INTRODUCING THE FUNDAMENTAL OF THREE CLASSIFICATIONS: DECISION TREE, MAXIMUM LIKELIHOOD & NEURAL NETWORK FROM DIFFERENT EXAMPLES

J.Y. Li
School of Computer Science and Engineering
Australian National University

Abstract: This paper presents the introductions and applications of three different algorithms, including decision tree, maximum likelihood, and neural network. In the neural network part, it will be applied into a wine dataset for classification with the detail, and in the decision tree part, the detail of how to use also will be shown with the data about purchases online. Besides, the decision tree and maximum likelihood will be compared with the neural network separately based on their features in the conclusion.

Keywords: Classification, Neural Network, Decision Tree, Maximum Likelihood

1 Introduction and background

The applications of Neural Networks are various, and six of those applications are distinguished, including the recognition of images, digital games, the generation and recognition of voices, the imitation of drawing, prediction, and the revision for better design on website. Take an example to make it explicit, commonly, we can distinguish objects artificially, but if the quantity is huge, for example, classifying the type of vegetation in a large range will consume a great deal of effort, time and money. Nowadays, the application of satellite data is mature than before, such as the satellite map on Google or Baidu, but it still not enough to supply the detailed information without neural network techniques (Omatu & Yoshida, 1991).

This study will focus on the classification on different type of data. For learning different methods, except for the basic example for how to implement Neural Network Classification, another two non-machine learning methods including Decision Tree Classification, Maximum Likelihood Classification also will be discussed and compared with Neural Network Classification.

2 Methodology

2.1.1 Introduction of Decision Tree Classification

Decision Tree Classification is one kind of the prediction models which represent the mapping relationship between the value of the object and the object. A data set will be divided into smaller parts as the different branches recursively, and each branch includes a set of tests, which is the definition of decision tree and it is

actually a classification procedure (Friedl & Brodley, 1997). ID3 algorithm and C4.5 algorithm are two major algorithms which are based on Decision Tree. We will introduce one of them below.

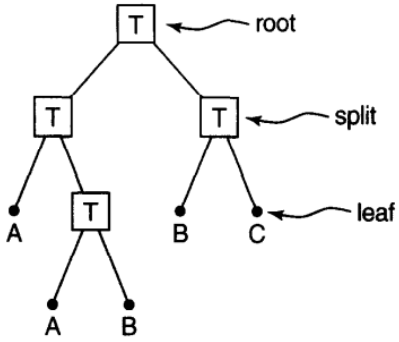

Figure 1. Decision tree classifier

2.1.2 Implement of ID3 algorithm

This section, the ID3, Iterative Dichotomiser 3 algorithm will be introduced to make everything as simple as possible which is the theory of Occam's Razor as well (Sheffer, 2014). The smaller decision tree generated by the algorithm will be better than the bigger one. To get a decision tress based on ID3 algorithm, there are two elements needed, entropy of information and information gain.

The formula of entropy of information, x represents the random variable, and the p is the probability of x.

$$H(x) = -\sum_{i=1}^{n} p_i log_2 p_i$$

The formula of information gain, S represents the assemblage of samples. The value(T) is the assemblage of value T, and v represents one of the values of T. S_v is the sample which is a specific value of T.

$$IG(S|T) = Entropy(S) - \sum_{value(T)} \frac{|S_v|}{S} Entropy(S_v)$$

The interpretation above seems to be a bit vague, but after introducing the example below, it will become more explicit (**Figure 2**).

Age	Income	Student	Level of credit	If buy PC
Youth	High	No	Fair	No
Youth	High	No	Excellent	No
Middle aged	High	No	Fair	Yes
Senior	Medium	No	Fair	Yes
Senior	Low	Yes	Fair	Yes
Senior	Low	Yes	Excellent	No
Middle aged	Low	Yes	Excellent	Yes
Youth	Medium	No	Fair	No
Youth	Low	Yes	Fair	Yes
Senior	Medium	Yes	Fair	Yes
Youth	Medium	Yes	Excellent	Yes
Middle aged	Medium	No	Excellent	Yes
Middle aged	High	Yes	Fair	Yes
Senior	Medium	No	Excellent	No

Figure 2. Dataset from the AllElectronics customer database.

