Dean&Francis

ISSN 2959-6157

Feature Analysis with K-Nearest Neighbor and Random Forest

Liang Deng[1, *]

[1] Sendelta international academy school, Shenzhen, China

*Corresponding author: dunnnaabcdefg@gmail.com

Abstract:

The performance of a machine learning algorithm depends on the algorithm's complexity, and the feature representation of data is another critical factor. A standard feature representation method is based on expert prior knowledge, which transforms data into discrete feature representation, and each attribute represents the vital information in prior knowledge for the task. However, in practical application, the prior knowledge and the actual performance of the model are often contradictory, so it is essential to identify the features that can improve the machine learning model in machine learning, data mining and statistical modelling. Specifically, feature selection aims to identify and delete irrelevant or redundant features in the data while retaining those features that contribute the most to the model's prediction performance. By removing irrelevant or noisy features, feature analysis can improve the model's accuracy, generalization ability and robustness. This paper discusses the influence of feature selection on the performance of K nearest neighbour (KNN) and random forest (RF) algorithms in machine learning applications. The model's accuracy change is observed and analysed by deleting the features in the data set one by one. Then, the feature importance ranking provided by the random forest algorithm is compared to reveal their correlation and difference in feature selection. The experimental results show that although the two methods differ in feature evaluation, they can effectively guide feature optimization and improve model performance.

Keywords: Feature analysis; K-Nearest neighbor; Random Forest; Machine learning.

1. Introduction

The catastrophic sinking of the Titanic is a painful history lesson, reminding people of the dangers of sea travel and the devastating consequences of negligence. So far, using various methods to analyze the survival probability of people in catastrophic events has become the research direction that researchers pay attention to. In machine learning, researchers collected the passenger list of Titanic, including detailed records of everyone on that disastrous voyage. This data set contains detailed information about passengers' age, gender, ticket class, paid fare and living status. Using machine learning technology; researchers can analyze the functional relationship between different attributes and passengers' living conditions and suggest similar events to improve the survival rate.

This paper uses two machine learning algorithms, k-nearest neighbour (KNN) and random forest, to reveal the role of different passenger characteristics in predicting survival status. This paper chooses KNN because, as a lazy learning method, the KNN algorithm only depends on limited parameters, such as the selection of the K value. Therefore, the influence of different feature settings on the results can be analyzed more cleanly [1-3]. On the contrary, the random forest algorithm belongs to multi-parameter integrated learning technology, which integrates the results of multiple decision trees to achieve accuracy and reliability. Random forest reduces excessive fitting and improves the robustness of the model by introducing randomness at the feature and sample levels. This paper analyzes the importance of features by comparing the performance of the above two models in different experimental settings [4].

Specifically, this paper systematically studies how the above two machine learning models respond to including or excluding specific features. By repeatedly stripping off each feature and reevaluating the model's performance, this paper can quantify the relative importance of each attribute in predicting survival rate. This method reveals how KNN and random forest algorithms perceive and process data and reveals the complicated interaction between features and prediction results. The experimental results show that although the two methods differ in feature evaluation, they can effectively guide feature optimization and improve model performance.

2. Methodology

2.1 Feature Analysis with KNN Algorithm

As a cornerstone of supervised learning, the K-nearest neighbour (KNN) algorithm represents a simple and powerful classification technique that operates primarily on the principle of distance, as shown in Figure 1. The method classifies unknown instances by comparing their features with known instances in the training set and uses k nearest neighbour labels for prediction. Crucially, the effectiveness of this distance-based approach depends heavily on the choice and quality of the features used. Therefore, selecting features directly affects the model's ability to distinguish patterns and ultimately determines its prediction accuracy.

