A Comparison of Supervised Learning Algorithms: KNN, SVM, ANN

Adrienne Chang (A12125503)

Abstract

In the paper An Empirical Comparison of Supervised Learning Algorithms, Caruana and Niculescu-Mizil performed large scale empirical comparisons between supervised learning methods published since the last comprehensive analysis in the 90's. We'll be looking to emulate three of the classifiers used in the paper which will then be used to train and test three datasets from the UCI repository. Accuracies for three different partitions will be reported along with the hyper-parameter.

1. Introduction

Since the Statlog Project concluded in the 90's, multiple new algorithms have introduced that are more efficient and accurate than previous algorithms. Caruana and Niculescu-Mizil concluded a large scale comparison of ten supervised learning algorithms: SVMs, neural nets, logistic regression, naive bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps. We'll be seeking to emulate Support Vector Machine with an rbf kernel, k-Nearest Neighbors, and Artificial Neural Network on the datasets Adult, Covertype, and Letter Recognition from the UCI Machine Learning Repository. These learning algorithms will be trained and tested on these datasets three times with different partitions to evaluate and compare the accuracy of the three learning algorithms.

2. Data and Problem Description

This section describes the three datasets that were selected from UCI's machine learning repository.

Table 1. Overview of Datasets

Name	# of Instances	# of Attributes	Missing Values?
ADULT	48,842	14	Yes
COV_TYPE	581,012	54	No
LETTER	20,000	16	No

2.1 ADULT Dataset

ADULT's data was extracted from the 1994 Census from the census bureau database and the goal is to predict whether the individual earns less than or greater than 50K/year based on the census data.

Table 2. Attributes of ADULT Dataset

Name	Description
Age	Continuous
Workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
fnlwgt	Continuous
Education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
Education Num	Continuous
Marital Status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Sex	Continuous
Capital Gain	Continuous
Capital Loss	Continuous
Hours per Week	Continuous
Native Country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

2.2 COV_TYPE Dataset

COV_TYPE's independent variables was derived from data originally obtained from US Geological Survey and USFS data. The study collects information from four wilderness areas Roosevelt National Forest of northern Colorado due to their lack of human disturbance. The instances collected are then used to determine which of the seven forest cover types the wilderness area has. Only twelve measures are listed but once broken down into their respective options, total to 54 attributes.

Table 3. Attributes of COV_TYPE Dataset

Name	Description
Elevation	Continuous
Aspect	0-180 (degrees)
Slope	0-180 (degrees)
Horizontal Distance to Hydrology	Continuous
Vertical Distance to Hydrology	Continuous
Horizontal Distance to Roadways	Continuous
Hillshade at 9PM	0-255 (meters)
Hillshade at Noon	0-255 (meters)
Hillshade at 3PM	0-255 (meters)
Horizontal Distance to Fire Points	Continuous
Wilderness Area	Rawah Wilderness Area (1), Neota Wilderness Area (2), Comanche Peak Wilderness Area (3), Cache la Poudre Wilderness Area(4)
Soil Type	1-40 (ELUs)

2.3 LETTER Dataset

LETTER's dataset is a collection of the English alphabet letters that are make of black and white rectangular pixels. The letters themselves are composed of 20 different fonts that are randomly distorted to create unique characters. Each character is then broken down into statistical moments and edge counts to create the attributes.

Table 3. Attributes of LETTER Dataset

Name	Description
x-box	Horizontal position of box
y-box	Vertical position of box
width	Width of box
heigh	Height of box
onpix	Total number of pixels
x-bar	Mean x of on pixels in box
y-bar	Mean y of on pixels in box
x2bar	Mean x variance
y2bar	Mean y variance
xybar	Mean x y correlation
x2ybr	Mean of x*x*y
xy2bar	Mean of x*y*y
x-ege	Mean edge count left to right
xegvy	Correlation of x-ege with y
y-ege	Mean edge count bottom to top
yegvx	Correlation of y-ege with x

2.4 Problem

With all of the new supervised learning methods that have been introduced, a comprehensive analysis was needed to compare the pros and cons each method would present depending on the dataset. Our model is limited to three learning methods but the datasets chosen are varying in the types and number of attributes. Depending on the dataset, we'll be able to compare the performances of the three methods and analyze which method is better for the type of dataset.

