RESEARCH ARTICLE OPEN ACCESS

Prediction of the Data Analysis Using Decision Tree

P.T.V.Lakshmi, B.Nivegeetha
Assistant Professor
NadarSaraswathi College of Arts and Science,Theni
India

ABSTRACT
Data mining is the process of discovering the patterns in large data set`s at the intersection of machine learning, statistics and database systems. Data mining is the analysis step of KDD. In this paper we are going to analyze the Iris data sets for sample and going to construct the decision tree based on their individual characteristics. The classification technique is used to accurately predict the target class for each case in the data. In this paper we are going to analyze the datasets using the powerful tool known as "Rapid miner".

Keywords:- *Data mining, classification, decision tree, rapid miner.*

I. INTRODUCTION

Data mining is a computer science term also known as Knowledge Discovery in Databases (KDD). It was about finding new information in a lot of data. The data is stored so that it can be used later. Data mining is primarily used by industries like retail, financial and marketing companies. If you have ever shopped at a retail store and received customized coupons that's the result of mining (i.e.) your individual purchase history was analyzed to find out what products you have been buying and what promotions you are likely to be interested in. By analyzing the datasets the decision tree was built based on the classification techniques using the "Rapid miner tool". Through this paper we are going to construct the decision tree for Iris datasets.

RAPID MINER TOOL

Rapid miner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation , machine learning, deep learning ,text mining and predictive analytics. It is used mainly for business and commercial applications. It is developed on a open core model. Its initial release was before 13 years. It was formally known as YALE (Yet Another Learning Environment). It uses the client/server model with the server offered as either on-premise, or in public or private cloud infrastructure. It is written in java programming language and it provides GUI to design and execute the analytical workflows. Those workflows are called "processes" and it consists of multiple operators. It provides learning schemes, models and algorithms using python scripts. About 50 developers participate in the development of the open source rapid miner.

II. CLASSIFICATION

Classification comes under the category of supervised learning (i.e.) the training data are accompanied by labels indicating the class of the observations. It is a data mining function that assigns items in a collection to target category. The main goal is to accurately predict the target class. The simplest type of classification problem is binary classification. Classification models are tested by comparing the predicted values to known target values in a set of test data. They are divided into two data sets

1. Building the model
2. Testing the model

III. CLASSIFICATION ALGORITHMS

- Linear classifiers
- Support vector machines
- Quadratic classifiers

- Kernel estimation
- Decision trees
- Neural networks
- Learning vector quantization

Among the many algorithms, in our view of perspective the best algorithm chosen is " Decision tree".

IV. DECISION TREE ALGORITHM

A decision tree is a structure that includes a root node, branches and a leaf node. Each internal node denotes a test on a attribute, each branch denotes the outcome of a test and each leaf node holds the class label. The topmost node in the tree is the root node. They are used for solving regression and classification problems.

- Place the best attribute of the dataset at the root of the tree.
- Split the training data set into subsets.
- Repeat steps 1 , 2 on each subset until you find leaf nodes in all the branches of a tree.

ABOUT IRIS

BOTANICAL NAME	Iris germanica
PLANT TYPE	Flower
SUN EXPOSURE	Full Sun, Part Sun
BLOOM TIME	Summer
FLOWER COLOR	Blue, Multicolor, Orange, Pink, Purple, White, Yellow

Table I

IRIS DECISION TREE USING RAPID MINER TOOL

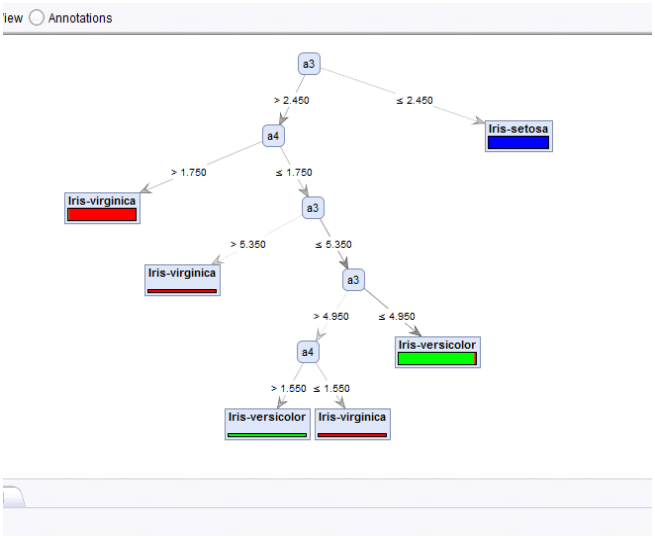


Fig 1

The attributes a1, a2, a3 and a4 are sepal length, sepal width, petal length and petal width respectively. They form the decision tree according their individual characteristics. Iris family is classified into three types namely,

