Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230683

Comparison study of data adaptability based on k-nearest neighbor(KNN), multilayer perceptron(MLP), and decision tree(DT)

Yitong Liu

Applied Mathematics, University of California Santa Barbara, United States

yitongliu@ucsb.edu

Abstract. In the current context of increasingly popular and developed human-computer interaction, machine learning (ML) and data mining are becoming more and more important in various areas, such as image detecting, medicine, and commercial companies. Different researchers from various fields have considered ways to improve the correctness rate for data mining. This paper used data sets from Kaggle and UCI machine learning repository, by applying k-nearest neighbor (KNN), multilayer perceptron (MLP), and decision tree (DT) classifiers to these data sets, the confusion matrix shows the result of the correctness rate for different classifiers under these data sets. As a result, the confusion matrices have shown that data adaptability is based on the data sets' characteristics which rely on different classifiers' specialties.

Keywords: machine learning, classification, KNN, multi-layer perception, decision tree, data mining.

1. Introduction

With the development of machine learning (ML), researchers from various fields and areas have worked on the problem of data mining. While creating classifiers is one of the widely used techniques in data mining [1], figuring the data adaptability for different classifiers is essential for data mining. Data mining is a technique that uses statistics and artificial intelligence to discover or extract new patterns from large data sets [2]. One of the wide research areas is data adaptability for different classifiers. The goal for classifiers is to predict the goal class with the highest precision and the classification algorithm help people find out the tie between the input attribute and output attribute to construct a model that is a training process [3]. While different machine learning-based classifiers perform differently for different situations. Sometimes, one of the machine learning algorithms gives better accuracy in rightfully classifying the fault in some datasets, whereas at another instant, it performs weakly for a different data set [4]. Therefore, this paper will focus on the suitability of different classifiers for four different data sets from medicine, music, material, and plant areas. By using k-nearest neighbor (KNN), multi-layer perception (MLP), and decision tree (DT) based classifiers to analyze four different datasets from Kaggle and UCI machine learning repositories in different areas, the constructed confusion matrices based on these three classifiers show their correctness rate for different data sets. After analyzing these results and comparing them with other research, the conclusion for this topic is that the

DOI: 10.54254/2755-2721/6/20230683

data adaptability for different data sets is different concerning different classifiers and every classifier has its specialty. This topic is meaningful not only for various areas to help them find the most suitable classifier for their dataset but also for future researchers to improve the classifiers based on every classifier's weaknesses.

2. K-Nearest neighbor

2.1. Method

K-nearest neighbor (KNN) is a non-parametric model which was invented by Dudani in 1976, and in 1983 Jozwik induced an enhanced version [5]. This kind of algorithm uses the formula (1) [6] of distance to measure the arbitrary distance between the target point and the closest clouds of neighbors for k¿=1. In other words, KNN uses the training instance as the predicted value label for the arbitrary instance. Since this kind of algorithm doesn't utilize the training data points to do any generalization, it is also called a lazy algorithm [6]. Based on this characteristic, it is also analytically tractable and very easy to implement [7]. As a result, KNN has lately been recognized as one of the top data mining algorithms [5]. By previous research, KNN performs well in disease prediction especially in HD prediction since it works with a single distance [7].

2.2. Result

Based on the result of our four data sets, the confusion matrix shows that the result (figure1, figure2, figure3) of KNN in glass genre [8] classification, the correctness for KNN is around 62.6%, whereas the result for MLP is around 42.99%, and the result for DT is around 58.9%. The performance of MLP is the worst while the performance of KNN is the best for this data set.