3. Method Description

Prior to analyzing the data, the datasets were obtained in the data files that were first converted into csv files for easier implementation. Before each trial is run, the dataset is first randomized and then the first 5000 instances are extracted for training and testing. The classification column where the instance is recorded as positive or negative, is then extracted as a separate column to be later used for testing. Cross-validation with five folds and grid search are then applied to each of the

methods. Three trials are then performed with the first training/validation set being 1000 instances, the second being 2500 instances, and the final third being 4000 instances and the rest used as the testing set. Within the training/validation set, three different partitions are applied: 20% training and 80% validation, 50% training and 50% validation, and 80% training and 20% validation. After the datasets are tailored for binary classification, which are detailed in the section below, the average training and validation score are calculated and the best parameters for the respective method are collected. Cross validation and grid search are again performed on the dataset so that the resulting fit with the testing set can give the best test accuracy.

3.1 Datasets

The datasets consisted of raw data with many variations that first needed to be adjusted before we could implement binary classifications. This section describes what changes were made to tailor the dataset.

ADULT: Unlike the other two datasets, the ADULT database has missing values in multiple attributes and instances. To ensure that the data is recreated like the dataset used in Caruana and Niculescu-Mizil's paper, any instance with a missing value will be removed, giving us a total of 45222 instances. For analysis, income of greater than 50K was labeled as positive and income of less than 50K was labeled as negative.

COV_TYPE: To convert our data into a binary problem, we chose to focus on just class 2 of the seven forest cover types. Of the 581,012 instances, we'll be observing whether the instance will be classified as a Lodgepole Pine (positive) or as one of the other six cover types (negative). According to the data description, a number of 283301 Lodgepole Pines were recorded.

LETTER: In Caruana and Niculescu-Mizil's analysis, two classifications were performed on the data. The first implementation labeled the letter 'O' as positive and the other 25 letters as negative. This analysis will look at the more balanced implementation where the letters 'A-M' are labeled positive and the letters 'N-Z' are labeled negative.

3.2 Method Descriptions

KNN: used 26 values of k that are evenly separated between 1 to the maximum fold.

SVM: only the rbf (radial basis function) kernel is implemented with parameters of C = {1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 1e1, 1e2, 1e3} and gamma = {1e-3, 5e-3, 1e-2, 5e-2, 1e-1, 5e-1, 1, 2}.

ANN: implemented with parameters of hidden units = {1, 2, 4, 8, 32, 128}, momentum = {0, 0.2, 0.5, 0.9} and epochs = {50, 100, 150, 200, 250, 300, 350, 400}.

4. Experiment

This section will report for each method, dataset, and partition, the training/validation accuracies and test accuracies averaged over three trials, followed by an an analysis of the data collected. For best parameters, see attached code.

4.1 KNN Results and Analysis

Table 4. KNN of ADULT Dataset

Partition (Training/ Validation)	Average Train Accuracy	Average Validation Accuracy	Average Test Accuracy
20% / 80%	74.59%	73.20%	76.48%
50% / 50%	75.75%	74.53%	76.68%
80% / 20%	76.71%	75.54%	74.90%

Table 5. KNN of COV_TYPE Dataset

Partition (Training/ Validation)	Average Train Accuracy	Average Validation Accuracy	Average Test Accuracy
20% / 80%	76.67%	76.22%	89.50%
50% / 50%	76.14%	75.67%	89.67%
80% / 20%	76.45%	75.99%	89.33%

Table 6. KNN of LETTER Dataset

Partition (Training/ Validation)	Average Train Accuracy	Average Validation Accuracy	Average Test Accuracy
20% / 80%	66.51%	66.19%	95.06%
50% / 50%	67.15%	66.56%	95.49%
80% / 20%	68.37%	67.96%	94.99%

For the KNN algorithm, it can be observed as a general trend that the larger the training set became, the more inaccurate the test accuracy became. At the same time, all methods performed the best when the training and validation set were equally split, although the differences were minute. We can conclude from this general observation that the best partition to use for KNN is 50% training and 50% validation.

Looking at the test accuracies of all three methods, KNN performed the best on the LETTER dataset. In comparison, the training and validation accuracies for the LETTER dataset were the worst out of the three. This anomaly where the training accuracy is less that the test accuracy, seen in both the COV_TYPE and LETTER dataset, can most likely be attributed to the fact that the extracted data is only a small fraction of the entire dataset. Had the entire dataset been used, the results would have been more accurate and consistent.

