Support Vector Machine Algorithms for the DNA Dataset

Gabrielle Cantor
Indiana University - Bloomington
Intelligent Systems Engineering
Bloomington, Indiana 47406
gcantor@iu.edu

ABSTRACT

This paper provides an in depth analysis of the experimentation done on the DNA Dataset, which is a dataset that divides 3,186 splice junctions into three classes which determine the boundaries between exons and introns. Using an SVM, or Support Machine Vector, Algorithm, the data is run through a REST service and classified. The Algorithm uses the partition factor, the γ value, and the C value to determine the prediction accuracy. The Algorithm can also be altered by which kernel is used. Using a Radial Basis Function Kernel with a partition factor of 0.95, a γ value of 0.015625, and a C value of 100, we were able to achieve the maximum prediction accuracy of 0.98.

KEYWORDS

hid-sp18-202, DNA Dataset, Support Vector Machine Algorithm, Linear Kernel, Radial Basis Function Kernel, Sigmoid Kernel

1 INTRODUCTION

This paper focuses on the analysis of the DNA dataset [2], which is a collection of 3,186 data points, or splice junctions, divided between 180 features, which in this case are binary indicator variables. This data is divided into three classes when analyzed, which relate to the boundaries between exons (the parts of a DNA sequence that remain after splicing) and introns (the parts of a DNA sequence that are spliced out). This dataset was ran through a REST service using SVM, or a Support Vector Machine, which produces the prediction accuracy of the REST service.

2 SVM ALGORITHM

This experiment uses an SVM [4], or Support Vector Machine, algorithm to analyze the dataset. SVM is a supervised learning module, meaning that the dataset is split into a training set and a testing set, and the algorithm learns how to process the data using the training set before applying it to the test set. One of the primary uses of SVM is classification, which in this example is used to classify the data into three groups, based on how the data relates to the boundaries between the exons and the introns. The exons are the parts of the DNA sequence that remain after it is spliced, and the introns are the parts of a DNA sequence that are spliced out.

Within the SVM algorithm, several factors could be changed which would in turn vary the prediction accuracy. The partition factor could be easily modified. This determines how much of the data is used for training, and how much is used for testing. In the majority of cases, the more data included in the training set the higher the prediction accuracy. This experiment varied the partition factor from 0.5 to 0.99, meaning that anywhere from 50% to 99% of the data is being used to train the algorithm, while the remaining data is being used for testing.

In addition, the kernel type can be modified to determine how the algorithm classifies the data. The three kernels used in this experiment include Linear, Radial Basis Function, and Sigmoid [1]. The Linear kernel operates on a linear model, where

$$K(X, Y) = X^T Y$$

The Radial Basis Function kernel sums the normal curves around data points so that the decision boundary can be defined by a type of topology condition, for example curves where the sum is greater then a certain value. This kernel uses the equation

$$K(X, Y) = exp(\frac{\|X - Y\|^2}{2\sigma^2})$$

The Sigmoid kernel is similar to the sigmoid function in the logistic regression model, which uses a logistic function to define curves based on where the logistic value is greater than a modeling probability value. This kernel is based on the equation

$$K(X, Y) = tan(\gamma * X^T Y + r)$$

The Linear, Radial Basis Function, and Sigmoid kernels each use a different equation and form to vary how the SVM algorithm clusters the data. In addition to pulling data from the datasets, each kernel also uses a γ value, which is the kernel coefficient, and a C value which is the penalty parameter of the error term [4]. We used two different γ values and two different C values, based off of the information provided in the study by Hsu et. al. which compared methods for multiclass SVMs [3].

3 EXPERIMENTATION

In order to determine which combination of variables would result in the highest possible prediction accuracy, we ran a series of experiments varying the kernel, partition value, C value, and γ value.

Linear Kernel

Partition	γ	C	Prediction
0.5	.0001	2	91.0
0.5	.0001	100	91.0
0.5	.015625	2	91.0
0.5	.015625	100	91.0
0.6	.0001	2	91.25
0.6	.0001	100	91.25
0.6	.015625	2	91.25
0.6	.015625	100	91.25
0.7	.0001	2	90.167
0.7	.0001	100	90.167
0.7	.015625	2	90.167
0.7	.015625	100	90.167
0.8	.0001	2	90.5
0.8	.0001	100	90.5
0.8	.015625	2	90.5
0.8	.015625	100	90.5
0.9	.0001	2	90.5
0.9	.0001	100	90.5
0.9	.015625	2	90.5
0.9	.015625	100	90.5
0.95	.0001	2	88.0
0.95	.0001	100	88.0
0.95	.015625	2	88.0
0.95	.015625	100	88.0
0.99	.0001	2	80.0
0.99	.0001	100	80.0
0.99	.015625	2	80.0
0.99	.015625	100	80.0

Radial Basis Function (RBF) Kernel

Partition	γ	C	Prediction
0.5	.0001	2	51.5
0.5	.0001	100	93.5
0.5	.015625	2	93.7
0.5	.015625	100	93.7
0.6	.0001	2	52.0
0.6	.0001	100	94.375
0.6	.015625	2	94.75
0.6	.015625	100	94.875
0.7	.0001	2	51.167
0.7	.0001	100	95.834
0.7	.015625	2	96.667
0.7	.015625	100	95.0
0.8	.0001	2	49.25
0.8	.0001	100	97.25
0.8	.015625	2	97.0
0.8	.015625	100	95.75
0.9	.0001	2	45.5
0.9	.0001	100	96.5
0.9	.015625	2	97.0
0.9	.015625	100	97.0
0.95	.0001	2	42.0
0.95	.0001	100	97.0
0.95	.015625	2	98.0
0.95	.015625	100	98.0
0.99	.0001	2	40.0
0.99	.0001	100	95.0
0.99	.015625	2	90.0
0.99	.015625	100	95.0

A unique factor that appears when the algorithm uses a linear kernel is that the γ and C values do not matter, only the partition value. This is seen in that the prediction accuracy is the same within each partition grouping.The reason for this is that the linear equation doesn't involve the γ or the C value, only the partition value. As the prediction accuracy doesn't take the γ or C values into consideration, the linear kernel will produce a lower and less accurate prediction as it is a generalized result.

