Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/51/20241467

Comparison of different machine learning models: Linear model, forest and SVM

Xinran Liu

Beijing-Dublin International College, Beijing University of Technology, Beijing, 100083, China

xinran.liu@ucdconnect.ie

Abstract. Over the past century, machine learning has developed rapidly with the fast evolution of AI. As high-performance computing (HPC) has popularized over the past few decades, huge differences have been made in the field of machine learning, from expert systems to deep learning algorithms nowadays. This study summarizes some well-known models and compares three of the current popular machine learning models: the Linear Model, the Forest Model, and the Supporting Vector Machine (SVM), which aims to give and clearer view of those models in several aspects. Forest and SVM are more powerful in performance than the Linear Model due to their ability to deal with more types and more complex relationships and patterns. However, Forest and SVM have a higher complexity than the Linear Model and are less interpretable. Through the comparison of the three models, the basic model and algorithms will be better understood, the appropriate model will be selected faster when solving the problem, saving training time and space, and effectively saving resources and costs.

Keywords: Machine learning, Linear Model, Forest, Supporting Vector Machine (SVM).

1. Introduction

With the rapid growth of the economy and the fast development of technology, artificial intelligence (AI) has been widely applied in various domains. From the 1950s to the present, artificial intelligence has experienced three major periods: the reasoning period, the knowledge period, and the learning period [1]. Machine learning became an independent discipline, and various machine learning techniques blossomed since the learning period in the 1980s. Similar to the evolution of AI, the development process of machine learning can be divided into four major periods [2]. The hot period aims to develop various self-organizing and adaptive systems from the mid-1950s to the mid-1960s. The cooling-off period mainly symbolically simulated the conceptual learning process of humans using semantic networks and predicate logic from the mid-1960s to the mid-1970s. The Revival period completed the expansion from learning a single concept to learning multiple concepts from the mid-1970s to the mid-1980s. During the Booming period, various learning methods and learning systems that integrate multidisciplinary knowledge have been widely applied to various technical fields from the 1980s to the present [2].

In the past few decades, new machine learning algorithms have continued to emerge, but once limited by computing power have not been able to achieve rapid development, with the emergence of high-performance computers, machine learning has developed rapidly, from the former expert system

DOI: 10.54254/2755-2721/51/20241467

evolution to the current deep learning [3]. By the 2000s, deep learning, falling under the category of machine learning, was advancing rapidly [1]. The development of computer hardware processing technology and data storage technology has brought broader development and application to deep learning, and deep learning technology has rapidly swept the entire field of artificial intelligence. Recently, different models of machine learning have been widely used in many different fields. Chen et al. carry out a comparison of the mainstream machine learning model applied to the identification effect of auto insurance fraud [4]. Wang et al. use Random Forest in Self-Paced Bootstrap Learning in Lung [5]. Sun conducted a study on stock selection strategy based on Random Forest and Support Vector Machine (SVM) [6].

Since the rapid growth in artificial intelligence, various machine learning models have been used to train models to make accurate predictions. Machine learning has been extensively used in a large number of different fields. However, choosing the appropriate machine learning model for a specific task is not a trivial problem. Different models may have different advantages and disadvantages in terms of performance, complexity, and interpretability. Using an unsuitable model may lead to low accuracy, high computational cost, poor generalization, or other negative consequences. Therefore, it is important to understand the characteristics and trade-offs of different machine learning models and select the best one for the given problem. This paper summarizes the results of some previous studies and will give some author's views on several models of machine learning. This paper will introduce the current mainstream machine learning methods. Three of the popular machine learning models will be introduced in detail: Linear Model, Forest, and SVM. Description will be carried out within their basic principle, algorithm, performance, and application scenarios. A comparison will be made among them in the aspects of performance, complexity, and interpretability. Limitations of this field and Future outlooks of machine learning will also be discussed in this paper.