The main point of creating a decision tree is achieving the feature which has the max information gain as the node. As the table shown, we have five features to decide if the people will buy a PC, and we will calculate the information gain of them step by step. First, the table shows that there are 9 of 14 decide to buy a PC, and 5 of 14 decide not to buy a PC. Then, the Entropy of it is Entropy (S) = - 9/14 $\log_2 9/14$ – 5/14 $\log_2 5/14$ ≈ 0.940. Beginning with Age, H(Youth) = -2/5 $\log_2 2/5$ – 3/5 $\log_2 3/5$ ≈ 0.971, H(Middle aged) = -1 $\log_2 1$ = 0, H(Senior) = -3/5 $\log_2 3/5$ – 2/5$\log_2 2/5$ ≈ 0.971. Then, calculating the whole Entropy of Age is H (outlook) = 5/14 * 0.971 + 4/14 * 0 + 5/14 * 0.971 ≈ 0.694. Finally, the information gain of Age is IG (Age) = 0.940 – 0.694 = 0.246. Following this process to continue to achieve the other information gain gradually is the way. The information gain of income is 0.029, the information gain of status that if the customer is student is 0.152, and the information gain of level of credit is 0.048. Through the comparison among them, it is clear that the information gain of Age is biggest one, so the root node is Age. Then, to decide the next node, the information gain of income, student, and level of credit separately based on the known condition that the age belongs to youth or senior are necessary, then doing the comparison following the method mentioned above. Finally, a final decision tree through ID3 algorithm come out (**Figure 3**). Besides, I try to apply the decision tree into the wine dataset, but it cannot work well, since each value of the attributes is different, which cannot use the probability as shown above.

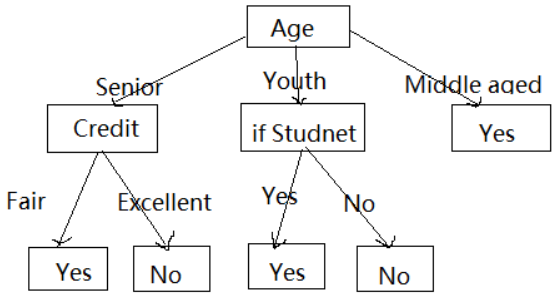

Figure 3. Result of the decision tree

2.2 The strengths and weaknesses of Decision Tree Classification

The computation complexity of Decision Tree algorithm is not that difficult, and the amount of necessary data is not huge as well, as the example shown in the above section, just like some mathematical problems. However, if there are too many matching options, it will cause the overfitting problem. Besides, the result from Decision Tree is not stable, because a tiny change in the data may make it generate a different Decision Tree.

3 Maximum Likelihood Classification

To achieve the maximum value in a known class distribution, Maximum Likelihood classification (MLC) is the suitable algorithm (Scott & Symons, 1971). Let me take an example to introduce the theory explicitly. Assumed that there are two roommates denoted A and B. They both cook after ten in the morning for an entire month. The frequency of which roommate A cook during the month was 20, while roommate B was 10. Someday, roommate C found there were someone cooking after 10 o'clock and the people there was supposed to be A.

For another example, assumed that two boxes, A and B. A contains 95 red balls and 5 black balls, while B contains 60 red balls and 40 black balls. Now, if someone takes out a ball from a box, and find it was a red one, others will suppose that box is A. Event x represent the probability of red balls. We already know P(x|A) is 0.95 and P(x|B) is 0.6. To compare P(A|x) and P(B|x), which can apply in Bayes'theorem which describes the probability of an event, using the previous experience that is probably relevant to the event. The formula for Bayes' theorem is P(A|B) = P(B|A) * P(A)/P(B). Then P(A) is equal to P(B) is 0.5. Then we know that P(A|x) is bigger. Maximum Likelihood Classification can be called Bayes classification.

