Algorithm primitives

Master Thesis

L.D. Stooker, 0819041

April 4, 2018

Abstract

To improve existing automated picking of a machine learning algorithm

Contents

1 Introduction **3**
1.1 problem description . . . 3
1.2 research question . . . 3
1.3 thesis structure . . . 3
1.4 Outline . . . 3

2 Preliminaries **4**
2.1 Sklearn/scikit-learn library . . . 4
2.2 Terminology . . . 4

3 Experimental setup **5**
3.1 Motivation . . . 5
3.2 Description . . . 6
3.2.1 Main method . . . 6
3.2.2 Strategy 1 . . . 6
3.2.3 Strategy 2 . . . 6
3.2.4 Strategy 3 . . . 6
3.2.5 Strategy 4 . . . 6
3.2.6 Strategy 5 . . . 6
3.2.7 Strategy 6 . . . 6
3.3 Realist model . . . 6

4 Experimental Results **7**
4.0.1 Main method . . . 7
4.0.2 Strategy 1 . . . 7
4.0.3 Strategy 2 . . . 7
4.0.4 Strategy 3 . . . 7
4.0.5 Strategy 4 . . . 7
4.0.6 Strategy 5 . . . 7
4.0.7 Strategy 6 . . . 7

5 Discussion **7**

6 Conclusion **8**

7 References **8**

1 Introduction

This report is the result of my graduation project which completes my Business Information Systems study at Eindhoven University of Technology. The project was performed internally at the University in the Data mining department. In this project we investigated annotations of primitives, more specifically primitives in the scikit-learn library. In the section 1.1 we will briefly explain more about primitives and the annotations. To elaborate on this we will outline the research questions and thesis structure

1.1 problem description

Machine learning is a growing field that can help process the increase of available data[4][3]. Python is a language which holds premade machine learning algorithms in libraries like scikit-learn[**?**]. In recent years python is also increasing in so called market share for machine learning[2]. To help machine learners using the scikit-learn library made a model to indicate what algorithm to use for what problem. In figure **??** you can see that depending on size of the data and early results different algorithms are recommend.

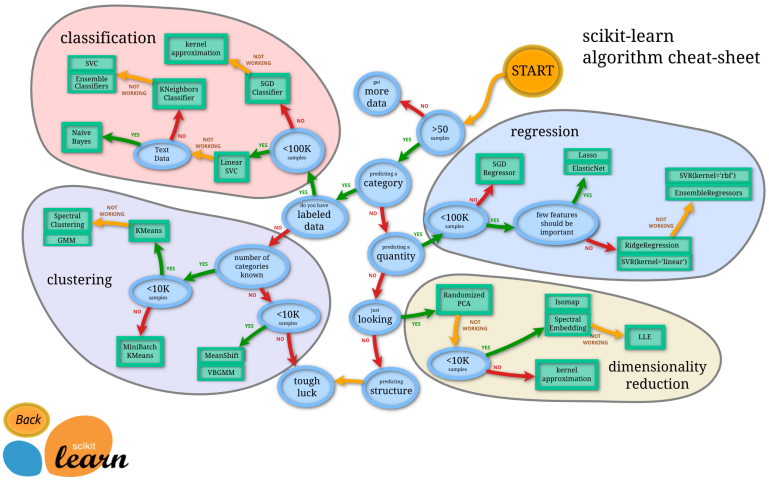

Figure 1: FlowChartML

1.2 research question

We base our research question on the work of Joaquin to give properties to algorithms[5]. More specifically we look more closely to the resilience properties. Earlier research has been done on scalability and resilience to irrelevant variables

1.3 thesis structure

1.4 Outline

2 Preliminaries

Before we discuss in detail the solutions for the steps of our approach, this chapter provides some background knowledge and definitions which are required for a good understanding of the remainder of this thesis.

2.1 Sklearn/scikit-learn library

The scikit-learn library is made in Python

2.2 Terminology

3 Experimental setup

3.1 Motivation

3.2 Description

3.2.1 Main method

3.2.2 Strategy 1

3.2.3 Strategy 2

3.2.4 Strategy 3

3.2.5 Strategy 4

3.2.6 Strategy 5

3.2.7 Strategy 6

3.3 Realist model

4 Experimental Results

4.0.1 Main method

4.0.2 Strategy 1

4.0.3 Strategy 2

4.0.4 Strategy 3

4.0.5 Strategy 4

4.0.6 Strategy 5

4.0.7 Strategy 6

5 Discussion

6 Conclusion

7 References

References

[1] The Popularity of Data Science Software, Robert A. Muenchen, r4stats.com, (2017)

[2] Most Popular Programming Languages For Machine Learning And Data Science,Adarsh Verma, fossbytes.com, (2016)

[3] Machine learning: Trends, perspectives, and prospects, M. I. Jordan, T. M. Mitchell, Science Volume 349 issue 6245 pages 255-260, (2015)

[4] Storage predictions: Will the explosion of data in 2017 be repeated in 2018?, Nick Ismail, www.information-age.com/, (2017)

[5] Understanding Machine Learning Performance with experiment databases, Joaquin Vanschoren, KU Leuven, (2010)

[6] Quantifying the resilience of inductive classification algorithms, M. Hilario, A. Kalousis, Proceedings of the 4th European Conference on Principles of data mining and knowledge discovery, pages 106-115, (2000)