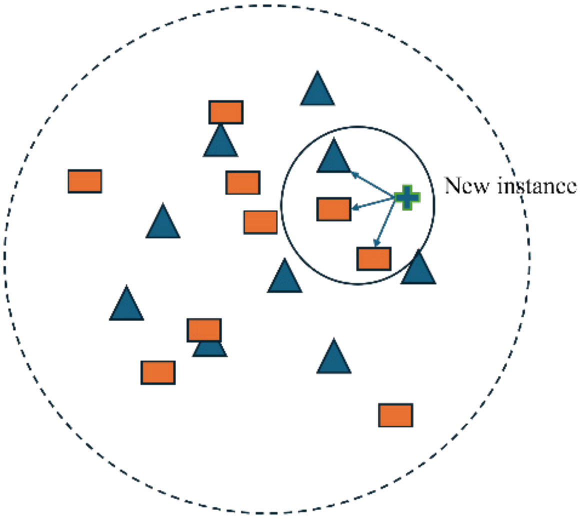

Fig. 1 The K-nearest neighbour method

As an illustrative example, consider removing the "number of family members" feature from a hypothetical data set designed to predict the probability of survival of an individual or family group in a disaster scenario. This characteristic is often considered vital to assessing resource allocation, support networks, and evacuation strategies and may contain valuable information about survival outcomes. However, the study hypothesizes that in some cases or specific contexts, this feature may contribute little to the model's accuracy or introduce noise, leading to suboptimal predictions. By eliminating this feature and observing subsequent changes in prediction accuracy, researchers can gain insight into its factual significance and assess whether its exclusion enhances or weakens the model's performance [5].

2.2 Feature Analysis with Random Forest Algorithm

Random forest is an ensemble learning technique that has gained widespread acclaim for its robust performance in various predictive modelling tasks and is a powerful strategy for enhancing prediction accuracy. By harnessing the collective intelligence of multiple decision trees, each trained on a different subset of data and randomly selecting features at each split, random forests mitigate the overfitting that can plague individual decision trees. This integrated approach not only improves the model's overall stability and generalization ability but also enhances its ability to capture complex relationships in the data.

A significant advantage of random forests is their inherent ability to gain insight into the relative importance of individual elements. Through the random forest algorithm, the contribution of each feature to model prediction can be systematically evaluated, as shown in Figure 2. This is achieved by measuring how much the model's prediction accuracy decreases when specific eigenvalues are randomly arranged or shuffled in a random (OOB) sample [6].

