NOUVELLE Léo

MOREAU-PERNET Baptiste

OLIVIER Antoine

GAUDIN Pierrick

Project Final Report
Avalanches

1. Abstract

This report presents a method to study avalanche risks. Avalanches are a catastrophic event that occur very frequently in high mountains. Hundreds of people die of avalanches every year, even though security means are set in place every year to prevent those accidents: the risk of avalanches is calculated, and people are deterred to go on the dangerous slopes. Consequently, preventing avalanches risks is an important matter of security. We have chosen to use machine learning to study precisely avalanches risks, aiming to provide an algorithm more precise to determine avalanche risks. Indeed, our physical knowledge of avalanches and the data we dispose of are insufficient to carry out reliable predictions of avalanches. Therefore, the usual approach to the prediction problem is statistical and that is why machine learning can be very useful: it can help improve statistical models.

2. Introduction

As winter sports enthusiasts, we wanted to tackle an issue that was linked to mountains. That is why we chose to work on the determination of the risk of avalanche depending on a large set of weather and topology records in ski resorts. Since the beginning of the study of avalanches in the 70s, this phenomenon keeps killing about 30 people per year only in France. Thereby it appears to be of upmost importance for the safety of mountain lovers. Thus, we want to give an answer to an issue that concerns every hiker, skier or snowboarder: how to predict the risk of avalanche? Machine learning applied to avalanche prevention may have many applications in ski resorts as a relevant way to improve security.

3. Problem Description

Working on avalanche prediction using machine learning opposes several difficulties. The first of them is the access to data which is very difficult to find. The database we found and used (reference) counts 402 avalanche records with few features:

Altitude of the avalanche start

Orientation of the slope at the avalanche start

Slope inclination at the avalanche start

Year and period of the year

Forecasted danger level

Furthermore, the database also presented information on avalanche's consequences:

Number of people caught in the avalanche

Number of people buried in the avalanche

Number of casualties due to the avalanche

Given the features available, the problems we tried to address are:

What is the avalanche predicted risk associated to a snow condition? To answer this question, we tried to implement a regression that would return a risk level.

What avalanches are likely to be dramatic (in terms of people buried and casualties)? In this case we chose to determine the chances of survival to an avalanche depending on its properties. Our choice was forced by the documents we found because the only large dataset available is a record of deadly avalanches, with precisions about the number of people caught in the avalanche and the number of people who died in the avalanche.

Those two questions also imply a certain classification of avalanches on which we worked. Concretely, we are going to apply a clustering algorithm on the dataset to try and distinguish clusters in which avalanches are more likely to lead to casualties.

4. Related work

Our first reference is a PDF report of a study called "Optimisation d'un logiciel de prévision d'avalanches à l'aide d'un algorithme génétique". Technically, the database is composed of weather records of 5 ski resorts in Scotland. The prediction is based on a k-nearest-neighbours method using weights to emphasize on specific features. Each point corresponds to a specific weather record for a location. Each sample contains eleven features (air temperature, wind speed, wind direction, etc) and two labels corresponding to the occurrence of an avalanche and to the risk scale. The particularity of this approach is the use of a genetic algorithm. This algorithm is based on the Darwin's principle of evolution. Initially, the weights of the algorithm are randomly allocated. At each generation, the best models are selected and reproduced with modifications. This process is repeated until reaching a stagnation point. According to this report, the previous kNN algorithm had an accuracy of 81%. This optimisation allows the algorithm to reach an accuracy of 84.5%.

However, an article from the Figaro specifies that it is impossible to predict very precisely the occurrence of an avalanche. A local avalanche can be triggered by a plate of ice or the passage of a skier. It is impossible to forecast these conditions, making very difficult the prediction of an avalanche at the scale of a mountain slope. Furthermore, the avalanches records are only based on human observations, it means that there remain many occurrences that are not signalled. This lack of information has an impact on the efficient prediction of avalanches.

5. Methodology

To address the problem, we had to find the best data base, collecting data on avalanches like frequency, characteristics of the start zone, number of people caught, the day of the avalanche, etc. We found a database containing data on avalanches that have been taking place in Switzerland since 1995. The features in the database are about the date, start zone elevation, inclination, orientation, and about the forecasted danger level, the number of caught, fully buried and dead people, along with the activity they were doing at that moment. We tried to find other datasets but there was no other dataset containing data on avalanches and characteristics of the slope around.

