Tec de Monterrey

Proyecto de sistemas inteligentes

Feature selection

Alejandro Cedillo Gámez

08/11/2022

Contents

[Feature selection 3](#_Toc120690506)

[Abstract: 3](#_Toc120690507)

[Introduction 3](#_Toc120690508)

[Background 3](#_Toc120690509)

[(1+1) 3](#_Toc120690510)

[Genetic Algorithm 3](#_Toc120690511)

[Methodology 4](#_Toc120690512)

[Palmer Penguins 4](#_Toc120690513)

[Mushrooms 5](#_Toc120690514)

[Internet advertisements 5](#_Toc120690515)

[Results 6](#_Toc120690516)

# Feature selection

## Abstract:

## In many classification models, one of the most frequently faced problems is the speed and performance of the model, as it has large volumes of information to handle. For this reason, it is sought to minimize the characteristics within a dataset while maintaining a high level of precision. This with the help of genetic and evolutionary algorithms, where its performance will be seen in different sizes of sample data and comparing its effectiveness before and after the reduction.

## Introduction

The management of data is a relevant topic in today’s world, it rules our day to day by being used for most of our digital apps, but sometimes we have more than we can comprehend leaving us with a waste or burden of data, with these two problems need to be address, the storage of massive volume of information and the manipulations and understanding of data.

That’s why the feature selection approach was investigates, to reach a minimum amount of information to make this data decisions, were we eliminate all the information that does not apport to the problem that the data scientist is addressing and the methodology to prove so started with different algorithms to reduce the vector of features, this being 1+1 Evolutionary algorithm [1] and a Genetic algorithm [2], being apply in 3 different datasets with unique singularities, sizes and type of data.

## Background

The idea started by assuming that some information in a dataset can be ignore or doesn’t provide that much relevant information, so in order to represent these columns we use a vector with values of 0 and 1 representing if the information is used or not believing that there is a ideal vector with the minimal amount of 1’s.

With this description the concept of genetic algorithms was the approach that fitted the most for this maximization/minimization problem.

### (1+1)

The concept is very simple, 1+1 algorithms work only with one parent and one child,

1. Start by a creating random vector V of size S, this being our parent
2. Convert V to V’ by doing a bitwise mutation with a probability P. This being the child
3. Make an evaluation F to V and V1
4. If F(V) < F(V’) then V’ will be our new parent
5. Repeat until meet a stopping condition.

### Genetic Algorithm

Similar to the concept of evolution the GA follow the same concept of evolution through generations where two parents give their best features to the son. For this performance the use of some concepts needs to be explained first.

**Creation of individual**

Make a random vector of 0 and 1 by a given length

**Evaluation**

This being the heuristic that changes depending on the problem you are trying to address, in this case is a minimization of columns with a maximization of precision in our decision tree classifier returning a value between 1 and 0.

**Select**

Works like a selection tournament where you return the winner between n contestants, having n-1 rivals and 1 winner, for this case the size worked on was 3.

**Combine**

From 2 selected winners you create their sun or offspring by combining their information from a random pivot.

**Mutate**

Through the length of the vector, the bit wise operation is performed with a probability p.

To perform the GA you select the amount of generations to evolve and their population, at the beginning the vector would have low values and unclear demonstration of value, but after some generation the best individual obtain will be a very precise model and so it’ll behave like it has evolve.

## Methodology

3 different datasets were selected to work on a classification model with the use of a decision tree classifier. These were selected by the information provided and the actions that needed to be performed to the data frames so the relevance of data could be assured as well as their manageability for the ML part.

### Palmer Penguins

This is a data frame with the information of three different species of penguins from Antarctica in the Palmer Archipelago [3]. The classification variable being the species of the penguin, having Adelie, Chinstrap and Gentoo. Working with a mix of categorical and numerical values implied doing some data science to clean the data and make it easy to work on it. By being a small recollection of data and having this more like a test to understand the feature selection value rather than a big demonstration of value, the 1+1 algorithm was the right one, especially by being a small dataset.

