Survey of Decision Tree and Logistic Regression

Miaomiao Zhao, Wenliang Hu, Zixun Zhang

Marquette University, WI

Abstract

The current era is an era of data explosion. Humans must seek more and more profits by analyzing large amounts of data and forecasts, such as bank marketing. However, due to the tedious and cumbersome processing and forecasting of large amounts of data, the staffs face great difficulties in this regard. Data mining work can greatly improve this dilemma. The methods of data mining work are various. This paper mainly introduces two kinds of algorithms, one is decision tree and the other is logistic regression. We will collect a large amount of relevant literature for comparison and then analyze the performance and differences between the two algorithms. In survey, we discuss two algorithms: Decision Tree and Logistic Regression, and make an overview of those two method. In the last part, we show the compare by some existed related projects)

Key words: *Decision tree, Logistic Regression, Data mining.*

I. Introduction

Nowadays, due to the rapid development of big data, the status of data mining technology has become more and more important in all walks of life, especially in the fields of e-commerce, medical care, and smart industry. The core content of data mining is its machine learning algorithm. There are many kinds of data mining algorithms available on the market. Commonly there are decision tree, logistic regression, k-means, k-nearest neighbor, etc.

Decision Tree [1] is a graphical method that uses probability analysis intuitively. It will determine the probability that the net present value will be greater than or equal to zero by constructing a decision tree based on the probability of occurrence of the various conditions. It can evaluate the risk of this project and then judge and analyze the feasibility of this project. In machine learning, a decision tree is a predictive model. He represents a mapping relationship between object attributes and object values. It uses a tree structure to establish a decision model based on the attributes of the data. Then, passing several layers of decision tree, we will get pure subset. Logistic regression [2] is a regression model where the dependent variable is classified. In the case of a binary dependent variable, its output can take only two values, namely "0" and "1", such as good/bad, right/wrong or healthy/sick. If the dependent variable has two or more results, its category can be analyzed and predicted in a logistic regression analysis.

II. Overviews

In this paper, we focus on two of the algorithms, one is the decision tree and the other is logistic regression.

Both Decision Tree and Regression (LR) are one of the classification and prediction algorithms. The probability of future outcomes is predicted by the performance of historical data. A deeply research is done in order to dig out all details of not only the decision tree but also logistic regression. Then we are going to compare two algorithms with advantages and disadvantages.

III. Methods

A. Decision Tree

Decision tree is a powerful prediction method which plays an important role in data mining. The reason why the decision tree algorithm so popular is that the final model is easy to understand by practitioners and domain experts. Decision trees are applied on both regression and classification problems.

The decision tree algorithm starts from attributes, which is as the basis of the algorithm, to separate different class. For example, in our bank marketing dataset, there are 16 attributes and one column of class. The values of class are "yes" and "no" so that this should be a binary decision tree. For binary decision tree, the decision tree algorithm separates the data into two subsets, and for each subset, we separate them into two subsets. We keep split the subsets until the criterion is satisfied.