2. Basic descriptions of the models

Various machine learning methods are used nowadays, based on some traditional learning models, such as Linear Regression, K-Nearest Neighbour, Decision Tree, some optimization based on basic models, or some comprehensive models combining multiple models have gradually emerged, such as Random Forest, Extreme Gradient Lifting Decision Tree Model. This section will provide a brief introduction to some current mainstream machine learning models: Logistic Regression, and K-Nearest Neighbour (KNN). Logistic Regression is a probability-based classification algorithm that maps an input variable to a probability value between 0 and 1 by fitting a Sigmoid Function and then determines the class based on the probability value [3]. Logistic regression is suitable for binary classification problems and can also be extended to multi-classification problems. The advantages of logistic regression are that the model is simple, easy to understand and implement, the output values have probabilistic significance and can deal with linear and nonlinear relations [7]. The disadvantage of logistic regression is that it can easily be affected by noisy data and outliers, and is sensitive to Multicollinearity. If too many explanatory variables are selected, the logistic regression model can easily be overfitted, resulting in a model that fits well in the training set but is not suitable for predictive analysis of other data sets [4]. Thus, feature selection and Regularization are required to avoid Overfitting. KNN is one of the typical examples that represent lazy learning, which was proposed by Cover and Hart in 1976 [4]. When deciding the type of an unknown instance in each data set, suppose this instance is called x, the distance of x and its neighbour will be computed, and its k nearest neighbours will be picked out to determine the type of x observe the majority rule. The value of k is partly decided by the scalar of the data set, finding the most suitable k value is the key to this algorithm. Too small a value of k will make the model overfit; the model only works well on the training set and performs poorly on other data sets due to the large variance of the machine learning model [4].

DOI: 10.54254/2755-2721/51/20241467

3. Linear model

Linear Model is an early-used machine learning model that can be used to deal with regression or classification problems [8]. It aims to generate a formula after training by the given data set to fit a line that best predicts unknown values. It can be expressed by a function:

$$f(x) = w_1x_1 + w_2x_2 + \cdots w_nx_n + b \tag{1}$$

with vector term:

$$f(x) = w^Tx + b \tag{2}$$

This model is attended to solve the parameters w and bias b to minimize the variance, in other words, the difference between the predicted value and the true value. To achieve this goal, two typical linear models are used to deal with different kinds of problems, linear regression models are used to deal with regression problems while logistic regression models are used to deal with classification problems [8]. Linear regression mainly covers two different linear regression models, the unary linear regression model and the multiple linear regression model [9]. The function expression of the unitary linear regression model is shown as follows:

$$\hat{y} = \beta_0 + \beta_1x \tag{3}$$

where $\hat{y}$ is the predicted estimate of the model output, β_0 and β_1 are the prediction coefficients. Here least square method is always used to calculate the output. Multiple linear regression uses multiple associated independent variables to solve the regression problem, and its model function expression is as follows:

$$\hat{y} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_mx_m \tag{4}$$

Here, m represents the total number of variables (features) and β_1 ... β_m is the prediction coefficient corresponding to each independent variable.

The generalized Linear Model (GML) is a generalization of linear regression, with greater flexibility to model by a wide range of different underlying statistical distributions and allow a wider range of outcomes [10]. The coefficients in GML are computed by maximum likelihood estimation, to determine the most likely outcome of the data, its expression is as follows:

$$f\big(E(Y)\big) = \hat{\beta}_0 + \hat{\beta}_1X_1 + \cdots + \hat{\beta}_nX_n \tag{5}$$

The Linear Model can be used to deal with relatively simple data sets, but it can't fit the nonlinear relationship, and sometimes it has the problem of poor fitting. Zhang compared two machine learning models of linear regression and logistic regression, based on the same diabetes data set, to explore data scenarios applicable to both methods and the following results were obtained as given in Fig. 1

