Binghong Yu

Professor Zhuowen Tu

COGS 118A

18 March 2020

Cogs 118A final project

Abstract:

Machine learning has become more and more popular in the last decade due to its broad application area. According to paper *An Empirical Comparison of Supervised Learning Algorithms*, the authors compare performance of ten learning algorithms including SVMs, neural nets, logistic regression, naive bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps across a large-scale of data. This paper reproduces the experiments Rich and Alexandru did with three different data sets from Kaggle and UCI dataset and three classification algorithms. Three classification algorithms including decision tree, SVM and KNN are evaluated in this paper.

1. Introduction

Machine learning gradually has a significant influence on many fields in technology. Its application is becoming broader, as Jordan stated in his article, "robotics and autonomous vehicle control , speech processing and natural language processing, neuroscience research, and applications in computer vision.(Jordan, 2015)" When performing a classification problem using machine learning, we use features that are continuous or categorical or binary to train these models. When training these models using certain algorithms, if the labels(the correct output) are

given, then it is called supervised learning. The application of supervised learning is also quite broad as more new learning algorithms emerged, including "cancer prognosis and prediction(Cruz,2016), population genetics(Schrider, 2018), real-time IP traffic classification(Nguyen, 2018), etc." As more new learning algorithms emerged, the significance of comparison and evaluation of these learning algorithms has been recognized. According to Loog, "A common method for comparing supervised ML algorithms is to perform statistical comparisons of the accuracies of trained classifiers on specific datasets." (Loog, 2007) This is exactly what Caruana and Niculescu-Mizil's paper and my paper has been working on. In their paper, Caruana and Niculescu-Mizil compared empirically 10 learning algorithms with 8 metrics across a large-scale of data, whereas in my paper, I use the overall accuracy metric to evaluate the performance of each classifier. This paper evaluates three classifiers with three different datasets from Kaggle and UCI dataset in different aspects. These datasets include data from the astronomy field: "predicting Pulsar Star", the medical field: "Cardiovascular Disease dataset" and finance area: "Bank Marketing Data Set". They are specifically chosen to perform binary classification as the last column of these dataset are binary outcomes. The number of observations for each dataset ranges from four thousand to ten thousand, therefore ensuring the size is enough to perform such classification.

In addition, I also want to propose that as more data are fed into the model, the more accurate the dataset will be. This assumption is testified by the result of the experiment, since there is an improvement in accuracy from smaller dataset to larger dataset.

2. Methods:

Dataset description:

The three dataset could be used for binary classification.

The first dataset is called PREDICTING A PULSAR STAR with 17,898 observations. Each observation has 8 features that describe the star. According to kaggel website, the first four features are simple statistics obtained from the integrated pulse profile. The last four variables are obtained from the DM-SNR curve.

The second dataset is called Cardiovascular Disease dataset with 5000 observations. Each observation has 12 features that describe personal information about the patient, such as age, gender, cholesterol, etc.

The third dataset is called Bank Marketing Data Set. It comes from UCI dataset and it has 41188 observations. Each observation has 20 features describing the job, education, etc.

The three dataset could be used for binary classification since their output are binary and could be used for binary classification.

In this section, a detailed introduction of three classifiers will be presented to readers.

Classification description:

Support vector machine:

SVM stands for support vector machine. The major pursuit of a SVM classifier is to find a hyperplane that maximizes the margin and splits the classes. The hyperplane should also maximize the distance to the nearest data samples(Shmilovici, 2005). The application of SVM spread from “recognizing fraudulent credit card activity to recognizing handwritten digits”(Noble, 2006). Another wide application is in the biological field: “the automatic

classification of microarray gene expression profiles". (Noble, 2006)

KNN:

The basic logic behind KNN is that given a dataset to be classified, KNN searches for the k nearest neighbors based on some similarity features. Then according to the similarity score, KNN ranks those neighbors(Phung et al., 2018). According to Phung, one of its applications is that it could be used for text categorization. In his book, he also mentions the drawback of kNN is that it is hard to determine the value of k and "complexity of computation time needed to traverse all the training documents."(Phung et al., 2018)

Decision tree:

Decision Tree is a classifier that is used to do classification which concludes the value of a dependent variable with given independent variables, which are labels. In the book, the author also mentioned the advantage of decision tree: it could deal with different kinds of input, and easier to achieve high accuracy with even a small dataset(Rokach, 2010). Because of its advantages, its application spread into "many areas such as data and text mining, information extraction, and pattern recognition."(Rokach, 2010)

3. Experiment

For each dataset, the observations are partitioned by 80% training data set versus 20% testing dataset , 50% training data set versus 50% testing dataset and 20% training data set versus 80% testing dataset. Then we repeat three trials by method shuffle. Therefore, there is a total of 3 trials/repeats X 3 classifiers X 3 datasets X 3 partitions (20/80, 50/50, 80/20) presented.

After loading these csv files into the jupyter notebook, we first clean our dataset by using isnull.sum() to check if there is any null element inside the dataset. Then, since some of the features are recorded in text, one-hot encoding method is used to encode these text data into numbers by using LabelEncoder. The heat map for each trail and each classifier is also drawn. Grid search is used to find the best hyper-parameter for the model that has the highest accuracy. The average training error, training accuracy, testing error, testing accuracy, validation error and validation accuracy are calculated and presented in the table below to help readers better visualize the result.

