# Ordered Sets in Data Analysis

Alkzir Nafe

December 2019

# Оглавление

# 1. Introduction

This work is devoted to solving the binary classification problem: based on training dataset (train dataset), for which the labels of the target class are known, it is necessary.

It is proposed to use the lazy classification method, that is, not to build the entire set of classifiers on the basis of the training sample, but to make a decision on the classification of a new object when it arrives.

To assess the quality of the methods used, we will use the following popular metrics:

| | | EXPERT REVIEW | |
|---|---|---|---|
| | | Positive | Negative |
| SYSTEM REVIEW | Positive | TP | FP |
| | Negative | FN | TN |

TP - a truly positive decision;

TN is a true negative decision;

FP - false positive decision;

FN is a false negative decision.

$$precision = \frac{TP}{TP+FP} \qquad recall = \frac{TP}{TP+FN} \qquad F = 2\frac{precision.recall}{precision+recall}$$

$$Accuracy = \frac{P}{N}$$

where, P is the number of documents by which the classifier made the right decision, and N is the size of the training sample. An obvious solution that you can start with.

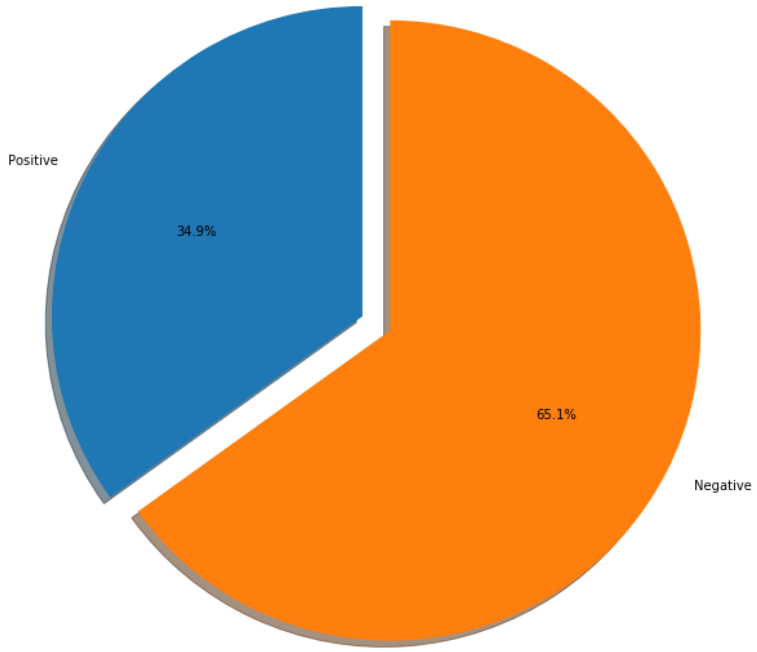
# 1.1. Data set

In this project I decided to use the following database: "Solar Flare Data Set". This database was download from "UCI" web site. The link on the database supporters site: https://archive.ics.uci.edu/ml/datasets/Solar+Flare

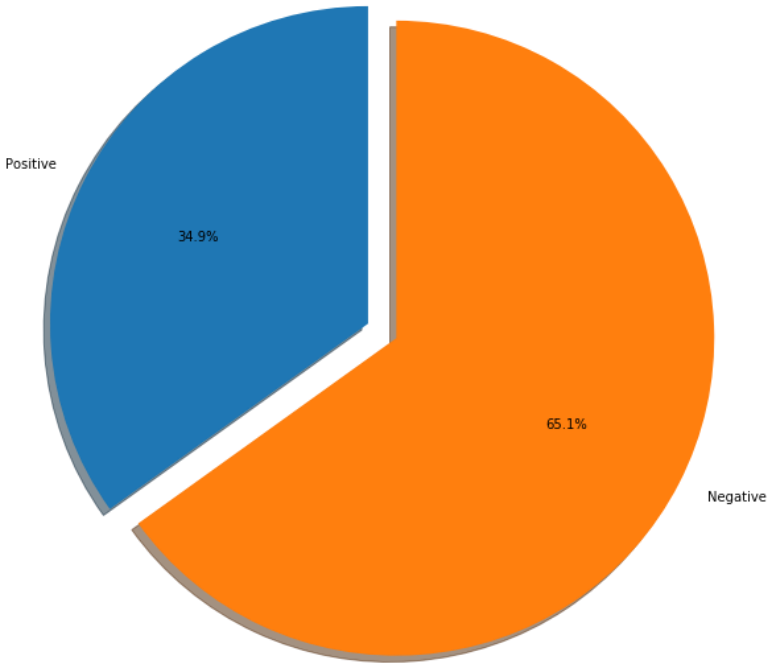
Data Set Information:

Notes:

- The data are divided into two sections. The first section (train.csv) has had much more error. The second section (test.csv) has less error, but he has less information.
- The train section has: 205 negative and 110 positive example.




- The test section has: 141 negative and 32 positive example. And the 32 positive example is so low.



