SABER PRO SCORE. USAGE OF DATA STRUCTURES THAT WILL TELL IF YOUR PAST COULD DEFINE YOUR FUTURE

Manuela Zapata Mesa
Universidad Eafit
Colombia
mzapatam1@eafit.edu.co

María Alejandra Moncada Agudelo
Universidad Eafit
Colombia
mamoncadaa@eafit.edu.co

Mauricio Toro
Universidad Eafit
Colombia
mtorobe@eafit.edu.co

ABSTRACT

The objective of this study is creating an algorithm based on decision trees, which predicts the Saber Pro test score, based on different social, economic and academic variables.

Finding the solution to this problem would be very useful for universities as they could change their approach in order to improve their average score and by using the information provided by the algorithm the adjustments will be much easier and efficient.

The four similar problems used in the report are algorithms frequently used in artificial intelligence with a very similar function among them.

Keywords

Saber pro test, decision trees, algorithm, analysis.

ACM CLASSIFICATION Keywords

Computing methodologies → Machine learning → Machine Learning algorithms → Dynamic programming for Markov decision processes → Value iteration

1. INTRODUCTION

Currently, the world's universities have an amount of information about their students never seen before, this information in most cases is not properly used by the universities.

With the data that is counted, there are endless things that could be done, which might be very useful for a university and even for a country; one of the most important things is the prediction of success in academic tests, information that would allow institutions to greatly improve its results and helps its students; Therefore, it is pertinent that each of the holders of this information have a system that can take the data they have and return predictions based on that material.

2. PROBLEM

The problem is based on crating an algorithm that uses the decision trees to predict the students score in the Saber Pro tests, for this there is a series of social, academic and economic variables, which will be used to perform the prediction.

Solving this problem would be very convenient for universities that wish to improve their Saber Pro scores since they would know which students should receive more attention and help in order to improve their tests results in the future.

3. RELATED WORK

3.1 ID3 algorithm

Iterative Dichotomiser 3 is an algorithm invented by Ross Quinlan that is typically used in the machine learning and natural language processing domains.

This algorithm is used in artificial intelligence. This algorithm works based in a set of examples, each example has attributes; the attribute is a binary objective, from these examples the algorithm tries to obtain a hypothesis that classified new instances.ID3 does that constructing a decision tree.

This algorithm does not guarantee an optimal solution and is harder to use on continuous data that on factored data.

3.2 C4.5 algorithm

C4.5 is an algorithm invented by Ross Quilan, this algorithm is an extension of the ID3 algorithm. The decision trees created by C4.5 can be used for classification.

C4,5 construct the tree in the same way that ID3 does, using examples; C4.5 choose an attribute that divides the example set in subsets.

A few of the things that C4.5 has and ID3 does not:

-manage of continuous and factored data.

-manage of the example set with missing values.

-manage of attributes with different cost.

-delete the parts of the tree that does not help

3.3 AQ algorithm

The algorithm quasi-optimal (AQ) is a powerful machine learning methodology aimed at learning symbolic decision rules from a set of examples and counterexamples. It has been applied to solve several problems from different domains, including the generation of individuals within an evolutionary computation framework. The current article introduces the main concepts of the AQ methodology and describes AQ for source detection (AQ4SD), a tailored implementation of the AQ methodology to solve the problem of finding the sources of atmospheric releases using distributed sensor measurements. The AQ4SD program is tested to find the sources of all the releases of the prairie grass field experiment. [1]

Like the ID3 algorithm CN2 need and example set, usually call training set, in order to generate a list of classification rules.

4. Matrix

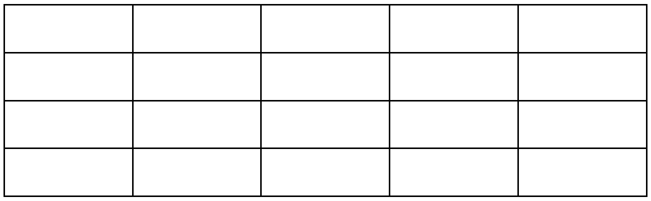


Figure 1: Sparce matrix n*m, where the number in each line is an individual and each column is the information we have about them.

4.1 Read_Data set

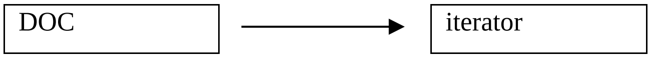
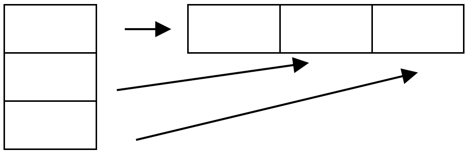


Figure 2: The first thing method read data set does is open the document with Open() where it converts the document to cvs.reader that converts the document into an iterator that is delimited by ;



REFERENCES

1. http://geoinf.psu.edu/publications/2010_WIRES_

   AQLearning_Cervone.pdf

2. https://people.eecs.ku.edu/~jerzygb/j24-sel.pdf 3.

   Anon, 2020. C4.5 algorithm. *En.wikipedia.org*.

4. Anon, 2020. Algoritmo ID3. *Es.wikipedia.org*.

3.4 CN2 algorithm

The CN2 algorithm is a learning algorithm for rule induction, its design based on ID3 algorithm an AQ algorithm a can even work when the example set is imperfect.

Figure 3: Then in a for, you take each line of the iterator, that converts into an array when it goes through the for. And it fills again a new arrangement called data, which will be an array of arrays, therefore it will be a matrix.

4.2 Design criteria of the data structure

We decided to choose a matrix as the structure to store our data, due to it´s low complexity to get access to the data (O(1)), which we think can reduce significantly the complexity of the entire program, because having access to the data will be one of the most used functions in our algorithm.

4.3 Complexity analysis

Method	Complexity
Read_Dataset	O(n*m)

4.4 Execution time

3.5	**3.6 5000**	**3.7 5000**	**3.8 5000**	**3.9 5000**	**3.10 35000**
3.11 Read_Dataset	**3.12 0.0838s**	**3.13 .2453s**	**3.14 .8513s**	**3.15 .2659s**	**3.16 .7016s**

Table 2: Execution time of the operations of the data