Project Proposal – Milestone 1

Qi Li

li.qi8@northeastern.edu

Percentage of Effort Contributed by Student : 100%

Signature of Student : Qi Li

Submission Date: Jan. 24. 2022

● Introduction

1) What is the business problem you are going to work on in this course?

I'm going to use Letter Recognition Data Set to finish my project. This data set is about classifying 26 English capital letters. By analyzing this data set, I'm going to identify a large number of black-and-white rectangular pixel as one of the 26 capital letters in the English alphabet.

2) Explain a few sentences about the business problem

The "letters recognition data" from the UCI machine learning repository contains summary statistics (from image analysis) of randomly-distorted English letters in 20 different fonts and each letter was randomly distorted to 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes. The research for this article investigates the ability of several variations of Holland-style adaptive classifier systems to learn to correctly guess the letter categories associated with vectors of 16 integer attributes extracted from raster scan images of the letters.

3) Explain one or two sentences on the data set

This data set has 17 columns (the first column is the capital letter from A to Z and the rest 16 columns are the numerical attributes). They trained on the first 16000 items and then use the resulting model to predict the letter category for the remaining 4000 items. The data is already mined and statistically scaled so it is possible to directly start with classification.

● Problem statement

1) Write down your hypothesis.

I'm going to apply different machine learning models to Letter Recognition Data Set in my project. By testing different models to the data set, I can find out which model applies to this data set best.

2) How your machine approach or automation is going to solve the business use case?

I'll separate the data set into training set and validation set: 16000 for training set and 4000 for validation set. Then build a variety of classification models (such as SVM, MLP, KNN models) using the training data set. After then, I'll use the validation data set to assess performance of the models and pick the best one.

3) What does the business get out of this solution?

By classify these 20000 unique stimuli from 20 different fonts, we can better distinguish the 26 capital letters in the English alphabet.