## 1.2. Attribute Information

1. Code for class (modified Zurich class) (A,B,C,D,E,F,H)
2. Code for largest spot size (X,R,S,A,H,K)
3. Code for spot distribution (X,O,I,C)
4. Activity (1 = reduced, 2 = unchanged)
5. Evolution (1 = decay, 2 = no growth, 3 = growth)
6. Previous 24 hour flare activity code (1 = nothing as big as an M1, 2 = one M1, 3 = more activity than one M1)
7. Historically-complex (1 = Yes, 2 = No)
8. Did region become historically complex on this pass across the sun's disk (1 = yes, 2 = no)
9. Area (1 = small, 2 = large)
10. Area of the largest spot (1 = <=5, 2 = >5)
11. C-class flares production by this region in the following 24 hours (common flares); Number
12. M-class flares production by this region in the following 24 hours (moderate flares); Number
13. X-class flares production by this region in the following 24 hours (severe flares); Number

# 2. Describe the binarization strategy or similarity operation

I decided to binaries my dataset. For instance, I have 7 different modified Zurich class groups, so each of them I account as a category.

For example, we have:

modified Zurich class witch have (A,B,C,D,E,F,H)

I convert this to

modified Zurich A: 1, 0

modified Zurich B: 1, 0

modified Zurich C: 1, 0

modified Zurich D: 1, 0

and …

I convert this data set to binary, using my code from Create_my_data.ipynb

# 3. Aggregation

My aggregation function works according to the voting logic. If there is an intersection between the validation element and the train element data, there is a large share of inclusions in the plus set, we vote for plus, otherwise with a minus.

Then we summarize all the votes and, if there are more votes for the plus, we say that the element is positive.

Also, this function has two constants, multiplying by them we can balance between positive and negative guessing. This function can be specified manually, as well as its algorithm can automatically change.

The second option is considered very slow and therefore was not reflected in my report.

This type of aggregation function can give us items with high or low intersection values. This very good work in image.

This algorithm work very slow its complexity is $O(n)^3$

$$\left|\left\{x \in G\pm : \frac{|(g' \cap x'^{\pm}) * C|}{|G \pm|}\right\}\right| / |G \pm|$$

where $\acute{g}$ is the description of the classified object (operator “prime”), $G+$ (G-) is the set of all objects with the target class "+" ( "-" ) , C is the constant.

# 4. The result of works

In this paper, I compared my algorithm with the xgboost algorithm.

First we needed to pick our constants:

C+ and C- is a constant

| C+ | C- | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|---|
| 1 | 1 | 0.693 | 0.90625 | 0.367 | 0.522 |
| 1 | 0.95 | 0.41 | 0.96875 | 0.234 | 0.378 |
| 1 | 0.9 | 0.20 | 1 | 0.189 | 0.318 |
| 1 | 085 | 0.184 | 1 | 0.184 | 0.312 |
| 1 | 0.80 | 0.184 | 1 | 0.184 | 0.312 |
| 1 | 0.75 | 0.184 | 1 | 0.184 | 0.312 |

| | | | | | |
|---|---|---|---|---|---|
| 0.95 | 1 | 0.82 | 0.875 | 0.509 | 0.643 |
| 0.9 | 1 | 0.87 | 0.75 | 0.631 | 0.685 |
| 0.85 | 1 | 0,878 | 0,375 | 0,923 | 0,533 |
| 0,8 | 1 | 0.855 | 0,21 | 1 | 0,35 |
| 0,75 | 1 | 0.826 | 0,0625 | 1 | 0,117 |

How we see the best result is when C+ = 0.95 and C- = 1

Now we compare with the xgboost algorithm

| Max_depth | eta | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|---|
| 1 | 1 | 0.658 | 0.09 | 0.0937 | 0,092 |
| 2 | 1 | 0.658 | 0.09 | 0.0937 | 0,092 |
| 1 | 3 | 0.34 | 0.20 | 0.90 | 0.33 |
| 2 | 3 | 0.34 | 0.20 | 0.90 | 0.33 |

How we see our algorithm work better than xgboost algorithm.

# 5. Conclusions

The tasks were completed. The algorithms considered in the work showed high accuracy compared to the algorithm xgboost.

It was also found that the data can be very specific, and therefore can vary significantly only optimal parameter values of certain algorithms, but also metric values qualities corresponding to the selected optimal values.

As we see in the niche data, there were significant problems with positive data, which is normal for this type of observation, and the classic approach cannot always help in solving such problems.

# 6. List of references


1. Mehdi Kaytoue, Sergei O. Kuznetsov, Amedeo Napoli, Sebastien Duplessis: Mining gene expression data with patter structures in formal concept analysis, published by Elsevier Inc 2010
2. Alexey Masyutin and Yury Kashnitsky: Query-based versus tree-based classification: application to banking data, National Research University HSE Moscow, Russia
3. Sergei O. Kuznetsov: Scalable Knowledge Discovery in Complex Data with Pattern Structures, School of Applied Mathematics and Information Science, National Research University HSE, Moscow, Russia