2.3. Analysis and discussion

The reason for this result is that the data set has enough training data for KNN and it also has a good discriminating distance between each genre. Since the KNN algorithm is based on distance, it is necessary to have a training set that is not too small, and a good discriminating distance. Therefore, it also performs well in multi-class simultaneous problem solving [3].

$$d(x_i, x_j) = (\sum_{s=1}^{p}(x_{is} + x_{js})^2)^2 \text{ [6]} \tag{1}$$

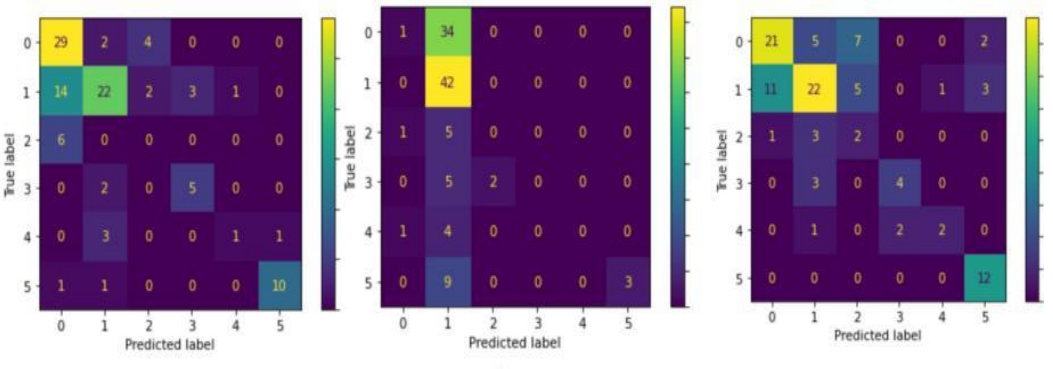

Figure 1. Confusion matrix for glass genre [8] in KNN, MLP, and DT.

DOI: 10.54254/2755-2721/6/20230683

3. Multi-layer perception

3.1. Method

The basic structure of the multi-layer perceptron (MLP) is composed of simulations and simplifications of the biological neuron model, which are composed of dendrites, cell bodies, axons, and other parts. MLP is commonly formulated as a multivariate nonlinear optimization problem over a very high-dimensional space of possible weight configurations [6]. Neurons are the main parts of multilayer perception, and there are clusters of neurons in each layer [5]. Classic MLP works with three types of layers. The prior layer receives inputs from its counterparts in the following layer. Before sending it to neurons in the subsequent layer, it analyzes the information by utilizing the activation function [4]. A hidden layer lies between the input layer and the output layer. These three layers are fully connected, which means every neuron in the upper layer is connected to all neurons in the lower layer. Based on these, formula (2) is the function for MLP.

3.2. Result

Based on the result of our data sets, MLP has the best accuracy in the liver disorder data set [9] (figure 4, figure 5, figure 6), which has 65.7% accuracy in MLP, 57.5% accuracy in KNN, and 62.2% accuracy in DT. In this data set, the performance of MLP is the best while the KNN is the worst.

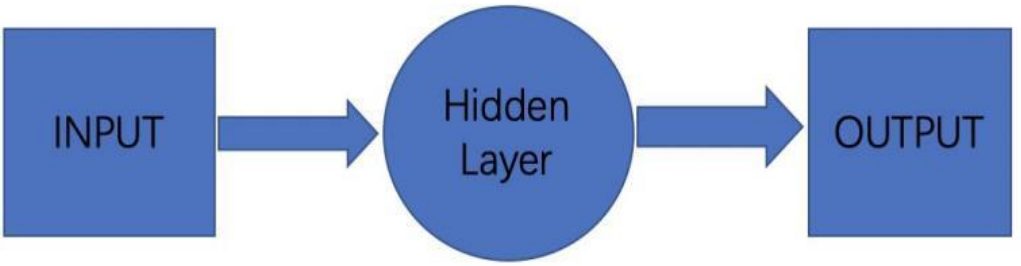

Figure 2. MLP working process.

3.3. Analysis and discussion

The result shows that MLP is good at predicting categorical and continuous variables [4]. The reason that KNN performs badly is that the data set does not have a discriminating distance. In addition, based on previous research, MLP was also found to be superior for predicting the production process of bio-diesel [4]. During the process of coding, it also shows that MLP has better accuracy for larger data sizes but the speed is slower.