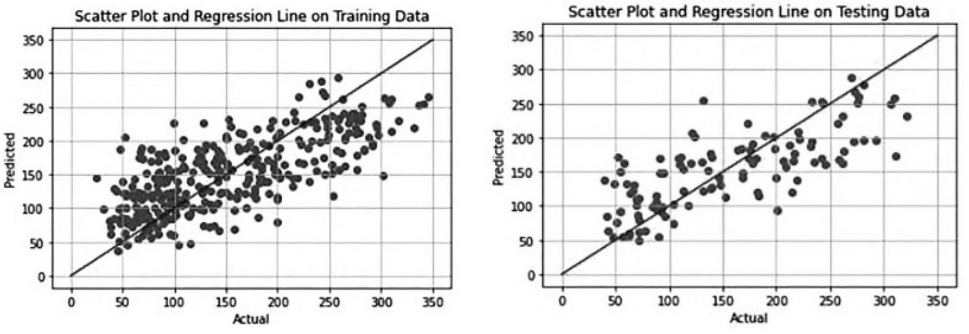

Figure 1. Scatter results of linear model [10].

Linear regression is often applied to machine learning problems with high linear connections, such as the prediction of house prices. Some models in Linear Models can deal with nonlinear relations better.

DOI: 10.54254/2755-2721/51/20241467

Logistic Regression Model has better performance in dealing with nonlinear data than Linear Regression Model [8]. Chen et al. established six machine-learning models based on four data sets from different countries and studied the application value of machine-learning methods in auto insurance fraud identification through a cross-analysis method. They found that the Linear Model is a shallow fitting model that performs well in fitting some poor-quality data sets since it does not overinterpret the relationship between independent and dependent variables [4].

4. Forest

The Forest Model is an ensemble learning method, which consists of multiple decision trees, each of which is a weak learner. By combining multiple weak learners, the prediction performance and stability can be improved [5]. A typical forest model in machine learning is Random Forest. Random Forest model is an integrated machine learning model obtained by applying the integrated learning method Random Patches to the decision tree model [4]. The basic idea of the decision tree model is to split the samples from the original training set into different groups according to the value of the explanatory variable to produce a prediction result. For instance, in dealing with a binary classification problem, the prediction probability value of a class of samples in the test set is the proportion of that class of samples in the training set in its group [4]. On the other hand, for regression problems, the output results of each node participate in voting, and the decision tree determines the final output results according to the minority rule [6]. Based on the participation of multiple decision tree single classifiers, the random forest can effectively reduce the variance of the predicted value of the decision tree model and greatly reduce the probability of classification errors, to achieve better prediction results. The Random Forest algorithm begins by randomly selecting N data by put-back draw from the original data set and generating a new training data set from these samples to train a decision tree. Repeat the above operations T times to form several decision trees, K features are randomly extracted from the feature set with put-back draw, and the decision tree divides the data based on these features which constitute a random forest, and the final output will be determined by their output results.

A Forest Model can often achieve high accuracy, due to involving several decision trees, As the weight of a single decision tree decreases, its influence on the result is reduced. The Forest Model can reduce the variance and bias of a single decision tree, prediction errors can be reduced. According to Chen et al, in research on the comparison of the mainstream machine learning model applied to the identification effect of auto insurance fraud, they used AUL to measure the prediction effect of the model based on two dimensions sensitivity and specificity [4]. The Forest Model has strong generalization ability because it is robust against overfitting and can also avoid underfitting through random sampling and random selection of features [5]. According to Wang et al., The Random Forest Model can be applied to gene selection and lung cancer prognosis, as well as to other cancer datasets. Moreover, the interpretability of forest models is low due to the large number of trees included in a forest, each of which is trained by a different data set, it is difficult to find a uniform standard to evaluate the contribution of each feature or sample to the model.