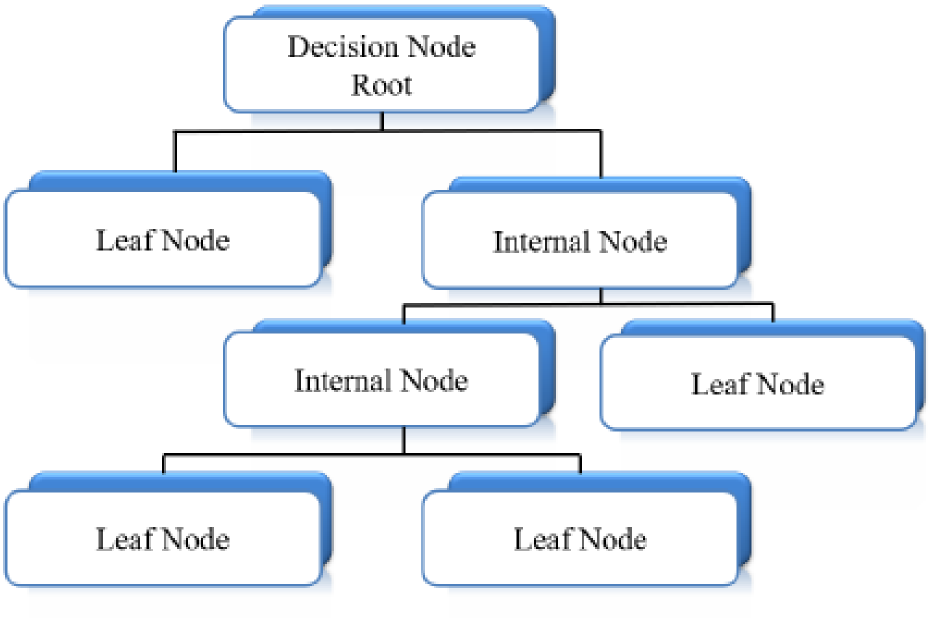


Fig.1 Decision tree model
(Image taken from our project)

Advantages of decision trees: [3][4]

- Decision trees are very easy to understand, explain, visualize.
- Since decision tree algorithm could handle both numeric and categorical data, decision tree model requires little data processing so that there is no need to do the

data normalization.

- Nonlinear relationships between attributes won't impact the tree performance.

Disadvantage of decision trees:

- Decision tree model leads to overfit noisy data.
- Decision tree is unstable since small variations give rise to a different tree which is generated. The variance should be decreased through the methods like bagging and boosting.
- It is easy to create a biased tree if some classes dominate.

Although there are several different type algorithms to build decision tree models, based on our project, we are going to focus on two of them: Classification and Regression Trees and Iterative Dichotomiser 3.

Iterative Dichotomiser 3 (ID3)

The ID3 algorithm was first proposed by Quinlan. This algorithm is based on information theory and uses information entropy and information gain as metrics to achieve data inductive classification. [5][6]

First of all, the problem that the ID3 algorithm needs to solve is how to select features as the criteria for dividing the data set. In the ID3 algorithm, the attribute with the greatest information gain is selected as the current feature to classify the data set. The concept of information gain will be described below, and the data set is continuously divided by continuous selection of features;

Second, the problem that the ID3 algorithm needs to solve is how to determine the end of the partition. There are two cases: The first one is the divided class belongs to the same class, the other is that there is no attribute to separate the data into subsets.

Finally, we get the final decision tree model.

Classification and Regression Trees (CART)

Regression: sum(y—prediction)^2

The cost function that minimizes the selection of the segmentation point is the sum of the squared errors of all the training samples that fall within the rectangle.

Classification: G = sum(pk * (1—pk))

Uses the Gini cost function, which provides a purity indication of the node, where node purity refers to how the training data assigned to each node is mixed.

The steps of doing the decision tree algorithm are: [7]

First, we need to compute the gini index for the dataset. Secondly, for every attribute, calculate gini index of all categorical values, than calculate the gini gain. Finally, we will get the tree we desired.

B. Logistic Regression

In most of survey or project, we could find Logistic Regression. For instance, in our group project where we try to predict if client will subscribe deposit in the next terms, this is a linking to our project https://github.com/huwenliang777/SpaceProject.git. In that project, we compare the Decision Tree and Logistic Regression Model and get the advantages and disadvantages of Decision Tree and Logistic Regression which will be mentioned in next section Compared with others. Therefore, we would like to show some concepts and knowledge points.

Logistic Regression is a kind of statistic method to analyze the datasets [8]. For the Logistic Regression model, the independent variables could be two or more, however, as we know, it is not more independent features to fit into model, the better accuracy that we will get. And for the prediction, the class or type would be binary “0” (false, fail, not) and “1” (true, success, yes) [9].

Logistic regression will estimate the coefficients and fit to the formula, it is a kind of linear function of partial or all dependent variables what we want to fit into logistic regression model [10].

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

So, this formula could be written as

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

P(x) is the probability of the dependent variable is success case rather than the false or fail, that could also say the probability is higher, the predicted result will be success case. The Fig.2 shows the standard logistic regression.

