HYBRIDIZATION OF DECISION TREES AND NEURAL NETWORKS

Utku Şirin, Özge Tokgöz
usirin@ceng.metu.edu.tr, zehra.tokgoz@ceng.metu.edu.tr
METU Computer Engineering Dept., 06531, Ankara, TURKEY

INTRODUCTION

Classification task is getting more and more attention throughout the recent years. Many scientific problems may be interpreted as a classification problem and can be solved by using very well developed and enhanced classification, machine learning algorithms. Although there are various types of classification algorithms for solving those tasks, the improvement of those algorithms is highly subjective, i.e., domain specific. In this work, we have tried to build a new general framework that is combining two different well-known methods for improving the success rate of methods. Our framework is composed of decision trees combined with neural networks. We have also collected several experiments with the limited experimental data served for KDDCUP ‘12 Track2.

Keywords: machine learning, neural networks, decision trees, hybrid algorithms, classifcation.

MODELS AND ALGORITHMS

Decision Trees

Decision tree as a predictive model maps observations about an item to conclusions about the item's target value. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

There are many specific decision-tree algorithms. In our project, we have used ID3 algorithm to select attributes and split nodes. ID3 algorithm makes use of information entropy as a measure of the average information content. It takes all unused attributes, counts their entropies and calculates **information gains** on attributes. The attribute which leads to more information gain (equivalently minimum entropy) is chosen as node attribute.

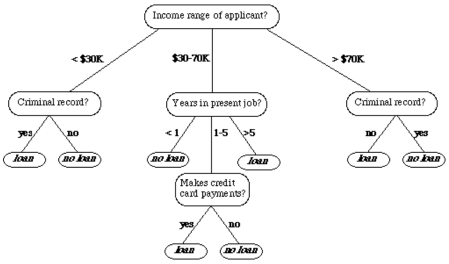


Figure 1: Sample Decision Tree

Neural Networks

Neural network (NN) is a mathematical or computational model that is inspired by the structure and functional aspects of biological neural networks. A neural network consists of a set of connected input/output units (perceptrons). Each connection has a weight associated with it. The NN learns by adjusting weights of the connections in the network to produce a desired output. After learning phase, the activation function converts a perceptron's weighted input to its output.

In our project, we have used single-layered perceptrons instead of multi-layer ones. For click and impression values which will be predicted, two different perceptron models are applied.

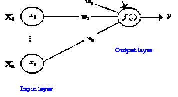


Figure 2: Sample Single-Layer Neural Networks (Perceptrons)

Hybrid Model

Hybrid Model is the model that we have combined decision trees and neural networks. The working principle of the hybrid model is as following. When the training data is available, firstly, one has to choose a subset of attributes that the data has. Those attributes should be the ones that are suitable for classifying with decision trees. For example, while UserID may not be a very suitable attribute for decision three classification, gender may fit perfectly for decision trees. After deciding the subset of attributes, the data is reduced to have only those attributes. Then the decision tree is built over those attributes and the ***first-level classifier*** is ready to use.

As we have only used a subset of the attributes that the data has, decision tree will classify the data only ***partially***. This is where the neural networks are integrated with the decision trees. For each training data, we firstly classify it with respect to the decision tree, if there are more than one possible results, the reduced form of the training data is used for training the neural networks. ***Hence, we have built a neural network model on the leafs of the decision trees.*** While testing the data, the same procedure is applied to each test data, either. If decision tree is able to produce single result, it is accepted, if it gives more than one results, the weights of the neural network that are trained, are loaded and the ultimate predicted result is reported.

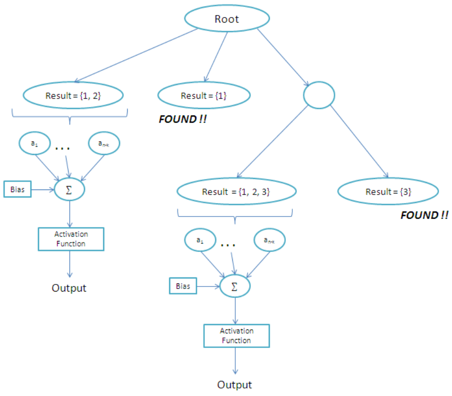


Figure 3: The Hybrid Model

EXPERIMENTS

Our experiments are based on the data provided in KDDCUP ’12 Track 2. The task is to predict the click thorough rates of the online advertisements. In order to calculate the click through rates, one has to divide the number of times that the ad is clicked to the number of times that the ad is displayed/impressed to the user. To calculate this rate, we have to predict two attributes, click and impression.

Data

Training data file from KDDCUP’12 Track 2 contains 150 Million records, and each record contains 12 attributes: click, impression, displayURL, adID, advertiserID, depth, position, queryID, keywordID, titleID, descriptionID, userID. ID attributes also have seperate files containing ID values as primary keys and hash values correponding to these IDs. Only userID reference file contains categorical variables corresponding to userID (gender and age) instead of hash values. Gender has two values and age values are categorized into six.

Test data file is smaller than training file. It contains 11 attributes except click and impression values to be predicted.

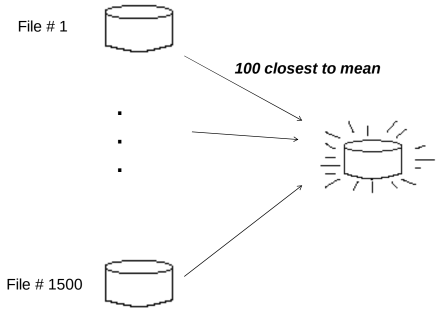
For building the decision tree we have chosen the four attributes, namely, depth, position, gender, and age. For applying neural network to the required leaf of the decision tree, we have decided to use a single-layer perceptron**.**

Sampling

Since training file is too big, we have divided it into about 1500 small files. These files are still to big to manipulate. We have sampled each ~1500 file by collecting 100 records from each file. For this purpose, firstly we have calculated mean values for each attribute in the file. Hash values’ means have been also considered for each attribute. Then, 100 records that are closest to mean values have been selected.

This sampling process time is calculated as about 5 months due to reference files’ size and computation limits. For the time being, we have collected 4000 records (from 40 files) and worked on that small training file.

We have taken some sample training records for reference and testing purpose.



Graphical User Interface

We have also designed a simple GUI. It was difficult to handle file paths on each team member’s own environment. Therefore, designed GUI gets file paths (training, test and reference files). Also, models are selected and can be trained/tested using GUI. Also, decision trees of click and impression for hybrid model and success results on bar chart improve visuality.

EVALUATION

We have applied our model to the explained data with the described model settings. We have compared the our results with only single-layer perceptron algorithms. Results are measured based on their accuracy values and are shown in Figure 4. Results show that our hybrid model and single layer perceptron have given same results. This we believe that because of the very high results and the domain-specific characteristics of the data – click is most of the time is 0 and the impression is most of the time is 1.

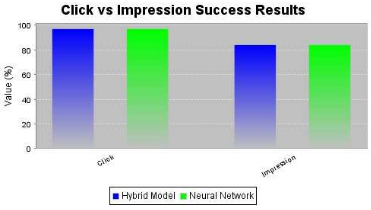


Figure 4: Comparative Results of Click and Impression

CONCLUSION & FUTURE WORK

In this work, we have built a new classification framework that is combining both decision trees and neural networks. The experimental results showed that our system and the single layer neural network system produced same results. As future work, we are planning to use another domain so that we can get more accurate results for comparing our system and the neural networks. Moreover, we are planning to enhance the neural network structure to be able to see its ability with decision trees much more effectively.