POLITECNICO
MILANO 1863

Politecnico di Milano

Scuola di Ingegneria Civile,Ambientale e Territoriale
Master in Geoinformatics

My Thesis

Bresciani Matteo

Referent professor: Brovelli Maria Antonia

April 20, 2022

Contents

Chapter 1

Introduction

1.1 section

Chapter 2

PreProcessing

Data is in most cases incomplete and noisy. Nowadays, dealing with big amount of informations, the probability of incorrect data is higher without a proper data preparation. But only high-quality data can generate accurate models and predictions. Hence, it's crucial to process data with the best possible quality. This step is called data preprocessing, and it's one of the essential steps in data science, artificial intelligence, and machine learning.

2.1 Feature Selection

Feature Selection is an important step in data pre-processing. It consists in selecting the best subset of input variable as the most pertinent. Discarding irrelevant data is essential before applying Machine Learning algorithm in order to:

- **Reduce Overfitting**: less opportunity to make decisions based on noise;
- **Improve Accuracy**: less misleading data means modelling accuracy improves. Predictions can be greatly distorted by redundant attributes;
- **Reduce Training Time**: With less data the algorithms will train faster;

Due to the fact that there isn't a best feature selection technique, many different methods are performed. The aim of this part is to discover by experimentation which one/ones work better for this specific problem. In this part, are shown the supervised methods used, which are classified into 3 groups, based on their different approach.

2.1.1 Filter Methods

Filter-based feature selection methods adopt statistical measures to evaluate the correlation/dependence between input variables. These select features from the

without machine learning algorithm. In terms of computation, they are very fast and are very suitable in order to remove duplicated, correlated, redundant variables. On the contrary, these methods do not remove multicollinearity. The filter-based feature selection methods used are the following:

- **Correlation coeffcients**: Correlation is a statistics measure that shows the strength of association between an independent variable and its target variable. The values range between -1.0 and 1.0. A coefficient of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no relationship between the movement of the two variables. In this study, I selected three types of correlation:
  - **Pearson correlation**: It's one of the most coefficients used in statistics. This measures the strength and direction of a linear relationship between two variables.
  - **Spearmanr correlation coefficient**: It's a non-parametric coeffcient which, instead of Pearson, measures the monotonic relationship between two variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. Hence, since monotonic relation is less restrictive than linear relation, Spearmanr coefficent can provide further informations;
  - **Kendall Correlation**: The Kendall correlation is similar to the Spearman correlation because is non-parametric too but it measures the dependence between two variables instead their correlations;
- **Fisher Score**: F-score is one of the most used supervised feature selection methods. In my study this is implemented using *SelectKBest method*, which compute the score for each variable using the *f_regression()* function to evaluate them.

2.1.2 Wrapper and Embedded Methods

Wrapper methods, as the name suggests, wrap a machine learning model, with different subsets of input features: In this way the subsets are evaluated following the best model performance. Embedded methods instead are characterised by the benefits of both the wrapper and filter methods, by including interactions of features but also having a reasonable computational cost.

- **Exhaustive Feature Selection**:
- **Recursive Feature Selection**:
- **Random Forest Importance**:

2.1.3 Multiscale Geographically Weighted Regression(MGWR)

Due to the fact that this study is related to geographic and spatial data, each pieces of data is very sensitive to the geographic distance between them. So the use of mgwr methods could be innovative, since multivariate models are increasingly encountered in geographical research to estimate spatially varying relationships between a targets and its predictive variables.

Chapter 3

References

3.1 Bibliography

3.2 Software used

- **LaTeX**: used to write and to build the document [`https://www.draw.io/`];
- **GitHub**: used to store and manage project repository [`https://github.com/`];
- **GitHub Desktop**: is the official GitHub application which allows us to contribute to the project repository in an easy way [`https://desktop.github.com`];