International Journal of Distributed Artificial Intelligence
Volume 13 • Issue 2

A Comprehensive Feature Selection Approach for Machine Learning

Sumit Das, JIS College of Engineering, India
https://orcid.org/0000-0002-8018-9151
Manas Kumar Sanyal, University of Kalyani, India
Debamoy Datta, JIS College of Engineering, India

ABSTRACT

In machine learning, it is required that the underlying important input variables are known or else the value of the predicted outcome variable would never match the value of the target outcome variable. Machine learning tools are used in many applications where the underlying scientific model is inadequate. Unfortunately, making any kind of mathematical relationship is difficult, and as a result, incorporation of variables during the training becomes a big issue as it affects the accuracy of results. Another important issue is to find the cause behind the phenomena and the major factor that affects the outcome variable. The aim of this article is to focus on developing an approach that is not particular-tool specific, but it gives accurate results under all circumstances. This paper proposes a model that filters out the irrelevant variables irrespective of the type of dataset that the researcher can use. This approach provides parameters for determining the quality of the data used for mining purposes.

KEYWORDS

Artificial Neural Network (ANN), Deep Learning (DL), Machine Learning (ML), Predictor Variable, Support Vector Machine (SVM), Variable Importance

1 INTRODUCTION

A data scientist usually starts with some hypothesis about what should go in a predictive model. Since a model is a simplified representation of the world, will never conclusively know that a variable causes some effect. The data scientist can only try to differentiate correlation and causation, by measuring a variable's predictive power across a long time-span. Perhaps even across the whole lifetime of the project. The most important characteristics of variable importance and feature selection are less data ensure simpler data layer; simpler models ensure faster machine learning; lower Variance ensures models generalize and comprehensive. Now there are some dedicated linear times algorithms for feature selection but none of them are generalized. This paper proposes a generalized algorithm that not only selects the most relevant features but also selects less number of variables whenever possible without sacrificing the accuracy. Firstly, the existing works are described in the literature survey section, and then the proposed algorithm is described in the methodology section along with

DOI: 10.4018/IJDAI.2021070102

the mathematical justification of the use of the algorithm as regarding its improved performance as compared to the present works. Then the results are compared for various test data and their interpretations are explained. This section also describes how the results support the data. Finally, conclude with the further scope of improvement in the conclusion section.

2 LITERATURE SURVEY

Any typical machine-learning algorithm involves extracting the output from a given set of input variables. These variables are also called as features in the input space, the input space can be represented as a vector in $\mathbb{R}^n$. The algorithms are developed assuming a functional relationship between the input vector and the output variable. $f : \mathbb{R}^n \rightarrow \mathbb{R}$. This assumption may not always be true; there may be some variables that may not have any influence on the output variables, to filter such variables numerous algorithms for ranking such variables are proposed. In a paper by Alain Rakotomamon new methods for variable selection are proposed for Support Vector Machines. Initial developments came from Guyon & Elisseeff (Guyon & Elisseeff, 2003). This paper contained an algorithm for selecting genes that are relevant for cancer classification problem. His goal was to find a subset of size r among d variables that maximizes the performance of predictor. The criterion $\left\|w\right\|^2$ is used where: $f\left(x\right) = w.\phi\left(x\right) + b$; ϕ is the kernel used in svm , w and b are the parameters for a particular model. The ranking criterion that has been used is $R_c\left(i\right) = \nabla\left\|w\right\|^2$ his algorithm runs in linear time.

Similarly other methods for variable selection has been proposed for neural networks also, the starting development in this area was done by Garson(Beck, 2018) which was later modified by Goh (Goh, 1995) to rank the variable importance, there a simple equation that is based on connection weights were proposed, Qik determined the relative importance of i-th input on k-th output. However, the main disadvantage of the Garson's algorithm was it used absolute values of weights that sometimes led to erroneous results. This disadvantage was removed by Olden(Olden et al., 2004). In a latest paper by Liu and Zaho (Liu & Zhao, 2017) a variable importance weighted random forest is used for classification and regression, the problem was the performance of random forest fall down when the number of features increased. In another paper by Kvalheim et al (Kvalheim et al., 2014) variable importance in latent variable regression model is proposed they presented some new graphical tools for improved interpretation of latent variable regression models that could assist in variable selection. Therefore, for different AI tools you have different algorithms. In this paper, present a novel algorithm that runs for any type of tool someone is using and provides the criterion to decide whether the results are meaningful.

Data Mining is a term coined to describe the process of sifting through large databases for interesting patterns and relationships. In this paper analyzes the cancer, Wisconsin performance of Decision Tree, Breast cancer classifier CART with and without feature selection in terms of accuracy, time to build a model and size of the tree on various. This paper is the Combination of attribute selection method with Logistic Regression; here the Quick Reduct Algorithm is used for attribute selection. The different attribute selection approaches defined for Logistic Regression, they are logistic regression with Backward Elimination, Logistic Regression with Forward Selection and Logistic Regression with Quick reduct Algorithm, and the results elaborates the Intercept, Coefficient and AIC measures for the Diabetes dataset(Samya, 2017). By distinctive the foremost salient options for learning, focuses a learning algorithmic rule on those aspects of the information most helpful for analysis and future prediction. A technique for correlation-based feature choice, supported concepts from take a look at theory, is developed. It evaluated using common machine learning algorithms on a variety of natural and artificial problems. It eliminates immaterial and redundant information and, in several cases,