tutorial on Support Vector Regression” , Royal Holloway College, London, U.K., NeuroCOLT Tech. Rep.TR 1998-030, 1998.

[10] S.K.Shevade, S.S.Keerthi, C.Bhattacharyya and K.R.K.Murthy, “Improvements to the SMO Algorithm for SVM Regression” , IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL.11, NO.5, Sep, 2000.

[11] David J.C. Mackay, “Introduction to Gaussian Processes” , Dept. of Physics, Cambridge University, UK, 1998.

[12] A.Macedo, T.B.Ferreira, R.Matias, “The Mechanics of Memory-Related Software Aging” , Second IEEE International Workshop on Software Aging and Rejuvenation , Nov2-2. 2010.