5. SVM

SVM is a supervised machine learning algorithm that is used to solve classification problems [6]. The SVM aims to identify a hyperplane that best separates the training data into different classes, whose core concept is to classify correctly plus maximize the margin. The distance between the hyperplane and the closest point from each class to this hyperplane, which is known as the margin, will be computed to determine the hyperplane [11]. Each point is a vector that contains data of several feathers used for production. Although infinite hyperplanes can separate data by features, there is only a unique optimal segmentation. The equation for linearly separable data in 1D and 2D is as follows:

$$f(x) = (wx + b) \tag{6}$$

Where x is the input (vector), w is the weight vector, and b is the bias. It is also possible to introduce a nonlinear interface to SVM by introducing the concept of a kernel function to generalize the linear

DOI: 10.54254/2755-2721/51/20241467

interface of the SVM classifier. Thus, data can be mapped to higher dimensions by kernel function [11]. In dealing with binary classification problems, the equation to solve the hyperplane is as follows:

$$min\varphi = \frac{1}{2} \|w\|^2 + C \cdot \sum_{i=1}^{n} \xi_i \quad (7)$$

The constraints on this expression are $yi(w \cdot xi) + b - 1 > \xi i$, where training input x is given as $x_i \in R$, where $i = 1, 2, \ldots, n$, and $y \in (-1, 1)$. ξi represents the stack value which transforms this inequation into an equation. $|| w ||^2$ represents the complexity and is used to evaluate the structure risk. Sigma represents the training error and is used to evaluate the experience risk. C is a factor used to balance the $|| w ||^2$ and sigma.

SVM is a novel and suitable small sample learning method with a solid theoretical foundation, which does not involve probability measure and the law of large numbers, and simplifies the usual classification and regression problems [4]. Cao et al. apply SVM to numerical rainfall forecast correction and state that SVM improved the accuracy of rainfall forecast most obviously among all models. Sun uses SVM to figure out a stock selection strategy and concludes that SVM can support a more accurate stock selection strategy than using random forest. Moreover, SVM adopts the principle of structural risk minimization for modelling, which minimizes the training error and reduces the generalization error, effectively avoids the overfitting phenomenon, and gives SVM have strong generalization ability [12]. Furthermore, The SVM algorithm determines the decision boundary by maximizing the interval and is robust to outliers, which can effectively avoid the influence of outliers on the result.

6. Comparison, limitations and prospects

Among these three models, the Linear Model is most suitable for small samples, while the Forest is more suitable for large samples in SVM [4, 11]. In the aspect of performance, in general, the performance of Linear Models is low since they can only capture the linear relationship of the data and cannot handle non-linear complex data [9]. Forest models and SVM have higher performance due to their improvements in the generalization and fitting ability of the model through ensemble learning or kernel functions. However, depending on the characteristics and quality of the data, sometimes a linear model can work well [4]. Linear Models are less complex because they involve only simple matrix operations and optimization methods such as gradient descent. Forest models and SVM are more complex because they need to build multiple basic models, such as training multiple decision trees in a random forest, which increases the training time and space overhead of the model. Forest and SVM are less interpretable than the Linear model manifested as SVM has a complex concept of kernel function while Forest has appreciable interpretability due to its intuitive decision tree structure.

Several limitations were discovered during this review. First, these machine learning models are applied in many fields, which results in very messy data, thus it is difficult to make a comprehensive and accurate comparison and summary. Moreover, the data sets used in different studies have different characteristics, and the quantity and quality of the data contained are uneven, which greatly affects the prediction effect of the model. According to the previous study, random forest is more suitable for small samples while SVM is more suitable for large samples, contrary to the study of Chen et al. Comparing the data sets of these two articles, it is found that the two data sets used by the former only have 33 features, and the records used for prediction are only 396 and 649 respectively. Compared with the data sets optimized by the SOMTE sampling method, the data sets used by the latter, the former are significantly smaller and of lower quality. Chen et al. argue that the quality of the dataset chosen by researchers will have a certain impact on the effect of machine learning. When the size of the data set and the quality of the data are different, the performance of the same model is different compared with its own, and the performance of different models will get different results. As a result, one cannot say which model is best, but only on a case-by-case basis when applying them to different problems, and conclusions drawn from one study may not necessarily generalize to the general situation. Looking ahead to the future, machine learning still has a long way to go. Deep learning, which belongs to the category of machine learning, has developed rapidly since 2000 and has gradually been widely used [2]. Next, researchers can consider how to design faster and more efficient deep learning algorithms to solve

