An Overview of Existing Machine Learning Methods for Gender Classification of Names

Anna Shleiko, Natalia Borysova, Zoia Kochuieva and Karina Melnyk

National Technical University "Kharkiv Polytechnic Institute", Pushkinska str., 79/2, Kharkiv, Ukraine

Abstract

The paper presents an overview of the existing machine learning methods for solving the problem of gender classification of the authors of the written texts by names: substantiates the relevance of the research topic, analyzes the existing methods of solving the task and selects the direction of further research.

Keywords [1]

Gender classification, supervised machine learning, methods of classification

1. Introduction

The problem of determining the gender of the authors of the Ukrainian corpora texts is of remarkable importance, since it enables increased functionality and improved performance, when using the corpora utilizing all of its functions. Due to the fact that the considered problem belongs to the area of classification problems, it can be solved using machine learning methods.

2. Machine learning methods for classification

A classification is the process of predicting the class of given data points. Classification requires machine learning algorithms in order to learn how to assign a class label to examples from the problem domain. In other words, during classification, a prediction is made for a class label for a certain example of the input data. A typical classifier uses a set of training data in order to understand how given input variables are related to a certain class [1]. Classification falls into the category of supervised learning, according to machine learning terminology, meaning that the learning occurs where a training set of correctly identified observations is available [2]. The analysis of works [1-3] showed that there are many classification methods for solving the task of determining (Fig. 1).

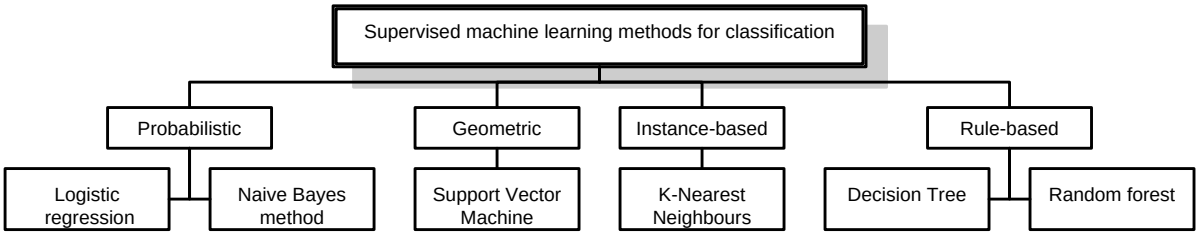

Figure 1: Supervised Machine Learning Methods for Classification

Comparison of the supervised machine learning methods for classification is presented in a more detailed way in Table 1 [1-3].

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: shleykoa@gmail.com (A. Shleiko); borysova.n.v@gmail.com (N. Borysova); aliseiko@gmail.com (Z. Kochuieva); karina.v.melnyk@gmail.com (K. Melnyk)
ORCID: 0000-0002-8834-2536 (N. Borysova); 0000-0002-4300-3370 (Z. Kochuieva); 0000-0001-9642-5414 (K. Melnyk)

Table 1
Comparison of the Machine Learning Methods for Classification

Method	Advantages	Disadvantages
Logistic regression	method is designed for this purpose (classification), and is most useful for understanding the influence of several independent variables on a single outcome variable	works only when the predicted variable is binary, assumes all predictors are independent of each other and assumes data is free of missing values
Naive Bayes method	method requires a small amount of training data to estimate the necessary parameters, it's extremely fast compared to more sophisticated methods	if we have the combination of features sometimes we can't explain the dependence of the classification result on them
K-Nearest Neighbours	method is simple to implement, robust to noisy training data, and effective if training data is large	it needs to determine the value of K and the computation cost is high as it needs to compute the distance of each instance to all the training samples
Decision Tree	method is simple to understand and visualize, requires little data preparation, and can handle both numerical and categorical data	it can create complex trees that do not generalize well, and decision trees can be unstable because small variations in the data might result in a completely different tree being generated
Random forest	reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases	slow real time prediction, difficult to implement, and complex algorithm
Support Vector Machine	effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient	method does not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation

As we can see, all the methods have both advantages and disadvantages. This must be taken into account when choosing the appropriate one or more methods.

3. Conclusions

After analyzing various methods of classification, it has been decided to use several supervised machine learning methods to solve the task of determining. This will provide an opportunity to compare the results of their work to choose the most fitting one.

4. References

[1] A. Sidath, Machine Learning Classifiers, 2018. URL: https://towardsdatascience.com/ machine-learning-classifiers-a5cc4e1b0623
[2] E. Alpaydin, Introduction to Machine Learning, third, 3rd. ed., MIT Press, Cambridge, MA, 2015.
[3] R. Garg, 7 Types of Classification Algorithms. URL: https://analyticsindiamag.com/7-types-classification-algorithms/