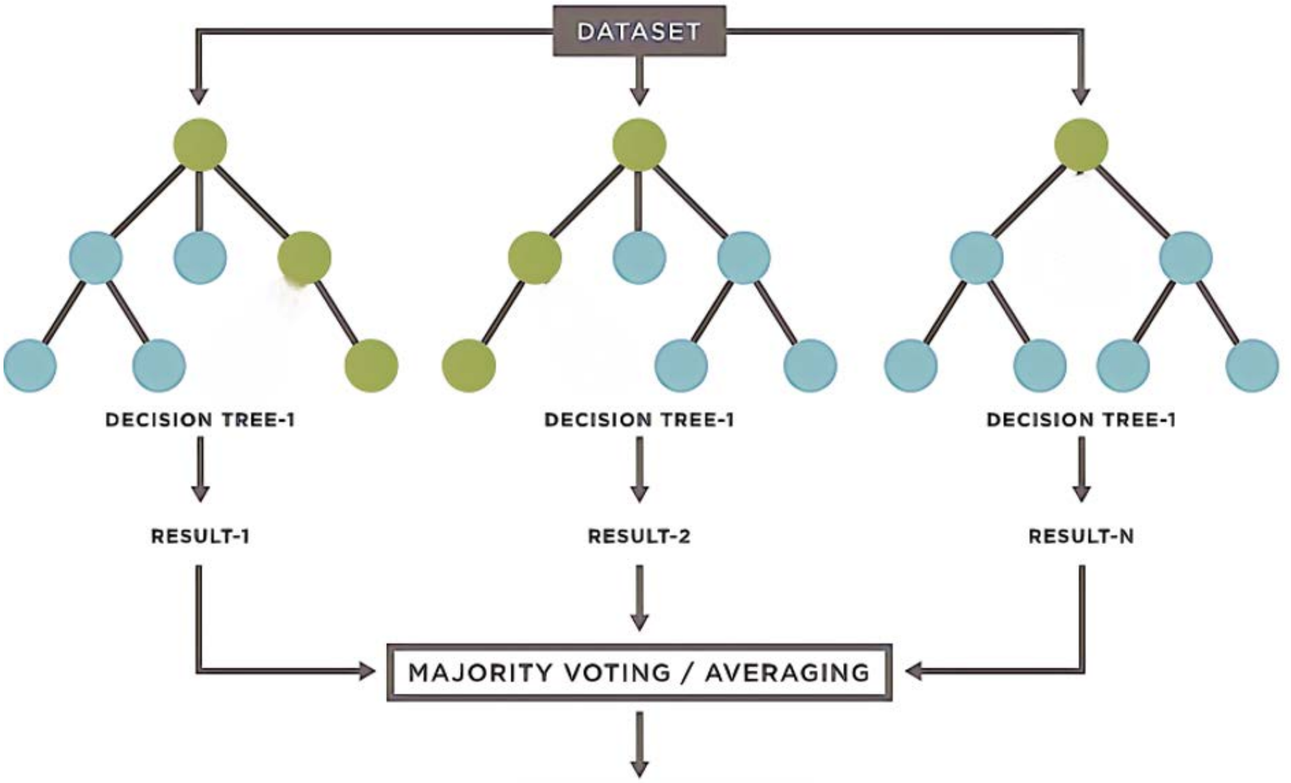

Fig. 2 The random forest algorithm

3. Result Analysis

3.1 DataSet Description

This paper uses a publicly available Titanic dataset containing passenger characteristics such as age, gender, ticket class, ticket price, number of family members, and a label indicating whether passengers survived. Before the model is implemented, the dataset goes through several pre-processing steps. The median of the respective features was used to estimate missing values in age and cabin characteristics. The data set is then split into a training set and a test set in a ratio of 80:20. The training set is used to train KNN and random forest models. In contrast, the test set is reserved for evaluating the model's performance.

3.2 Model Implementation

The KNN algorithm is implemented using the sci-kit-learn library in Python, and the neighbourhood number (k) is optimized by cross-validation. The random forest algorithm is also implemented using scikit-learn, with the number of trees set to 100 and the maximum depth limit of the tree to 10. These parameters are selected based on initial testing to ensure optimal performance.

Each feature is individually removed to assess the impact of feature elimination on the KNN model, and the model is retrained and evaluated on the test set. Performance is measured using accuracy, precision, recall, and F1 scores. The process is repeated for all features, and the results are compared to baseline model performance, which includes all features. The random forest algorithm inherently provides a feature importance score based on the average reduction of impurities. These scores were compared with the effect of feature elimination on the KNN model to derive correlations between the two algorithms' assessments of feature importance.

3.3 Results

The results show that excluding certain features, particularly the "kind of vote," significantly declined the model's prediction accuracy. This phenomenon underscores the importance of these characteristics, in particular "ticket class", in accurately predicting the probability of passenger survival. This finding highlights the need to incorporate these features into the modelling process, as they contain valuable information that can help distinguish between survivors and non-survivors.

As a complement to the KNN algorithm, subsequent applications of the random forest algorithm reveal nuances in the importance of features. The results show that "ticket price" and "age" become the key characteristics that significantly affect the model's predictive ability. These findings echo the results of the KNN analysis, suggesting a commonality in the correlation of these features across different machine-learning methods. However, random forest algorithms also reveal potential differences, hinting at each approach's unique strengths and perspectives. The ensemble nature of a random forest combines multiple decision trees, allowing it to capture complex interactions and nuances in data that may not be immediately obvious to a single model like KNN. Thus, while both algorithms identify "fare" and "age" as key features, random forests may assign slightly different weights or importance scores based on their ability to assess their combined impact on the entire dataset.

