Journal of Knowledge Information Technology and Systems

ISSN 1975-7700

http://www.kkits.or.kr

A Study of Model Selection for Electric Data using Cross Validation Approach

Saraswathi Sivamani, Saravana Kumar Venkatesan, Changsun Shin, Jangwoo Park, Yongyun Cho*

Department of Information and Communication Engineering, Sunchon National Univeristy, Suncheon-si, Jeollanam-do, Republic of, Korea.

ABSTRACT

In this paper, the appropriate model is selected for the risk assessment of the electric utility pole data with the help of cheat sheets and k-fold cross validation. In order to analyze, predict and forecast the data, the appropriate model has to be selected. The major issue is the declination of the accuracy in the model fitting, which may result in poor model selection. There are different type of machine learning algorithm, which makes it difficult to conclude the model selection. To ensure the proper selection of the model, we undergo a two-step process. Firstly, the basic model is selected with the existing model selection cheat sheets named as Scikit learn and Microsoft azure, by understanding the available input and required output of the data. After getting through the multiple question, the respective models such as Generalized Additive Model, Generalized Linear Model, Linear Regression and Support Vector Machine are obtained. In order to attain the appropriate model, we perform k-fold cross validation to estimate the risk of the algorithms, by comparing 2-fold, 8-fold and 10-fold cross validation. Between the three set, the 10-cross fold validation of generalized additive model is selected with the least risk error. Using k-fold cross validation, we estimate the accuracy of the model that is suitable for the data, by using the electric power data set.

KEYWORDS: Model selection, K-fold cross validation, Machine learning, Model fit, Electric power

ARTICLE INFO: Received 18 October 2017, Revised 24 November 2017, Accepted 8 December 2017.

*Corresponding author is with the Department of Information & Communication Engineering, Sunchon National Univeristy, Suncheon-si, Jeollanam-do, Republic of, Korea.

E-mail address: yycho@sunchon.ac.kr

1. Introduction

With the tremendous amount of data obtained from different areas, deriving the useful information from the set of available data, predicting and forecasting the data is difficult process. Machine learning was developed to construct computer algorithms that automatically improves with experience [1].

Bishop et al [2] defines that the field of pattern recognition is concerned with the finding of regularity through automation by using computer algorithms and with the use of these regularities, the data are classified into different categories. The pattern recognition and machine learning are closely related with minor differences. Similarly, numerous algorithms are related and use similar methods. This explains the fact that these activities has product of two different disciplines. Unlike machine learning, Pattern recognition has a primitive origins in engineering. The other method that is closely related is data mining. Data mining finds the unusual relationship form the observed data to organize the data and find useful information for the user [3]. Jain et al. [4] addresses the situation as difficult process to teach a machine to solve a pattern recognition problem, but the solution may be trivial to the human eye. Still, it is impossible to find the pattern through human eyes, as there are million of pattern. With this, the difficulty of appropriate algorithm selection is also increased. In order to find a solution, many researches focus on the classification of the algorithm [5] or validate the model selection [6]. Some selection process uses the error functions to validate and find the required model [7], while others uses residuals [8]. To solve this purpose, many model selection cheat sheets were designed, where some the most popular ones are Scikit-learn [9] and Microsoft azure [10]. The existing sheets can only identify a generalized or basic algorithm, which cannot help us to obtain the best fit model. Therefore, in this paper, we have combined the cheat sheet along with the k-fold cross validation to validate the chosen algorithm that helps to find the best fit model for the electric pole risk assessment process. In other words, the cheat sheet helps to extract the basic model. Then, the cross validation is used to find the exact model through the error factors such as Root Mean Square Error and R-Squared.

2. Background Study

2.1 Regression

Regression is used to find the relationship between the independent variables on a single dependent variable [11]. Commonly, the test includes the deviation of the mean and its scales the interval. The analysis is conducted by using the raw data matrix and sometimes correlation matrix are used.

In other words, regression analysis measures the degree of influence of the independent variables on a dependent variable. To find the dependency of the single independent variable, the simple equation (1) is used.

$$y = a + bx \quad \text{-> (1)}$$

where, a is the intercept and b is the slope, which can be explained as shown in equation (2) and (3)

$$a = \frac{\sum Y - b\sum X}{N} \quad \text{-> (2)}$$

$$b = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} \quad \text{-> (3)}$$

Where, N is the number of observation, X is the year index and Y is the annual size of the annual years. Although, the equation points the single variable, it can also be extended to find the relationship between multiple variable. The relationship predicted is always linear, whether it is for a single variable or for multiple variables.

