Letter Recognition using SVM, KNN, and Decision Trees

Ujjawal Saini

Guru Gobind Singh Indraprastha University

B.Tech AIML

Instructor*: Prof. Amit Choudhary*

Date: 18-05-2023

Basic Information

Title of Project: Letter Recognition using SVM, KNN, and Decision Trees

Student Name: Ujjawal Saini

Enrollment Number: **REDACTED**

Email ID: **REDACTED**

Contact Number: **REDACTED**

Google Drive Link: **REDACTED**

Google Website Link: https://spignelon.co/mlminor1

YouTube Video Link: **REDACTED**

Note: The links provided herewith might be empty or non-responsive yet, they will soon be updated with relevant information.

Title: Using SVM, KNN, and Decision Trees for Letter Recognition

1. Introduction

Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Tree algorithms will be used in the research project to create a machine learning model for letter identification. The dataset utilized in this work is made up of 20,000 rectangular black-and-white pixels that represent the capital letters of the English alphabet and was taken from the UCI Machine Learning Repository [1]. Based on 20 different fonts, each letter picture is randomly warped. 16 basic numerical attributes (statistical moments and edge counts) from the dataset are provided, and they have been scaled to integer values from 0 to 15. The model will be tested on the remaining 4,000 samples after being trained on a subset of 16,000 examples.

2. Methodology

The project will use the well-known Python machine learning framework scikit-learn to put the following algorithms into practice and assess them:

a. SVMs, or support vector machines: SVM is a potent classifier that creates hyperplanes to divide classes in a high-dimensional feature space. It has a lot of applications in many different fields, like pattern recognition and image categorization.

b.K-Nearest Neighbors (KNN) is a non-parametric classifier that assigns a data point to the class that is most prevalent among its k nearest neighbors in the feature space. Due to its efficiency and simplicity, KNN has been effectively used in letter recognition applications.

c. Decision Trees are hierarchical, tree-like structures that recursively divide the feature space according to the values of the attribute. They are renowned for being easy to understand and having the capacity to work with both category and numerical data.

3. Data Preprocessing

The dataset will go through preprocessing procedures before the machine learning techniques are applied. The handling of missing data, the elimination of any superfluous attributes, feature scaling, and the division of the data into training and testing sets are some of these procedures. The first 16,000 samples will be used as the training set, and the final 4,000 samples will be used to test how well the model performs.

4. Model Development and Evaluation

Utilizing scikit-learn, the chosen machine learning methods will be put into practice. Using relevant performance criteria, such as accuracy, precision, recall, and F1-score, the models will be trained on the training set and assessed on the testing set. The goal is to evaluate the letter recognition performance of SVM, KNN, and Decision Trees and determine which method performs the best.

5. Conclusion

Using SVM, KNN, and Decision Trees techniques, the research project intends to construct and compare machine learning models for letter recognition. These algorithms will be used in the research to examine the efficacy of several methods for correctly identifying capital letters in the English alphabet. The most effective method will be chosen after the models have been assessed using the proper performance indicators.

References:

1. Dua, D., & Graff, C. (2017). UCI machine learning repository.

2. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.

3. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

4. Breiman, L. (2017). Classification and regression trees. Routledge.