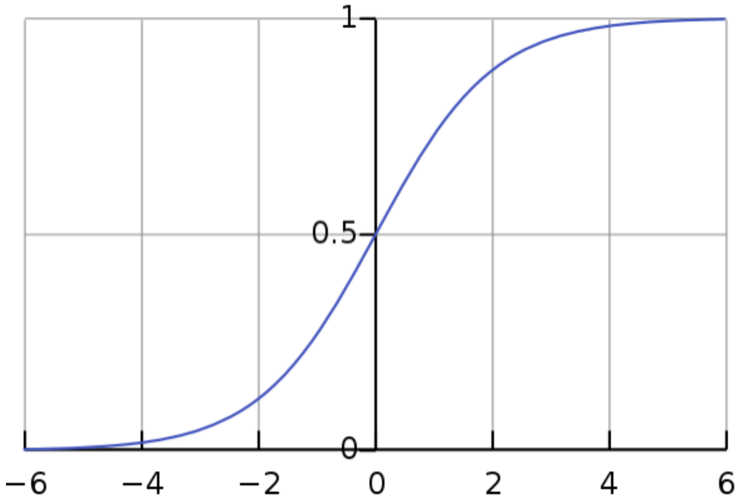


Fig.2 The Standard Logistic Regression

From this graph, we could find the y value will be between 0 and 1, that is not difficult to understand since we get the probability of success case. From this point, we will discuss why do we want to choose logistic regression rather than linear regression? Jonathan Bartlett [11] told us why linear regression is not a perfect idea when the predicted variable is a binary variable by considering linear regression assumption. Paul Von Hippel [12] discussed this by different cases in paper: Linear vs. Logistic Probability Models: Which is Better, and When? First of all, the result of linear regression could be bigger than 1, however, the probability of logistic regression is between 0 and 1. By this way, we could be easy to check and analyze the result.

In the process of building the logistic regression model, the important step should be estimating the parameters of formula of logistic regression. Here, we would like to discuss the Maximum Likelihood Estimation which is a popular method to estimate the coefficients.

The MLE is derived from the probability distribution of the dependent variable [13].

$$f(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i - y_i)!} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

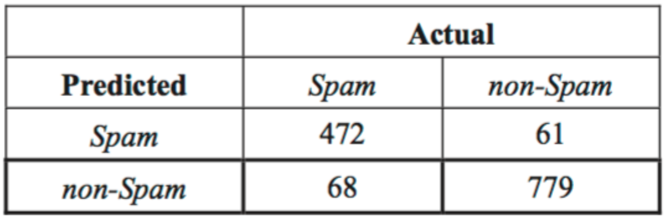
In this paper, Scott A. Caepiel* shows how to use likelihood method to estimate the parameters of logistic regression by deriving from the probability distribution of the dependent variables. There are some other methods to estimate [14], the familiar method could be Least Square which is similar with MLE. Another one is robust methods that try to balance the efficiency and desirable properties of least squares and maximum likelihood with a lower sensitivity to outliers.

IV. Compare

In this section, we try to compare Decision Tree and Logistic Regression Algorithms by some projects and experiments. Here, we would like to talk our project "Bank Deposit", in this project we build two models to predict if client will subscribe deposit in next term. The result is the accuracy of Decision Tree is 0.81, however Logistic Regression is up to 0.90. By our project, we find the accuracy of Logistic Regression is higher, and it is easy to build model. However, in fact, it depends on the datasets, and how to build model (method and algorithm).

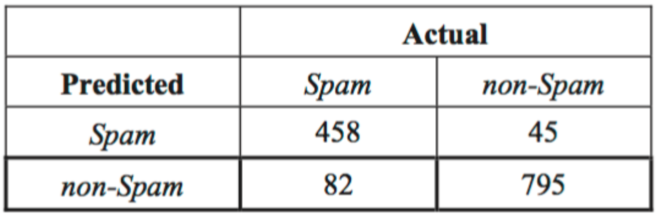
For instance, Adi Wijaya and Achmad Bisri [15] classifier for Email spam detection using decision tree and logistic regression, Tabel1 shows the Confusion Matrix of Decision Tree and Table2 shows the Confusion Matrix of Logistic Regression.

Tabel1 Confusion Matrix of Decision Tree



	Actual	
Predicted	*Spam*	*non-Spam*
Spam	472	61
non-Spam	68	779

Tabel2 Confusion Matrix of Logistic Regression

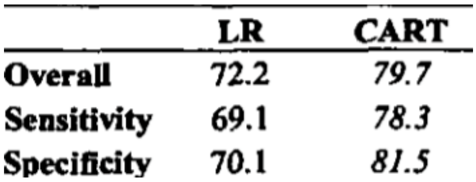


	Actual	
Predicted	*Spam*	*non-Spam*
Spam	458	45
non-Spam	82	795

By calculating accuracy, they got accuracy of decision tree is 90.65% and accuracy of decision tree is 90.79%. Those two almost are same.