3.1 The strengths and weaknesses of Maximum Likelihood Classification

This algorithm provides a high speed on dealing with a great deal of training and inquiring data, but this algorithm also assume that each sample is independent, which means the samples will have some bad influences, if they have some relationships.

4 Neural Network Classification

The neural network based on back propagation for classification problem is set to be in the feed ward direction, and has three layers, including one input layer, one hidden layer and one output layer (Yee, 2016).

4.1 Back Propagation algorithm

Back Propagation algorithm normally is used for prediction and the applications on classifications, and a set of known data for input and samples for output are necessary.

Back Propagation algorithm aim to minimize the error of output in total and achieve the minimum error sum of squares (Hameed, Karlik, & Salman, 2016). Back Propagation algorithm has two processes, which are forward and backward. In the process of feed-forward, it goes through the hidden layer from input layer to output layer and calculate the value of input and the value of output as the **Figure 4** shown. Wij represent the weights between neuron I and j, and Oj represent the output of the neuron. The backward part is for the feedback of error, and it goes in opposite direction.

$$I_j = \sum_i W_{ij} O_i$$

$$O_j = sigmod(I_l) = \frac{1}{1 + e^{-I_l}}$$

Figure 4. The relationship between input and out put

4.2 Experiment

Taking a simple example based on Wine dataset. A list of different chemical analysis was given by the Wine dataset to classify the type of different wines, such as Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyannis, Color intensity, Hue, OD280/OD315 of diluted wines and Proline. Then, 160 samples for training in the neural networks and 40 samples for testing in the neural networks were set. **Figure 5** was gotten, and after training the model, normally the misclassification will be reduced, which shown in **Figure 6**.

In the experiment, the original dataset was divided into two parts for training and testing. In the dataset, the last column represents the type of the three different wines. Because in computing language, it starts from 0, type 1 to 3 was changed into type 0 to 2 at the beginning. Then the dataset was loaded for training and a two layers network was defined. The activation function of the network was set to be Sigmoid function as the formula shown in the above part (**Figure 7**). Regarding the neurons for each layer, 13 neurons were decided to add in the input layer which was also the amount of the attributes of the dataset, and 27 neurons were decided to add in the hidden layer which was based on the Kolmogorov theorem. If there are x neurons in the input layer, the number of neurons in the middle layer should be 2x+1 (Hecht-Nielsen, 1987). Finally, 3 neurons were decided to add in the output layer which was equal to the amount of type of wines. Besides, the learning rate was set to be 0.01 because we tried different learning rates, such as 0.1, 0.001 and find the loss value in 0.01 change a lot, and the epoch was set to be 500. However, as the confusion matrix shown, the outcome is not the ideal situation. Only type 1 and type 2 here can be distinguished well while comparing to the type 3. When applying the program into another dataset, like Iris Data Set (Marshall, 1936), which is a data set using 4 attributes to distinguish three types of flowers, the program worked well to distinguish the three types of flowers with same epoch and neurons in each layer as Figure 9 shown. So, maybe some problems occur in the program to the dataset that haven't been found out. This will be continued to work on in the future work.

```
Confusion matrix for training:

 46   8   0
  5  60   0
  4  37   0
[torch.FloatTensor of size 3x3]

 Accuracy: [25.625
```

Figure 5. Confusion matrix for training wine data

```
Confusion matrix for testing:

 18   1   0
  1  16   0
  0   4   0
[torch.FloatTensor of size 3x3]

Testing Accuracy: [60.
```

Figure 6. Confusion matrix for testing wine data

```
Confusion matrix for training:

 37   0   0
  0  40   4
  0   0  39
```

Figure 7. Confusion matrix for training Iris data

4.3 The strengths and weaknesses of Neural Network Classification

Neural Network Classification has a high accuracy, and it can be trained to learn well. But, the process is in the black box which cannot be viewed, and the parameters of them are quite a few and complicated. Besides, Neural Network Classification need a long time to learn, which may cause the local minimum.