DOI: 10.54254/2755-2721/51/20241467

more problems that are currently unsolvable [1]. Furthermore, researchers can further extend the success of deep learning from some static tasks to complex dynamic decision-making tasks.

7. Conclusion

With the rapid development of machine learning, several machine learning models and algorithms have emerged, and are appropriate in different scenarios. Through comparison, it is found that the linear model is relatively simple, and it is enough to solve some relatively simple linear models. Forest and SVM models are more complex and contain parameters that are difficult to interpret, but these parameters improve their robustness and they can fit better. The Forest is more suitable for solving problems such as identifying the effect of auto insurance fraud and student achievement prediction, while SVM is more suitable for solving problems such as stock prediction and weather prediction. Nevertheless, several limitations might confuse researchers, different data sets may give different results, so a large number of research and statistical results in different fields are needed to draw conclusions that can be generalized. Due to the wide application of deep learning nowadays, further research can be carried out on finding solutions for faster and more efficient deep learning algorithms based on more complicated tasks. This article gives a clearer perspective on the machine learning model by comparing the three machine learning models, furthermore, comparing their performance and principles can lead to a better understanding of the basic concepts and methods of machine learning, the selection of suitable models to solve practical problems, and the exploration of the direction and future challenges of machine learning.

References

[1] Liu M and Li T 2020 Overview and the prospect of machine learning technology development. Integrated Circuit Applications vol 10 pp 56-57.
[2] Chen Y, Guo X, and Tao H 2018 A Brief introduction to the research status and development trend of machine learning. China New Communications vol 8 p 173.
[3] Kamath U and Liu J 2021 Explainable artificial intelligence: An introduction to interpretable machine learning (Berlin: Springer International Publishing)
[4] Chen K and Li B 2022 A comparative study on the effectiveness of mainstream machine learning methods in identifying auto insurance fraud. Insurance research vol 12 pp 90-102.
[5] Wang Q, Zhou Y, Ding W, et al. 2020 Random Forest with Self-Paced Bootstrap Learning in Lung Cancer Prognosis. ACM Trans Multimedia Comput Commun Appl vol 16 p 34.
[6] Sun Y 2020 Stock selection strategy based on random forest and support vector machine. The 4th International Conference on Software and E-Business pp 53–56.
[7] Dumitrescu E, Hué S, Hurlin C and Tokpavi S 2022 Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research vol 297(3) pp 1178-1192.
[8] Gross K 2020 Machine learning and linear models: How they work. Retrieved from: https://blogdataikucom/top-machine-learning-algorithms-how-they-work-in-plain-english-1
[9] Zhang H 2022 Scenario analysis for linear regression and logistic regression. Automation and instrumentation vol 10 pp 1-4+8.
[10] Arnold K F, Davies V, de Kamps M, Tennant P W G, Mbotwa J and Gilthorpe M S 2021 Reflection on modern methods: Generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. International Journal of Epidemiology vol 49(6) pp 2074-2082.
[11] Alamri L S, Almuslim R S, Alotibi M K, Alkadi D, Ullah K I and Aslam N 2020 Predicting student academic performance using support vector machine and random forest. 3rd International Conference on Education Technology Management pp 100–107.
[12] Cao Z, Li Y, Hu Y, et al. 2020 Numerical rainfall forecast correction based on machine learning model. Water Diversion and Water Technology (Chinese and English) vol 10 pp 1-10.