4.2 SVM Results and Analysis

Table 7. SVM of ADULT Dataset

Partition (Training/ Validation)	Average Train Accuracy	Average Validation Accuracy	Average Test Accuracy
20% / 80%	85.41%	77.07%	75.35%
50% / 50%	84.65%	75.85%	75.56%
80% / 20%	84.49%	75.58%	75.90%

Table 8. SVM of COV_TYPE Dataset

Partition (Training/ Validation)	Average Train Accuracy	Average Validation Accuracy	Average Test Accuracy
20% / 80%	80.16%	68.95%	72.20%
50% / 50%	80.50%	69.48%	69.60%
80% / 20%	80.61%	69.63%	68.6%

Table 9. SVM of LETTER Dataset

Partition (Training/ Validation)	Average Train Accuracy	Average Validation Accuracy	Average Test Accuracy
20% / 80%	68.07%	60.06%	89.80%
50% / 50%	69.32%	63.12%	93.8%
80% / 20%	69.69%	64.03%	95.30%

For the SVM algorithm with an rbf kernel, we can observe a lack of a general trend for the test accuracies and the partitions. The SVM stays about the same for ADULT, decreases for COV_TYPE, and increases for LETTER. Both ADULT and COV_TYPE datasets followed the convention where training sets perform better than testing sets. LETTER continues to be the anomaly where the accuracies are the exact opposite. This can still be explained by the fact that the extracted data is a small portion of the entire dataset.

With training accuracies, we can see the trend between COV_TYPE and LETTER dataset that as the training sets grew, the training and validation accuracies also increased accordingly. This can be attributed to the fact that the larger the training set, the more support vectors that can be drawn to the decision boundary, thus increasing the testing accuracy. Generally speaking, using a larger training set works better with SVM.

4.3 ANN Results and Analysis

Table 10. ANN of ADULT Dataset

Partition (Training/ Validation)	Average Train Accuracy	Average Validation Accuracy	Average Test Accuracy
20% / 80%	58.83%	58.50%	80.03%
50% / 50%	62.14%	62.10%	73.76%
80% / 20%	64.60%	64.61%	80.80%

Table 11. ANN of COV_TYPE Dataset

Partition (Training/ Validation)	Average Train Accuracy	Average Validation Accuracy	Average Test Accuracy
20% / 80%	65.22%	64.84%	82.55%
50% / 50%	69.80%	69.61%	80.12%
80% / 20%	72.34%	72.08%	86.40%

Table 12. ANN of LETTER Dataset

Partition (Training/ Validation)	Average Train Accuracy	Average Validation Accuracy	Average Test Accuracy
20% / 80%	70.53%	69.88%	89.4%
50% / 50%	71.93%	70.89%	87.60%
80% / 20%	64.78%	63.03%	84.28%

Lastly, the ANN algorithm produces the most consistent testing accuracies although they are not the highest. For all three datasets, we can see that the training accuracies are lower than their respective testing accuracies. Almost all datasets generally performed worse when the training and validation sets were split equally. All of these inconsistencies can possibly be explained by the small amount of data that was used. Since ANN is a neural network, it performs better as it's given more data to learn from. In order to get the best results for each dataset, all of the dataset's instances should be used.

4.4 Overview of Results

Table 12. Mean Test Accuracies

Method	Dataset	Average Testing Accuracy/Dataset	Average Testing Accuracy/Method
KNN	ADULT	76.02%	86.90%
	COV_TYPE	89.50%	
	LETTERS	95.18%	
SVM	ADULT	75.60%	79.57%
	COV_TYPE	70.13%	
	LETTERS	92.97%	
ANN	ADULT	78.20%	82.77%
	COV_TYPE	83.02%	
	LETTERS	87.09%	

5. Conclusion

From all three methods, the LETTER dataset performed the best with its testing accuracy going above 90% with the KNN algorithm and the ADULT dataset generally performing the worst. Comparing our results with Caruana and Niculescu-Mizil's results, our dataset wasn't too far off from their accuracies.

Table 13. Test Accuracies Comparison

Method	Results	Literature Data (Caruana and Niculescu-Mizil)
KNN	86.90%	81.50%
SVM	79.57%	78.10%
ANN	82.77%	84.20%

The differences in the testing accuracies can be attributed to the varied implementation of parameters and the use of only 5000 instances of the entire datasets. Overall, the results obtained could be greatly improved by implementing the algorithms on the entire dataset with a larger number of metrics and partitions to really observe the pros and cons of the three algorithms for each dataset.

References

Blackard, Jock A, et al. UCI Machine Learning Repository: Covertype Data Set, 1 Aug. 1998, archive.ics.uci.edu/ml/datasets/covertype.

Caruana, Rich, and Alexandru Niculescu-Mizil. An Empirical Comparison of Supervised Learning Algorithms. Cornell University, 2006, www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf.

Kohavi, Ronny, and Barry Becker. *UCI Machine Learning Repository: Adult Data Set*, 1 May 1996, archive.ics.uci.edu/ml/datasets/Adult.

Slate, David J. UCI Machine Learning Repository: Letter Recognition Data Set, 1 Jan. 1991, archive.ics.uci.edu/ml/datasets/Letter+Recognition.