A correlation matrix was used taking advantage of the number of features this allowing us to have a visual understanding of how the features would affect our model and having an idea on witch variables would more important to complete the task;

A picture containing graphical user interface

Description automatically generated

Figure . Correlation matrix for palmer penguins’ data.

For this dataset the instances were following a normal distribution, so to check outside values we search for those outside 3 times the standard deviation plus the mean.

### Mushrooms

In here we grow on volume, having 23 different species of mushrooms and gathering 22 attributes and 8124 instances, working on description values such as odor, gill size, cap shape, and more [4]. All of these being categorical attributes, so decoding was needed. The data was almost even distributed allowing to preform without problem the division for test and train.

Chart, pie chart

Description automatically generated

Figure . Mushrooms distribution of values.

For our model it must classify if a instance is poisonous or edible, this was created with a decision tree classifier and using a genetic algorithm with a selection tournament.

### Internet advertisements

For the final set the amount of data management had a big augment, having 1558 features and 3279 instances [5], in here the task is to determine if an instance is an advertisement or not, for this information of the element is provided, width, height, and more encoded information of the image. By being our largest dataset, it is also the one with the most missing values, so the cleaning and selection had to be performed.

Another issue found working with this was the way the data is distributed, having more than 3 times instances classified as *not an add* than *add*.

Chart, pie chart

Description automatically generated

Figure . Distribution of Advertisement results.

To continue working with this data and ensure the precision of the model, the under-sampling technique was used leaving us with a more even relation. Where 1 means not an ad and 0 being an ad.

Chart, pie chart

Description automatically generated

Figure . Distribution of Advertisement results after under-sampling.

## Results

Palmer Penguins had a behavior more than acceptable with 30 executions of 100 generations each and only using 8 individuals. Always ending up with values above 90%, in part this thanks to the nature of the size of the dataset.

After being executed 30 times with 100 generations we get the following repeated values, having as the most used feature body mass being included 2/3 of the executions. And despite what one might think the year was not the least used feature, it was the bill length with approximately a 150 difference of occurrence.

Chart, line chart

Description automatically generated

Figure . Accuracy score of Palmer penguins in 30 executions.

The mushrooms classification model throws almost always an evaluation of 1.0, where 1 individual execution looked like the figure 6, giving us a hint of how important a specific variable is. In these 30 executions we kept a register of the most used variables throughout all and end up having *veil-type* as the most used variable.

Chart, histogram

Description automatically generated

Figure . Behavior of GA in Mushrooms dataset.

For internet advertisement by being the largest dataset worked on, it gave us more data to analyze, having in 30 execution the behavior shown in Figure 7.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Chart, line chart, histogram  Description automatically generated  Figure . Accuracy scores obtain in 30  executions for Internet Advertisement. | |  |  | | --- | --- | | **count** | 30.000000 | | **mean** | 0.975983 | | **std** | 0.004551 | | **min** | 0.969432 | | **25%** | 0.973799 | | **50%** | 0.975983 | | **75%** | 0.978166 | | **max** | 0.982533 |   Table . Statistical values for Internet Advertisement in 30 executions. |

# References

[1] Droste, S., Jansen, T. and Wegener, I. (2002) *On the analysis of the (1+1) evolutionary algorithm*, *Theoretical Computer Science*. Elsevier. Available at: https://www.sciencedirect.com/science/article/pii/S0304397501001827 (Accessed: November 30, 2022).

[2] Eiben, A.E. and Smith, J.E. (2003) *Introduction to Evolutionary Computing*. 2nd edn. Berlin, Heidelberg: Springer (Natural Computing Series).

[3] Gorman KB, Williams TD, Fraser WR (2014) *Ecological Sexual Dimorphism and*

*Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis)*. PLoS ONE: e90081. <https://pypi.org/project/palmerpenguins/>

[4] Mushroom. (1987). UCI Machine Learning Repository.

<https://archive-beta.ics.uci.edu/dataset/73/mushroom>

[5] Kushmerick,Nicholas. (1998). Internet Advertisements. UCI Machine Learning Repository.

<https://archive-beta.ics.uci.edu/dataset/51/internet+advertisements>