In addition, random forests' inherent feature importance assessment mechanism provides a more fine-grained understanding of how each feature affects overall predictive performance. This confirms the importance of "fare" and "age" and opens the way for further exploration of how other characteristics interact or influence these key predictors. Ultimately, the juxtaposition of KNN and random forest algorithm results provides a more comprehensive view of the forecast landscape, informs data-driven decisions, and enhances the accuracy and interpretability of predictive models.

4. Discussion

In this paper, representation learning plays a more critical role in improving model performance besides studying the algorithm principles of different machine learning algorithms. By analyzing the performance of the K- the nearest neighbour algorithm and random forest algorithm on the Titanic passenger survival data set and comparing the effects of removing different features on the model performance, this paper systematically expounds on the importance of different features in the survivor prediction task. This paper holds that an important bottleneck to improving the machine learning model's performance comes from the consistency between feature engineering based on expert prior and the machine learning model based on statistical learning. Among the two machine learning algorithms selected in this paper, the KNN algorithm relies on the proximity of adjacent data points, highlighting the characteristics that make the most significant contribution to distinguishing different categories or results. On the other hand, through the set of decision trees, the random forest algorithm evaluates the influence of each feature on multiple decision paths. The differences observed in evaluating the feature importance between the two algorithms emphasize the complexity and versatility of predictive modelling. These differences remind researchers that relying only on the results of a single algorithm may lead to a narrow or biased understanding of data and its potential patterns. Therefore, it is imperative to adopt the multi-algorithm method in practical application, in which the results from multi-algorithms are considered and compared to achieve a more comprehensive and robust understanding of the importance of features. In order to ensure the generalization of the results, in future work, this paper will design a complete experimental scheme, including different data sets and adopt various analysis methods to ensure that the experimental conclusions are not limited to any single data set or the details of algorithm implementation.

5. Conclusion

Before constructing a specific machine learning model, analysing the different features in the data set is indispensable to obtain a high-performance machine learning system. The main goal of feature selection is to identify and eliminate irrelevant or redundant features while carefully retaining those features that have significantly contributed to enhancing the model's prediction ability. Because different features in the data set have different degrees of importance and influence on the model results, by systematically pruning unnecessary or noisy elements, feature analysis can enhance the accuracy, generalization ability and robustness of the model and eliminate the noise generated by irrelevant features in the model learning process. Based on two typical machine learning models, KNN and random forest, this paper studies the influence of different characteristics on the prediction of passenger survival in the passenger data set retained in the Titanic incident. In this paper, the characteristics of the data set are deleted by controlling variables, and the subsequent fluctuations of model accuracy are carefully observed and analyzed. The experimental results show that although KNN and random forest are based on different algorithm principles, the results are satisfactory. In addition, the results of KNN show that in the data set studied in this paper, all kinds of features are indispensable in predicting the probability of survivors, and the effects of different features on the model results are slightly different.

ISSN 2959-6157

References

[1] Abdul Samad, Salih TAZE, Ucar M. Enhancing milk quality detection with machine learning: A comparative analysis of KNN and distance-weighted KNN algorithms. International Journal of Innovative Science and Research Technology, 2024, 9(3).

[2] David Wijaya, Anastasia Rita Widiarti. Batik classification using KNN algorithm and GLCM features extraction//E3S Web of Conferences. EDP Sciences, 2024, 475: 02012.

[3] Ahmed M Neil, Eman Shabaan, Mervat El Qout, et al. Machine Learning Based Approaches For Android Malware Detection using Hybrid Feature Analysis//2024 6th International Conference on Computing and Informatics (ICCI). IEEE, 2024: 158-165.

[4] Tony Wayne Wang. Survival Prediction and Comparison of the Titanic based on Machine Learning Classifiers. Transactions on Computer Science and Intelligent Systems Research, 2024, 5: 443-450.

[5] Jiale Li. Survival prediction and analysis of Titanic based on logistic regression and KNN. Transactions on Computer Science and Intelligent Systems Research, 2024, 5: 568-572.

[6] Yang Liu. Casualty on the Titanic based on Machine Learning Methods. Highlights in Science, Engineering and Technology, 2023, 39: 1364-1376.