# CDS 403 Midterm Project: Letter Recognition

Alessa Ogork

October 18, 2022

## 1 Introduction

The objective of this project is to evaluate the accuracy and efficiency of the K-Nearest Neighbors algorithm (KNN) in the classification of the letter-recognition data set from the UCI machine learning repository. We will compare these results with that of Support Vector Machine algorithms (SVM) that have been developed by other data scientists using this data set. The process of letter recognition using machine learning algorithms involves classifying black and white pixel displays into groups representing one of the 26 letters of the English alphabet. Note that these pixel images are of capital letters of varying sizes. As mentioned in the opening statement, the data set that will be used is the original letter-recognition data set from the UCI machine learning repository. The letter-recognition data set is a multi-class classification data set that is most widely known for its experimentation and affinity for SVM algorithms. While many projects involving the letter-recognition data set have focused on using, improving, and evaluating SVM algorithms, there is a lack of exploration regarding other methods of classification, including KNN algorithms. This is mainly because SVM algorithms are fast and can classify images in approximately ten seconds, as opposed to KNN algorithms, which take anywhere from forty to fifty seconds to classify the same image. The assumption for the hypothesis is that the KNN algorithm will perform slightly better in accuracy than the SVM algorithm.

## 2 Literature Review

1. *Stephen D. Bay, Nearest neighbor classification from multiple feature subsets, Intelligent Data Analysis, Volume 3, Issue 3, 1999, Pages 191-209, ISSN 1088-467X, https://doi.org/10.1016/S1088-467X(99)00018-9. (https://www.sciencedirect.com/science/article/pii/S1088467X99000189)*

The aim of this project was to improve the accuracy of the nearest neighbor algorithm (NN). While there are many algorithms, like boosting, bagging, and error correcting output coding, that can increase the accuracy of classifiers, like decision trees, neural networks, and rule learners, these are ineffective against

NN classifiers. The researchers focused on the MFS algorithm, which is a combing algorithm that was designed specifically to improve the accuracy of the NN classifier. The MFS algorithm created was effective because it combined several NN classifiers that only use a random subset of features from the given data. From their research, it was discovered that the MFS algorithm was effective in increasing the NN algorithm accuracy on 25 data sets used from the UCI repository, including the letter-recognition data set. The MFS algorithm significantly outperformed the NN algorithm in general and several of its variants. It is also significant to note that this algorithm was competitive with boosted decision trees, mainly because it is robust to irrelevant features and can reduce bias with respect to components of error. One limitation of this project was that they did not create the NN algorithms that were used as comparisons to the resulting MFS algorithm. Instead of attempting to improve a singular NN algorithm or creating an original NN algorithm, they opted to use a combination of other researchers' NN algorithms that had been previously established. The MFS algorithm was a combination of NN algorithms that had their own flaws, respectively; a more thorough review would address the specific holes these NN algorithms were missing that were sufficiently accounted for with the MFS algorithm.

## 3 Sources

1.
http://odds.cs.stonybrook.edu/letter-recognition-dataset/:~:text=The%20original%20letter%20recognition%20dataset%20from%20UCI%20machine,of%20the%20alphabet%20are%20represented%20in%2016%20dimensions.

2.
https://scholarworks.boisestate.edu/icur/2017/PosterSession/139/: :text=K-Nearest%20Neighbor%20%28KNN%29%20and%20Support%20Vector%20Machine%20%28SVM%29,KNN%20and%20SVM%20each%20have%20strengths%20and%20weaknesses.

3.
https://www.ibm.com/topics/knn

4.
*Stephen D. Bay, Nearest neighbor classification from multiple feature subsets, Intelligent Data Analysis, Volume 3, Issue 3, 1999, Pages 191-209, ISSN 1088-467X, https://doi.org/10.1016/S1088-467X(99)00018-9.*
*(https://www.sciencedirect.com/science/article/pii/S1088467X99000189)*