2.2 Classification

In case of the classification methods, the problems admits only the unordered, unlisted and discrete values. Basically, it predicts the categorical class label, by classifying the data through the training set [12]. Classification is closely related to the Regression, where the Logistic regression is considered to be part of the classification. At first, the classifier is fed with the training data that has been labeled properly with the respective class. Later, this data is trained to the learn the algorithm, which helps to create a classifier with similar traits.

2.3 Clustering

Clustering is one of the unsupervised learning methods which groups or categorize the data set through constant observation [13]. Clustering is widely used algorithm in many areas such as customer segmentation, commercial grouping, and social network analysis. As the clustering is unsupervised, the result can only be evaluated through visualization. But, it is possible to evaluate the grouping by training the data with the available set.

The most commonly used algorithm in the clustering is k-means clustering [14] that evaluates the distance between the two coordinate points. The method is most popular due to its simplicity and the flexibility for the data processing. The only difficulty lies in the selection of cluster numbers which can determine the quality of the model.

3. Model Selection Methods

Among various cheat sheets and solutions, the most popular ones are Scikit learn sheet and Microsoft Azure Sheet that helps to narrow the selection process.

3.1 Scikit Learn

Scikit learn [15] is widely used in the many machine learning algorithm problems. It is an integrated python module that includes most recent algorithm for solving problems from various applications. With the expert knowledge, the cheat sheet was designed, which helps the

narrow down the search of the machine learning, by knowing the simple characteristics of the data. The algorithm is divided into four categories such as Regression, Classification, Clustering and Dimensionality Reduction, which subdivides further with the more specifications.

3.2 Microsoft Azure

Microsoft Azure [16] emphasis that each algorithm has its unique traits and cannot be categorized as decision on the data input and expected output. Although, the algorithms are categorized into four types that is similar to the Scikit learn, instead of the decision approach, the characteristics are defined for each algorithm. It helps to identify and compare the algorithm suitable for the data modelling and analysis.

4. Model Selection Process

Based on the Scikit learning sheet and Microsoft azure sheet, the machine learning methods are categorically separated based on the nature of the algorithm. Although the selection process is complicated, both the cheat sheet gives a rough outlier of the machine learning categorical values that represent the most capable model selection for the respective data. An electric utility pole data of the southern district has been used for the analysis process. The Pitch and Roll data are the observed data from the electric utility pole. The five devices are transformer, load switch, load balance, utility pole, and communication enclosure.

The electric power utility pole data includes the pole information that contains acceleration and pitch for these devices. According to the cheat sheets such as scikit learn and Microsoft azure, the selection of the model selection is made easier as shown in the figure 1. By following the cheat sheets that are currently available, it was easy to scrutinize the model selection to a certain extend. But the real problem starts with the numerous number regression models.

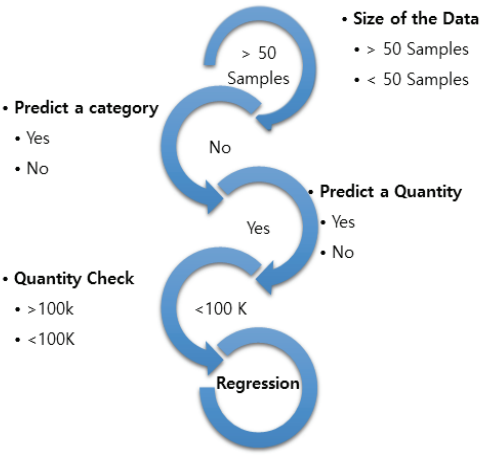

그림 1. 치트 시트를 기반으로 한 기본 모델 선택

Figure 1. Basic Model Selection based on Cheat Sheet

According to the available data, we selected three models that are closely related to the data, namely Linear Regression, Generalized Additive Model, Support Vector Machine and Generalized Linear Model, that can be used for model fitting. The common problem that resides in the data modelling is over-fitting which leads to the improper model fit.