The first step was to pre-process the data. To begin with, we suppressed the irrelevant features such as the quality of the coordinates of the avalanche. We also modified unusable features like the aspect of the slope where the avalanche began. For example, we assigned an integer value to a direction.

Also, we wanted to keep the date in our features, to find links between events at the same period in the year, but we wanted to discretize it and assign one integer to one day of the year: We decided to give to June, 1st the number 1, and to assign the integer 31x(month number-6)+day number. We created a scale that doesn't interrupt during the winter, when the matches between dates are the most important.

Then, we determine whether or not there had been survivors thanks to the "number.caught" and the "number.dead" feature corresponding to the number of people concerned by the avalanche and the number of people who died in it. We created a "label" vector, with the row values 1 if there was at least a survivor and 0 if there was not. Then we deleted these features from our dataset.

We split our dataset and our labels list in two parts of 201 and 200 rows. The first one corresponded to the training data and the second one to the test data.

Nearest neighbours' method

Our first idea was to use a nearest neighbours' method as used in the document "Optimisation d'un logiciel de prévision d'avalanches à l'aide d'un algorithme génétique" [1].

The idea is to evaluate the risk of a deadly avalanche by calculating a form of "density" of those avalanches close to a given avalanche. To calculate this density, we use the distance to the 20 nearest neighbours (20 because our dataset has 402 elements and by taking 402's square root we are close to 20). The important thing is then to define the distance that allows to calculate a density. For that, we must weigh the importance of each parameter. This is where the genetic algorithm intervenes.

The genetic algorithm uses the principles of natural selection to select the best solutions. The key point being the possibility to evaluate how good a solution is. This is where our problem wasn't adapted to the method as compared to the study [1]. Indeed, in [1] it was possible to use a fitting function independent from the algorithm's predictions thanks to the data which gave both the conditions when the avalanches occurred but also conditions when an avalanche didn't occur. Furthermore, their goal was easier since they were looking for an algorithm which would predict if an avalanche would occur or not. Contrary to them we are looking for an algorithm providing a continuous risk function.

Clustering Method

Our goal was to find a risk associated to a given situation with parameters of the slope, like for example the elevation of the avalanche start zone, its inclination, etc. A high risk is associated to a high number of caught people, along with fully buried and dead people. We wanted at first to find a risk that an avalanche would take place and would cause deaths, but we missed the data on places where avalanches would have occurred and caused no deaths, or just where the conditions to create an avalanche were present, but without any avalanches.

Then we could have compared to this data and have found a probability for a given situation to produce a deadly avalanche or a non-deadly avalanche, counting the number of avalanches among all the occurrences of the conditions. But we had to find another way of scaling the risk an avalanche had to occur and injure people. Thereby we thought about creating clusters only based on area conditions and see whether these clusters bring information on the risks associated to the avalanches; the risk being based on the other parts of the data, which are the number of dead people, caught people, and the fully buried people. So, we used the KMeans algorithm, that created clusters with the Euclidean distance, on normalized data (data moved between 0 and 1). We based the clusters on the area conditions only, as if we could predict the risks of an avalanche only given a set of local conditions.

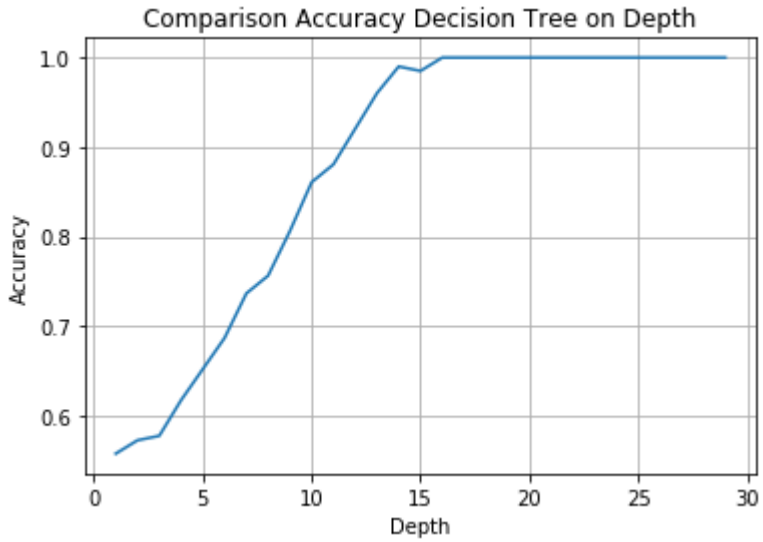
To predict the best number of clusters, I tried to maximize the distances between the clusters, and to find clusters where the number of dead people, fully buried or caught people is discriminant as compared to the others. Consequently, we could then identify sets of avalanches that are dangerous, and why, within the conditions on the local area.