- Iris-Setosa
- Iris-Versicolor
- Iris-Virginica

ALGORITHM

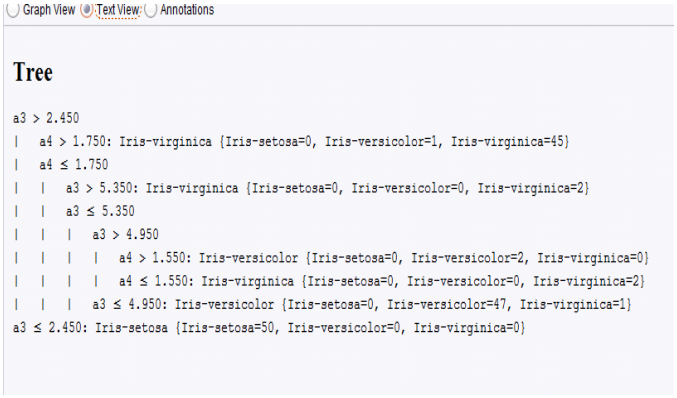
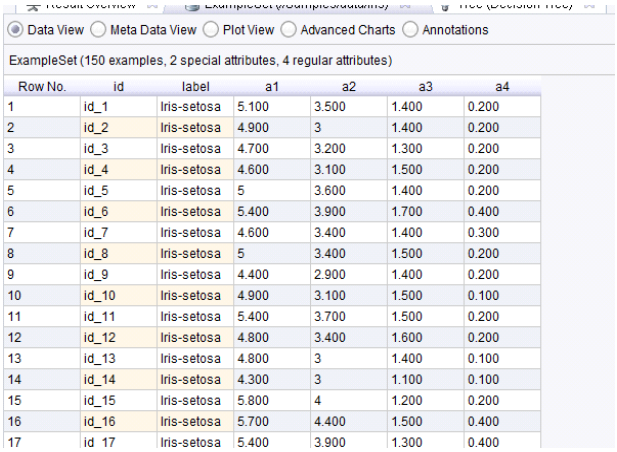


Fig 2

The algorithm used here is the "Decision tree algorithm".

DATA SET



Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (150 examples, 2 special attributes, 4 regular attributes)

Row No.	id	label	a1	a2	a3	a4
1	id_1	Iris-setosa	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	4.900	3	1.400	0.200
3	id_3	Iris-setosa	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	5	3.600	1.400	0.200
6	id_6	Iris-setosa	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	5	3.400	1.500	0.200
9	id_9	Iris-setosa	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	4.900	3.100	1.500	0.100
11	id_11	Iris-setosa	5.400	3.700	1.500	0.200
12	id_12	Iris-setosa	4.800	3.400	1.600	0.200
13	id_13	Iris-setosa	4.800	3	1.400	0.100
14	id_14	Iris-setosa	4.300	3	1.100	0.100
15	id_15	Iris-setosa	5.800	4	1.200	0.200
16	id_16	Iris-setosa	5.700	4.400	1.500	0.400
17	id_17	Iris-setosa	5.400	3.900	1.300	0.400

Fig 3

This shows the dataset for Iris.

V. CONCLUSION

It produces the accurate result for prediction of the datasets. They enable us to understand easily. Rapid miner tool provides the best pathway to mine the data in the effective manner. We analyzed the data for Iris with more accuracy. By analyzing the algorithm we found that decision tree produces the best results when compared to linear classifiers.

REFERENCES

[1] Data Mining: Concepts and Techniques, Third Edition by Han, Kamber & Pei (2013).

[2] Data Mining and Analysis Fundamental Concepts and Algorithms by Zaki & Meira (2014).

[3] Data Mining Techniques: For Marketing, Sales, and Customer Support, Michael J. A. Berry, Gordon S. Linoff, Wiley, 1997.

[3] Yahoo Data Mining Club.

[4] UCI HYPERLINK "http://www.ics.uci.edu/~mlearn/MLRepository.html" "http://www.ics.uci.edu/~mlearn/MLRepository.html"Machine "http://www.ics.uci.edu/~mlearn/MLRepository.html" "http://www.ics.uci.edu/~mlearn/MLRepository.html"Learning "http://www.ics.uci.edu/~mlearn/MLRepository.html" "http://www.ics.uci.edu/~mlearn/MLRepository.html"Repository, databases, domain theories and data generators for the empirical analysis of machine learning algorithms.