Train accuracy	Dataset 1 star			Dataset 2 cardio			Dataset 3 bank		
	80/20	20/80	50/50	80/20	20/80	50/50	80/20	20/80	50/50
SVM	0.9817	0.9801	0.9800	0.7605	0.8885	0.8104	0.9051	0.8872	0.9229
KNN	0.9745	0.9761	0.9760	0.6824	0.6966	0.6668	0.8962	0.8828	0.8986
Decision tree	0.9790	0.9817	0.9786	0.7357	0.7324	0.7398	0.9050	0.8931	0.9101

Table 1.1 This table displays the average train accuracy for different classifiers, different dataset and different partitions. It is calculated by summing up train accuracy of each trial (3 in total) then taking its average.

Train accuracy (avg)	Dataset 1	Dataset 2	Dataset 3	Total avg
SVM	0.9806	0.7262	0.9051	0.8706
KNN	0.9755	0.6819	0.8925	0.8499
Decision tree	0.9798	0.7359	0.9027	0.8728

Table 1.2 This table displays the performance of train accuracy of a classifier on three datasets and its overall performance. It is calculated by sum up for three partitions and then take its average.

Test accuracy	Dataset 1 star			Dataset 2 cardio			Dataset 3 bank		
	80/20	20/80	50/50	80/20	20/80	50/50	80/20	20/80	50/50
SVM	0.9791	0.9744	0.9786	0.6755	0.6276	0.6577	0.9144	0.8956	0.8914
KNN	0.9741	0.9721	0.9723	0.5586	0.5567	0.5623	0.9077	0.8827	0.8796
Decision tree	0.9793	0.9767	0.9781	0.703	0.6965	0.7136	0.9078	0.8836	0.8836

Table 2.1 This table displays the average test accuracy for different classifiers, different dataset and different partitions. It is calculated by summing up the test accuracy of each trial (3 in total) then taking its average.

Test accuracy (avg)	Dataset 1	Dataset 2	Dataset 3	Total avg
SVM	0.9773	0.6536	0.9004	0.8438
KNN	0.9728	0.5592	0.89	0.8073
Decision tree	0.9778	0.7044	0.8916	0.8579

Table 2.2 This table displays the performance of test accuracy of a classifier on three datasets and its overall performance. It is calculated by sum up for three partitions and then take its average.

Validation accuracy	Dataset 1 star			Dataset 2 cardio			Dataset 3 bank		
	80/20	20/80	50/50	80/20	20/80	50/50	80/20	20/80	50/50
SVM	00.9791	0.9758	0.9786	0.6755	0.6276	0.6577	0.8872	0.8956	0.8914

KNN	0.9739	0.9764	0.9746	0.5679	0.592	0.5653	0.8844	0.8875	0.8883
Decision tree	0.9772	0.9781	0.9744	0.7299	0.7170	0.7103	0.8883	0.8938	0.8924

Table 3.1 This table displays the average validation accuracy for different classifiers, different dataset and different partitions. It is calculated by summing up the validation accuracy of each trial (3 in total) then taking its average.

Validation accuracy(avg)	Dataset 1	Dataset 2	Dataset 3	Total avg
SVM	0.9778	0.6536	0.8914	0.8409
KNN	0.9750	0.5750	0.8867	0.8122
Decision tree	0.9766	0.7190	0.8915	0.8624

Table 3.2 This table displays the performance of validation accuracy of a classifier on three datasets and its overall performance. It is calculated by sum up for three partitions and then take its average.

4. **Conclusions**

This paper proves the hypothesis that as dataset size increases, the more accurate the classifiers will be. The result also provided a similar result from the paper An Empirical Comparison of Supervised Learning Algorithms by Rich and Alexandru, which the decision tree classifier has the higher training, testing, validation accuracies than SVM and SVM has higher training, testing, validation accuracy than KNN. According to three tables, overall, the decision tree method has higher testing, training and validation accuracy than two other classifiers. The dataset 2 has the lowest training, testing and validation accuracy for all three classifiers as its dataset size is the smallest. It could also because there is no huge relationship between the feature extracted and the label predicted. Meanwhile, almost all the 80% training 20% testing has

higher accuracy than 50% training 50% and 20% training 80% testing, which might imply that as the ratio of training versus testing increases, the accuracy rate would increase as well.

Reference

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning - ICML 06*. doi: 10.1145/1143844.1143865

Cruz, J. A., & Wishart, D. S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics*, *2*, 117693510600200. doi: 10.1177/117693510600200030

Jordan, M. I., & Mitchell, T. M. (2015, July 17). Machine learning: Trends, perspectives, and prospects. Retrieved from https://science.sciencemag.org/content/349/6245/255.full

Loog, M. (2007). Supervised Classification: Quite a Brief Overview. *Machine Learning Techniques for Space Weather*, 113–145. doi: 10.1016/b978-0-12-811788-0.00005-6

Nguyen, T. T. T., & Armitage, G. (2008). Clustering to Assist Supervised Machine Learning for Real-Time IP Traffic Classification. 2008 IEEE International Conference on Communications. doi: 10.1109/icc.2008.1095

Noble, W. What is a support vector machine?. *Nat Biotechnol* 24, 1565–1567 (2006). https://doi.org/10.1038/nbt1206-1565

Phung, D., Tseng, V. S., Webb, G. I., Ho, B., Ganji, M., & Rashidi, L. (2018). *Advances in Knowledge Discovery and Data Mining*. Cham: Springer International Publishing.

Rokach, L., Maimon, O. Z., Rokach, L., & Maimon, O. (2010). Data mining with decision trees: theroy and applications. Singapore: World Scientific.

Shmilovici A. (2005) Support Vector Machines. In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA

Schrider, D. R., & Kern, A. D. (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. Trends in Genetics, 34(4), 301–312. doi: 10.1016/j.tig.2017.12.005