We performed many other kinds of learning algorithms such as the Support Vector Machine, the Neural Network and the logistic regression.

```
##SVM:

from sklearn import svm

t = time()
clf = svm.SVC().fit(data_train, labels_train)
accuracy_svm = np.mean(cross_val_score(clf, data_train, labels_train))
Y_train_svm = clf.predict(data_train)
Y_test_svm = clf.predict(data_test)
deltat_svm = time() - t

print("accuracy Support Vector Machine :", accuracy_svm)
print("time Support Vector Machine :", deltat_svm)

## Random Forest
```

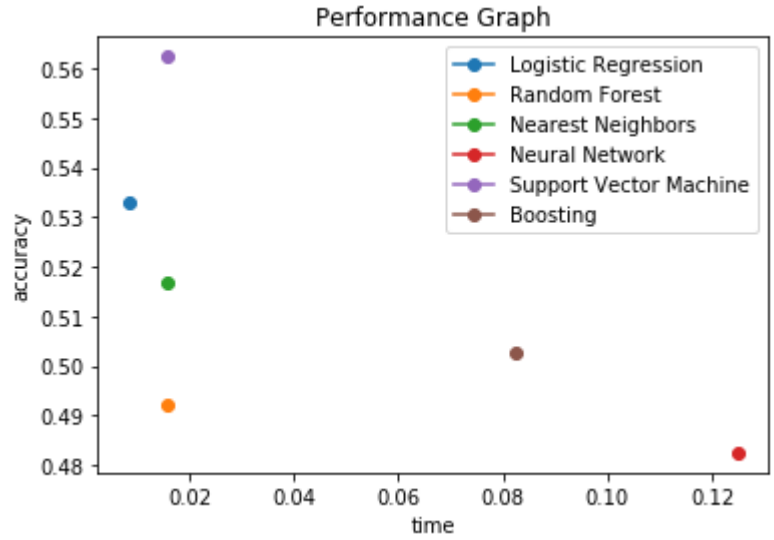
In this example, we calculate the computation time, we train the algorithm on the train dataset and then, we perform a cross validation to determine the accuracy of the algorithm. Finally, we calculate the predicted labels of the test dataset in order to compare them later with the true labels of the test data. This methodology was applied to all our algorithms.

In the case of the Random Forest, we optimised the characteristics of the random trees. We drew the accuracy of the algorithm depending on the depth of the decision trees. We noticed that the accuracy reaches a maximum with a depth of approximately 17. So, we decided to set the maximum depth at 18.



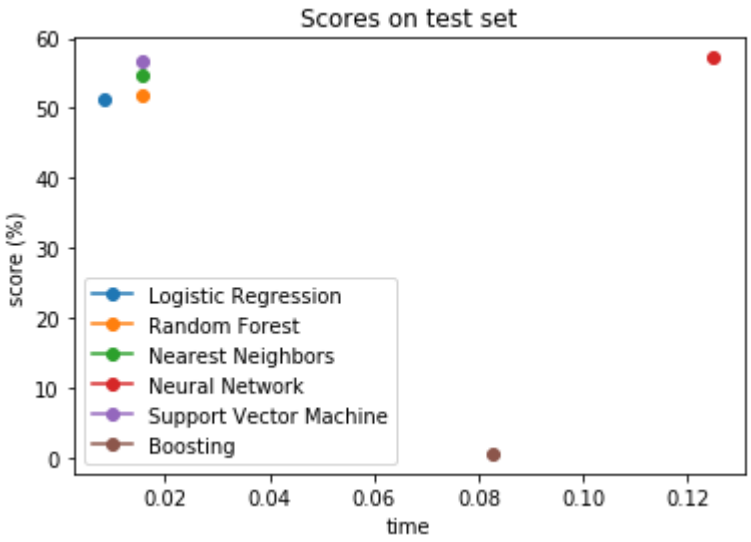
6. Evaluation

First, we plotted the accuracy of the algorithms with the cross-validation method. It helped us to foresee which algorithms were more efficient.