표 1. 2 배 교차 검증 리샘플링 결과
Table 1. 2-fold Cross Validation Resampling result

	RMSE	R Squared
Linear Regression	0.3316	0.6363
Generalized Linear Model	0.2962	0.6725
Generalized Additive Model	0.1958	0.5736
Support Vector Machine	0.2914	0.6395

To solve the over-fitting issue, the data is cross validated with the k-fold technique. In the range of 1 to 10, the cross validation is performed, from which the best three folds are compared with the four models to find the best fit result. For each model, the number of folds such as 2, 8 and 10 are tested, to find the Root Mean Squared Error (RMSE), and RSquared.

표 2. 8 배 교차 유효성 검사 리샘플링 결과
Table 2. 8-fold Cross Validation Resampling result

	RMSE	R Squared
Linear Regression	0.4819	0.6511
Generalized Linear Model	0.5828	0.6013
Generalized Additive Model	0.3876	0.5685
Support Vector Machine	0.5996	0.8713

<Table 1> shows the two fold cross validation of the electric utility pole data. Similarly, <Table 2> and <Table 3> shows the 8- fold and 10-fold cross validation for all the four models.

The model is good, if the range is < 0.5. Similarly, the R-Squared value should range between 0 to 1, which is the statistical measure of how close the data are fit to the regression line. By using the RMSE and R-Squared from the three tables, the best number of folds is identified as the ten fold <Table 3> which has low RMSE values less than 0.3 and high R-Squared values that ranges between the 0.7 and 0.9. Generalized additive model from the 10 fold cross validation gives low RMSE value with and the precise R-Squared values that justify the best fit model. With this, the model selection is cut down to Generalized addictive model using the simple cheat sheet and k-fold cross validation.

표 3. 10 배 교차 검증 리샘플링 결과
Table 3. 10-fold Cross Validation Resampling result

	RMSE	R Squared
Linear Regression	0.1938	0.7006
Generalized Linear Model	0.2272	0.8497
Generalized Additive Model	0.1668	0.8991
Support Vector Machine	0.2771	0.7666

5. Conclusion

In this paper, we have used the basic model selection cheat sheet and cross validation approach to find the best fit model for the electric power utility pole data set. First, the initial model selection is decided through the

available cheat sheet and then the cross validation is used to finalize the suitable model for the data set through the Root Mean Square Error and R-Squared Error, which had a less risk error factor. For the future work, we plan to improvise the model selection process through machine learning methods.

References

[1] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, (Eds.). *Machine learning: An artificial intelligence approach,* Springer Science & Business Media, 2013.

[2] C. M. Bishop, *Pattern recognition and machine learning (information science and statistics)* springer-verlag new york. Inc. Secaucus, NJ, USA, 2006.

[3] D. J. Hand, H. Mannila, and P. Smyth, *Principles of data mining*, MIT press, 2001.

[4] A. K. Jain, R. P. W. Duin, R. P. W. and J. Mao, *Statistical pattern recognition: A review*, IEEE Transactions on pattern analysis and machine intelligence, Vol. 22, No. 1, pp. 4-37, 2000.

[5] T. O. Ayodele, *Types of machine learning algorithms*, In New advances in machine learning. InTech, 2010.

[6] S. Arlot, and A. Celisse, *A survey of cross-validation procedures for model selection*, Statistics surveys, 4, pp. 40-79, 2010.

[7] B. Gu, and C. X. Ling, *A new generalized error path algorithm for model selection*, In International Conference on Machine Learning, pp. 2549-2558, 2015.

[8] R. R. Bies, M. F. Muldoon, B. G. Pollock, S. Manuck, G. Smith, and M. Sale, *A genetic algorithm-based, hybrid machine learning approach to model selection,* Journal of Pharmacokinetics and Pharmacodynamics, Vol. 33, Issue 2, pp. 195-221, 2006.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and J. Vanderplas, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12(Oct), pp. 2825-2830, 2011.

[10] S. Mund, *Microsoft azure machine learning*, Packt Publishing Ltd., 2015.

[11] S. Chatterjee, and A. S. Hadi, *Regression analysis by example*, John Wiley & Sons., 2015.

[12] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, *Machine learning, neural and statistical classification*, 1994.