5 Dataset

The used dataset is about classifying the wine into three different classes based on 13 attributes, including Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, hue, OD280/OD315 of diluted wines, Proline. According to the donor of the dataset Stefan Aeberhard (1991) stated, these used data were ensured that the environment variable of the chemical analysis of the different wines was limited, and they all collected in the same region in Italy. But the initial data from Forina, M. et al actually had 30 variables, and the rest of the data was lost by the donor. The reason for choosing this data is because this data set is sorted into dealing with the task which is related to classification in the website of UCI Machine Learning Repository. It can be a basic dataset of classification for beginner, which is suitable for this study.

6 Conclusion and Discussion

Compared to the neural network, decision tree can deal with the non-digital data better, as the example shown above (**Figure 2**), while the neural network does well in distinguishing different features in a great deal of complex dataset, as the experiment shown above. Basically, both of them are not replaceable to each other. Besides, except other factor, if we just focus on the quantity of data. The decision tree can perform better in

the minor dataset and with the amount of data increasing, they both preform equally. But the superiority of neural network will gradually be shown, if the amount of data continues to increase. Because the demand of model capability is more important, while the risk of over fitting will reduce.

Compared Maximum Likelihood to neural network, as the statement shown in section 3.1, the sample provided for training or testing should be independent, or the accuracy of Maximum Likelihood method can be influenced. Thus, for selecting samples, neural network can have a wider scope. According to the report by Paola and Schowengerdt (1995), the data they used contain the mixed part of two vegetation classes, which possibly leaded to the lower accuracy of Maximum Likelihood method than the accuracy of neural network method. However, the large amount of time for iterations is needed for the neural network. The time for neural network to achieve the same accuracy of Maximum Likelihood in the same data is approximately the 15 times (Paola, & Schowengerdt, 1995).

As a beginner, I only use some different examples to introduce the features and theory of these three algorithms, and try to show the difference between Maximum Likelihood and neural network, as well as the difference between Decision Tree and neural network. For the future work, I am going to learn more to do the more practical comparison among these algorithms, like image classification about geography, and it can be the suitable dataset for these three type of algorithms for comparison.

References

Aeberhard, S. (1991). Wine Data Set. Retrieved from http://archive.ics.uci.edu/ml/datasets/Wine

Hameed, A. A., Karlik, B., & Salman, M. S. (2016). Back-propagation algorithm with variable adaptive momentum. *Knowledge-Based Systems, 114*, 79-87. doi:10.1016/j.knosys.2016.10.001

Hecht-Nielsen, R. (1987). Kolmogorov's mapping neural network existence theorem. In *Proceedings of the international conference on Neural Networks* (pp. 11-14). IEEE Press.

L.K. Milne, T.D. Gedeon & A.K. Skidmore (2018). Classifying dry sclerophyll froest from augmented satellite data: comparing neural network, decision tree and maximum likelihood. Retrieved from https://wattlecourses.anu.edu.au/mod/folder/view.php?id=1357298

Marshall, M. (1988). Iris Data Set. Retrieved from http://archive.ics.uci.edu/ml/datasets/Iris

Paola, J. D., & Schowengerdt, R. A. (1995). A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification. *IEEE Transactions on Geoscience and remote sensing*, *33*(4), 981-996.

Sheffer, J. (2014). Occam's razor. *Biomedical Instrumentation & Technology, 48*(2), 1.

Scott, A. J., & Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics, 27*(2), 387-397.

Yee, N. (2016). Principal component selection for neural network classification of active ingredients from near infrared spectra. *The Review of Socionetwork Strategies, 10*(2), 91-103. doi:10.1007/s12626-016-0066-7