Then, we plotted the score of each algorithm on the test set. The score represents the comparison between the predicted labels and the real label of the test dataset.

```
score_nn = 0
for i in range(200):
    if Y_test_NN[i] == labels_test[i]:
        score_nn += 1
score_nn = score_nn / 201
```



The results of the clustering algorithm are good, because we found some relevant clusters that group avalanches having the same risks (high or low) and the same characteristics. We computed for each cluster the total of avalanches in the cluster, the total number of dead, caught, or fully-buried people, and the ratios dead/count, caught/count, and fully-buried/count. It enabled us to compare the danger of the clustered avalanches in an appropriate way. Then it was interesting to focus on the relevant clusters centers. One of the four clusters was far more dangerous than the others (as the ratios in the following picture show) and the slope characteristics were a high zone elevation, a huge zone inclination, and overall the mean of the date is in august (some people die of summer avalanches occurring high and that are far more dangerous because in more extreme conditions). We also see that the last fully-buried ratio is low, because of the extreme danger and rapidity of this kind of avalanches.

[{'dead': 222.0, 'count': 174, 'caught': 387.0, 'burried': 245.0, 'activity': 432.5513784461153, 'dead/count': 1.2758620689655173, 'caught/count': 2.2241379310344827, 'burried/count': 1.4080459770114941}, {'dead': 156.0, 'count': 131, 'caught': 265.0, 'burried': 175.0, 'activity': 334.0, 'dead/count': 1.1908396946564885, 'caught/count': 2.022900763358778, 'burried/count': 1.3358778625954197}, {'dead': 100.0, 'count': 76, 'caught': 183.0, 'burried': 110.0, 'activity': 196.55137844611528, 'dead/count': 1.3157894736842106, 'caught/count': 2.407894736842105, 'burried/count': 1.4473684210526316}, {'dead': 35.0, 'count': 20, 'caught': 54.0, 'burried': 14.0, 'activity': 60.0, 'dead/count': 1.75, 'caught/count': 2.7, 'burried/count': 0.7}]

7. Conclusion

As we can see in the “Score on test set” graph, the best score we reached is about 60 %. This is quite disappointing because, if we allocate randomly a label to each avalanche of the test set, we statistically obtain a score of 50 % with a range of 200 avalanches.

We can deduce two conclusions of this results. The first one is that our dataset may lack of features. After data analysis, we obtain a dataset with only 6 relevant features. Thereby, our algorithms are underfitted and their predictions are not relevant. The second one is that maybe the characteristics of the avalanche that determine if there will be a survivor are not represented in this dataset. Or worse, this phenomenon could be due to chance and, consequently, could be unpredictable.

We tested this methodology on new databases, to distinguish between the activity type, or adapt it to compare a risk prevision to the current forecasted danger level. However, it revealed to be too imprecise. Finally, the avalanche probably depends on other features that are caracteristic of the slopes themselves (ground compisition, topography for example).

8. Bibliographic References

[1] - Ing.T. LENFANT and Ir R. LESCROART PIERRARD - Virton, article published in the scientific review of I.S.I.L.F. n°19, 2005: "O*ptimisation d'un logiciel de prévision d'avalanches à l'aide d'un algorithme génétique*"

[2] - Marc CHERKI, article published in the Figaro, "*Pourquoi est-il si difficile de prévoir avec précision une avalanche* ?"

http://www.lefigaro.fr/sciences/2017/01/19/01008-20170119ARTFIG00231-pourquoi-est-il-si-difficile-de-prevoir-avec-precision-une-avalanche.php

[3] - P. BOIS and C. OBLED, scientific article published in La Houille Blanche - Revue internationale de l'eau, *"Prévision des avalanches par des méthodes statistiques"*

[4] DATA-AVALANCHE, internet website dedicated to the study and prediction of avalanches,

http://www.data-avalanche.org/

(The first database we found and used)

[5] EnviDat, database "*Fatal avalanche accidents in Switzerland since 1995-1996*"

https://www.envidat.ch/dataset/fatal-avalanche-accidents-switzerland-1995

(The database we used during our project)