$$f(x) = G(b^{(2)} + W^{(2)}(s(b^{(1)} + W^{(1)}x))) \text{ [10]} \quad (2)$$

DOI: 10.54254/2755-2721/6/20230683

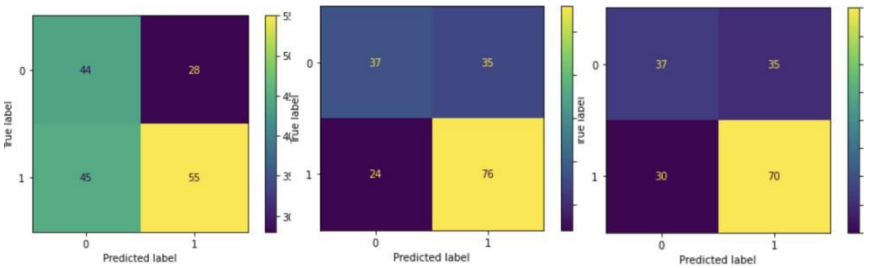

Figure 3. Confusion matrix for liver disorder [9] in KNN, MLP, and DT.

4. Decision Tree

4.1. Method

The decision tree is a widely used classification method in data mining which is commonly used in marketing, surveillance, fraud detection, and scientific discovery [2]. Since it is a tree-based classifier, it follows the same structure as a normal tree. It is structured by the root node, branches, and leaf node [11]. This kind of classifier is a technique in which any path beginning from the root is described by a data separating sequence until a Boolean outcome at the leaf node is achieved [3]. In other words, the training data are partitioned into several subsets according to the values of the splitting attribute. It is easier to see how it works from graph 7 [12].

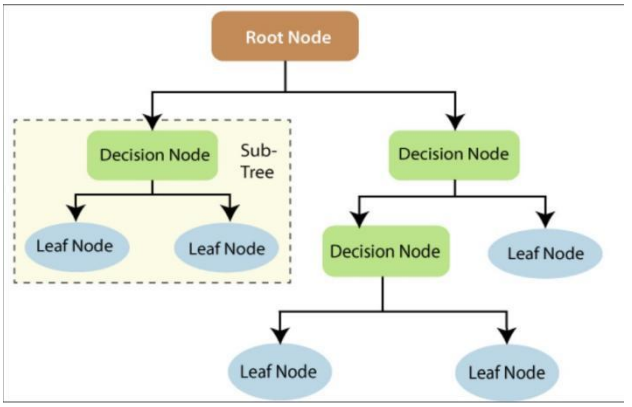

Figure 4. Decision tree algorithm [12].

4.2. Result

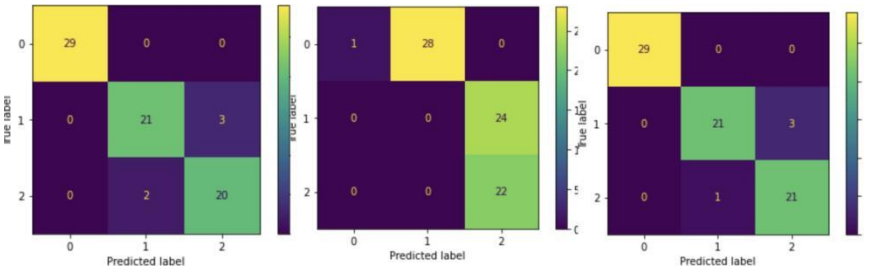

Figure 5. Iris data set confusion matrix in KNN, MLP, and DT.

DOI: 10.54254/2755-2721/6/20230683

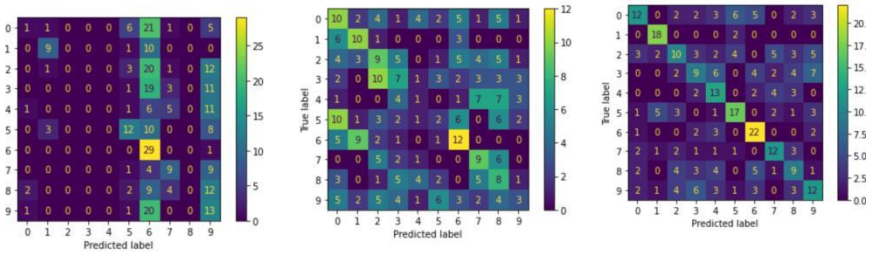

Figure 6. Music genre data set confusion matrix in MLP, KNN, and DT.

