Proposal

Objective

To build a Machine Learning model based on classification using J48 algorithm on the dataset 'Iris' taken from the most famous online repository UCI.

Dataset

The dataset named 'Iris' is taken from UCI Repository of datasets. It can be accessed at -

https://archive.ics.uci.edu/ml/datasets/iris.

Dataset Description

Iris dataset is a plant database. It is a classification database that classifies iris plants into 3 different categories based on 4 parameters. So, it allows us to identify the class of an iris plant if we have information about 4 parameters. It is one of most famous and best datasets to study classification with the most number of web hits.

It has 5 attributes namely:

1) Sepal length- It gives the length of the sepal of any category in cm.
2) Sepal width- It gives the width of the sepal of any category in cm.
3) Petal Length- It gives the length of the petal of any category in cm.
4) Petal Width- It gives the width of the petal of any category in cm.
5) Class: There are 3 types of class of plant- Iris Setosa, Iris Versicolor, Iris Virginica.

   So, predicted attribute is 'class' of iris plant.

The dataset has 150 instances in total.

Pre-processing

The dataset has no missing values hence not much data cleaning and pre-processing is required. The dataset will be imported in Python environment and then further modeling and analysis will be done.

Methodology

The algorithm used to build the model is decision tree by **J48** algorithm. It will study the classification of iris plant in any one of the class based on other 4 attributes. So, it will be very easy to predict the class of any new iris plant by looking at the value of these 4 parameters.

All the algorithms are run in Python notebook. The libraries used are ‘sklearn’ and to visualize the tree, ‘graphviz’ is used.

It can be asked that why this model is beneficial - to understand classification in the best and easiest way. Our brain perceives things faster if they can be visualized so firstly, the output of this model is in such a way that makes it easy to understand. Secondly, the way it classifies, it considers individual value of all the parameters. Thirdly, comparison is always better than calculation to get a concrete result. So, to understand and predict results, this is the best and the easiest algorithm.

Results

J48 is a tree-based classification algorithm. The output will be a tree that can be visualized as well as probabilities for lying in specific category based on the values of attributes. Thus, classification can be studied.

This dataset and the model will be very useful to under classification techniques and how to make good observations.

Explanation of the result obtained

The decision tree obtained is called so because it has the structure of a tree. Like a tree, it has branches and these branches can be traced down based on the value of the parameters that we have. This classification is done by using 'InfoGain' technique which gains information about all the instances and uses it for classification. For example: If sepal length> 'a' and sepal width > 'b' then we will follow one branch and it will have another condition leading us to a class at last. So, tracing down to a branch and comparing the values of a parameter can help us go down to a class it belongs to.

GitHub link for the project

All the codes will be available on the GitHub platform with the following link:

https://github.com/hadeel-atheer/Machine-Learning

Weekly schedule and milestones