[13] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, *Supervised machine learning: A review of classification techniques*, 2007.

[14] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, *An efficient k-means clustering algorithm: Analysis and implementation*, IEEE transactions on pattern analysis and machine intelligence, Vol. 24, No. 7, pp. 881-892, 2002.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and J. Vanderplas, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12(Oct), pp. 2825-2830, 2011.

[16] Microsoft Azure, https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice, Oct 2017.

교차 검증 접근법을 이용한 전력 데이터 모델 선택에 관한 연구

Saraswathi Sivamani, Saravana Kumar,
신창선, 박장우, 조용윤
순천대학교 정보통신공학과

요 약

이 논문에서는 치트 시트 및 k-교차 검증 기법을 이용하여 전주 상에서 검출되는 전력 데이터에 대한 최적 위험 평가 모델 기술 연구를 제안한다. 일반적으로 효율적인 빅데이터 분석 및 예측을 위해서는 데이터의 특성과 상황이 고려된 적절한 분석 모델의 선택이 중요하다. 빅데이터 분석을 위한 기계 학습 알고리즘은 모델유형이 다양하여 모델 선택이 어려울 수 있으며, 이러한 과정에서 모델 정합의 정밀도가 낮아지면 모델 선택의 오류가 발생될 수 있다. 제안하는 방법은 최적의 데이터 분석 모델의 선택을 보장하기 위해 2단계 모델 선택과정을 포함한다. 첫째, 기본 모델 과정은 데이터의 사용 가능한 입력 및 필요한 출력을 이해함으로써, Scikit learn 및 Microsoft azure와 같은 선택 치트 시트를 통한 데이터 모델 선택과정이다. 이때, 다중 질문을 거친 후 일반화 된 가산 모델, 일반화 된 선형 모델, 선형 회귀 및 지원 벡터 머신과 같은 각각의 모델이 얻어진다. 두 번째 단계에서, 2배, 8배 및 10배의 교차 검증 비교를 통해 선택된 모델의 오류위험을 평가하는 k 배 교차 검증을 수행한다. 제안하는 모델의 시뮬레이션 실험을 통해 전력데이터에 대한 3가지 세트 사이에서, 일반화 된 모델의 10-교차 검증이 가장 적은 위험 오류로 선택되었음을 보였다. 따라서, 제안하는 k-배 교차 검증 방법을 사용하는 경우, 전력 데이터 세트뿐만 아니라 다양한 데이터 셋에 대해 해당 데이터에 적합하고 오류가 최소화된 데이터 모델을 선택할 수 있다.

Acknowledgments

This paper was supported by(in part) Sunchon National University Research Fund in 2017.

Saraswathi Sivamani completed bachelor degree in Information Technology from India. She received Master degree on Information and Communication Engineering in South Korea and currently pursuing Doctorate degree in the same department. Her area of interest includes Ontology model, Ubiquitous Computing and Big data analysis and modelling.

E-mail address: saraswathi86@gmail.com

Saravana Kumar Venkatesan completed bachelor degree in Mathematics from India. He is currently studying Master degree in Information and Communication Engineering in South Korea His area of interest includes Data analysis and Modelling

E-mail address: skkumarvsk061@gmail.com

Changsun Shin received the Ph.D. degree in computer engineering at Wonkwang University. Currently, he is an associate professor of the Department of information & communication engineering in Sunchon National University. His main research interests include Distributed Real-Time Computing, Distributed Object Modeling, Ubiquitous Agriculture and Ubiquitous Sensor Network (USN).

E-mail address: csshin@sunchon.ac.kr

Jangwoo Park received the B.S., M.S. and Ph.D. degrees in Electronic engineering from Hanyang University, Seoul, Korea in 1987, 1989 and 1993, respectively. In 1995, he joined the faculty member of the Sunchon National University, where he is currently a professor in the Department of Information & Communication engineering. His research focuses on Localization and SoC and system designs and RFID/USN technologies.

E-mail address: jwpark@sunchon.ac.kr

Yongyun Cho received the Ph.D. degree in computer engineering at Soongsil University. Currently, he is an assistant professor of the Department of Information & communication engineering in Sunchon National University. His main research interests include System Software, Embedded Software and Ubiquitous Computing.

E-mail address: yycho@sunchon.ac.kr