In many cases, a linear kernel will not be used beyond preliminary estimations as it is a generalized formula, and is not designed to classify complex datasets.

The Radial Basis Function kernel saw the largest range of prediction accuracy values, from 40.0 to 98.0. The wide range of values within each partition value as well highlight the importance of the γ values and the C values in the Radial Basis Function kernel. The lowest prediction accuracy within each partition value occurred when the γ value was 0.0001 and the C value was 2, which indicates that this particular γ and C value combination are not well suited for the analysis done when a Radial Basis Function kernel is used.

Sigmoid Kernel

Partition	γ	C	Prediction
0.5	.0001	2	51.5
0.5	.0001	100	93.0
0.5	.015625	2	92.8
0.5	.015625	100	87.9
0.6	.0001	2	52.0
0.6	.0001	100	91.25
0.6	.015625	2	93.625
0.6	.015625	100	89.625
0.7	.0001	2	51.167
0.7	.0001	100	94.834
0.7	.015625	2	95.167
0.7	.015625	100	90.5
0.8	.0001	2	49.25
0.8	.0001	100	95.25
0.8	.015625	2	95.0
0.8	.015625	100	95.25
0.9	.0001	2	45.5
0.9	.0001	100	95.5
0.9	.015625	2	90.0
0.9	.015625	100	94.5
0.95	.0001	2	42.0
0.95	.0001	100	96.0
0.95	.015625	2	90.0
0.95	.015625	100	97.0
0.99	.0001	2	40.0
0.99	.0001	100	95.0
0.99	.015625	2	75.0
0.99	.015625	100	95.0

The Sigmoid kernel saw a large range of prediction accuracy values, from 40.0 to 97.0. The wide range of values within each partition value as well highlight the importance of the γ values and the C values in the Sigmoid kernel. The lowest prediction accuracy within each partition value occurred when the γ value was 0.0001 and the C value was 2, which indicates that this particular γ and C value combination are not well suited for the analysis done when a Sigmoid kernel is used.

4 ANALYSIS

After conducting all experiments, the highest prediction accuracy we were able to achieve was 98.0, which was achieved using a Radial Basis Function kernel, a partition of 0.95, a γ value of 0.015625, and a C value of 100. This was an interesting finding, as it was done using not the highest partition value as would be expected. This result also mixed γ and C values, as the original experiments were done using set pairs, where a γ value of 0.0001 was paired with a C value of 2, and a γ value of 0.015625 was paired with a C value of 2.

The lowest prediction accuracy we achieved was 40.0, which was achieved using a Radial Basis Function kernel, a partition of 0.99, a γ value of 0.0001, and a C value of 2. This result was one of the more interesting ones as with a partition of 0.99, the algorithm is training on 99% of the data before testing on the remaining 1%. In this situation, one would expect the algorithm to achieve a high prediction accuracy, as it has encountered almost all of the data in training before being tested.

Based off of the knowledge that when analyzing the DNA dataset with this SVM algorithm the lowest prediction accuracy comes from a partition value of 0.99 while the highest prediction accuracy comes from a partition value of 0.95, we can infer that the SVM algorithm is most accurate for the DNA dataset when it is trained on less than 99% of the data. This differs from many datasets, where having a higher partition factor leads to a higher prediction accuracy. The most likely explanation for this is that the DNA dataset likely contains several outlying data values which would not be predicted by the algorithm, and if those outlier are included in the data that the algorithm is tested on it would be less likely to result in an accurate prediction.

5 CONCLUSION

Using the DNA Dataset, we worked to determine which variation of the SVM Algorithm would provide the highest prediction accuracy. By varying the kernel, partition value, γ value, and C value we were able to achieve prediction accuracy's ranging from 40.0 (Radial Basis Function and Sigmoid kernel, 0.99 partition value, γ value of 0.0001, and a C value of 2) to 98.0 (Radial Basis Function kernel, 0.95 partition value, γ value of 0.015625, and a C value of 100.

From these experiments we are able to conclude that the DNA dataset likely contains several outlying data points, which would explain why the prediction accuracy is higher at a 0.95 partition then at a 0.99 partition value. This is valuable information, as it can impact future work with the DNA dataset. When working with this in the future, researchers may consider sifting through the data to remove outliers, or incorporate a degree of error into their findings to account for these outliers.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski, Dr. Geoffery Fox, Tyler Balson, Vibhatha Abeykoon, and Miao Jiang for their support and suggestions to write this paper.

REFERENCES

[1] [n. d.]. The difference of kernels in SVM? ([n. d.]). https://stats.stackexchange.com/questions/90736/the-difference-of-kernels-in-svm

[2] [n. d.]. LIBSVM Data: Classification (Multi-class). ([n. d.]). https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html

[3] Chih-Wei Hsu and Chih-Jen Lin. 2002. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks* 13, 2 (Mar 2002), 415–425. https://doi.org/10.1109/72.991427

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.