Based on our data sets, the music genres [13]and iris [14] perform best in the decision tree algorithm. The accuracy for the music genre in KNN is 23.67% and in MLP is 24.3%, while the accuracy for DT is 44.67%. This data set performs best in the decision tree and has the worst performance in KNN. On the other hand, the accuracy for iris in KNN is 93.3% and 30.67% in MLP, while the accuracy in DT is 94.67%. This data set performs best in the decision tree and has the worst performance in MLP.

4.3. Analysis and discussion

Based on the result of the accuracy test in the music genre [13], it is normal to have such a low accuracy because there are too many labels in this data set since the more labels, the lower the acc will be. The reason that the decision tree performs best is that there is a lot of noise in the features in this data set, which makes it sparse. Therefore, it takes many calculations for KNN and MLP. On the contrary, since the decision tree (DT) is based on the information gain ratio, it can find the best feature directly.

5. Conclusion

Based on all these results of data sets, we can draw the conclusion that these three classification algorithms have their specialty in data adaptability. KNN works better with data sets that have a not too small data size and a good discriminating distance. It performs well in multi-class simultaneous problem-solving. On the other hand, MLP is good at predicting categorical and continuous variables. Moreover, when there is a lot of noise in the features in the data set, it can bring lots of calculations for KNN and MLP, which leads to results of low accuracy. However, since the decision tree (DT) is based on the information gain ratio, this cannot affect the accuracy of this kind of classification. This paper is not completely perfect since the data sets are not completely suitable for the comparison. Future research can focus on how to improve these classification algorithms so that they can be suitable for more types of data sets.

References

[1] B. Jijo and A. Mohsin Abdulazeez, "Classification based on decision tree algorithm for machine learning," Journal of Applied Science and Technology Trends, vol. 2, pp. 20–28, 01 2021.

[2] R. PatelBrijain and K. K. Rana, "A survey on decision tree algorithm for classification," International Journal of Engineering Development and Research, vol. 2, pp. 1–5, 2014.

[3] S. Amendolia, G. Cossu, M. Ganadu, B. Golosio, G. Masala, and G. Mura, "A comparative study of k-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening," Chemometrics and Intelligent Laboratory Systems, vol. 69, no. 1, pp. 13–20, 2003.

[4] S. Qi, W. Zhao, Y. Chen, W. Chen, J. Li, H. Zhao, W. Xiao, X. Ai, K. Zhang, and S. Wang, "Comparison of machine learning approaches for radioisotope identification using nai(ti) gamma-ray spectrum," Applied Radiation and Isotopes, vol. 186, p. 110212, 2022.

[5] X. Sun, M. J. C. Opulencia, T. P. Alexandrovich, A. Khan, M. Algarni, and A. Abdelrahman, "Modeling and optimization of vegetable oil biodiesel production with heterogeneous nano catalytic process: Multi-layer perceptron, decision regression tree, and k-nearest neighbor methods," Environmental Technology Innovation, vol. 27, p. 102794, 2022.

DOI: 10.54254/2755-2721/6/20230683

[6] Kayabasi, "Mlp and knn algorithm model applications for determining the operating frequency of a-shaped patch antennas," International Journal of Intelligent Systems and Applications in Engineering, vol. 3, pp. 154–157, 09 2017.
[7] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," Computers in Biology and Medicine, vol. 136, p. 104672, 2021.
[8] B. German, "UCI machine learning repository," 2017.
[9] R. S. Forsyth, "UCI machine learning repository," 2017.
[10] wepon, "Deep learning tutorial3mlp multilayer perceptron principle introduction + code details," 1 2015.
[11] H. Patel and P. Prajapati, "Study and analysis of decision tree based classification algorithms," International Journal of Computer Sciences and Engineering, vol. 6, pp. 74–78, 10 2018.
[12] C. Janikow, "Fuzzy decision trees: issues and methods," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 28, no. 1, pp. 1–14, 1998.
[13] Olteanu, "Gtzan dataset - music genre classification," 2020.
[14] R. Fisher, "UCI machine learning repository," 2017.