However, Donald Brown, Justin R. Stile and other two analyze Accident data by using those methods, get the result like Table3 shows. F

Tabel3 Result



	LR	CART
Overall	72.2	*79.7*
Sensitivity	69.1	*78.3*
Specificity	70.1	*81.5*

For these two methods, they have their own advantages and disadvantages. In fact, when we would like to build model to analyze dataset and predict, choosing which method depends on what we need.

For example, decision tree is trending to consider consequences, which means we could use the framework to follow the probability. In the same time, we could use backward method to get the condition when we know the consequences [16].

Considering Logistic Regression, the first advantage is we do not assume the linear relation between independent variables and dependent variable [17], in the same time, that is why we choose logistic regression rather than linear regression. The second point is dependent variable need not to be normally distribution.

All in all, there are a lot of factors we need to consider when we choose a model. Or combining those two methods is a good way to improve our model and improve the accuracy. Building a model is just a start, the important and though work is to know how to improve this model.

V. Conclusion

In this survey, we deeply studied two popular data mining techniques: Decision tree and Logistic Regressive. What's more, we compared both algorithms by some projects and experiments. Although the results of our project show an obvious difference between decision tree model and logistic regression model, however, in the experiments for Email spam detection using decision tree and logistic regression shows very close values of accuracy and confusion matrix. We need to add more meaningful attributes to decision tree model in our project so that the accuracy might increase and get close to the logistic regression model.

The survey also shows the experiment which the accuracy of logistic regression is lower than the decision tree, which strongly proved that, besides the algorithm we use, how we improve the model is the most important thing in data mining.

VI. Reference

[1] The mind tool content team. Decision Tree Analysis. Retrieved https://www.mindtools.com/dectree.html?route=article/newTED_04.htm

[2] M. Strano; B.M. Colosimo (2006)."Logistic regression analysis for experimental determination of forming limit diagrams". Retrieved

https://www.sciencedirect.com/science/article/pii/S0890695505001598?via%3Dihub

[3] Rebecca Njeri. What Is A Decision Tree Algorithm? Retrieved https://medium.com/@SeattleDataGuy/what-is-a-decision-tree-algorithm-4531749d2a17

[4] Prashant Gupta. Decision Trees in Machine Learning. Retrieved https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

[5] Decision Tree- Classification. Retrieved http://www.saedsayad.com/decision_tree.htm

[6] Simply study machine algorithm: Decision Tree ID3 algorithm. Retrieved https://blog.csdn.net/google19890102/article/details/28611225

[7] Jason Brownlee. How To Implement The Decision Tree Algorithm From Scratch In Python. Retrieved https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/

[8] Jason Brownlee, "Logistic Regression for Machine Learning", April 1, 2016. Retrieved https://machinelearningmastery.com/logistic-regression-for-machine-learning/

[9] Wikipedia, "Logistic Regression". Retrieved https://en.wikipedia.org/wiki/Logistic_regression

[10] MEDCALC easy-to-use statistical software, "Logistic Regression". Retrieved https://www.medcalc.org/manual/logistic_regression.php

[11] Jonathan Bartlett, The Stats Geek, "Why shouldn't I use linear regression if my outcome is binary?", January 17, 2015. Retrieved https://www.medcalc.org/manual/logistic_regression.php

[12] Paul Von Hippel, "Linear vs. Logistic Probability Models: Which is Better, and When?", July 5, 2015. Retrieved https://statisticalhorizons.com/linear-vs-logistic

[13] Scott A. Caepiel*, "Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation". Retrieved https://czep.net/stat/mlelr.pdf

[14] "How are estimates of the unknown parameters obtained?". Retrieved https://www.itl.nist.gov/div898/handbook/pmd/section4/pmd43.htm

[15] Adi Wijaya and Achmad Bisri, "Hybrid Decision Tree and Logistic Regression Classifier for Email Spam Detection". Retrieved https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7863267&tag=1

[16] Gregory Hamel, "Advantages & Disadvantages of Decision Trees". Retrieved https://www.techwalla.com/articles/advantages-disadvantages-of-decision-trees

[17] Fang, "Advantages & Disadvantages of Logistic Regression". Retrieved https://victorfang.wordpress.com/2011/05/10/advantages-and-disadvantages-of-logistic-regression/