Homestead High School

Improving Neural Network Generalization Ability Using Outlier Analysis & Voronoi Tessellation

M. Ho[1], D.M. McIntosh[2], A.N. Srivastava[2]
[1]NASA Summer High School Apprenticeship Research Program, Moffett Field, CA, USA
[2]NASA Ames Research Center, Moffett Field, CA, USA

Introduction

Artificial neural networks (ANNs) serve to process, learn, and predict information using layers of interconnected computational units resembling the neurons that comprise the biological nervous system of the human brain. The main contrast between ANNs and the human brain lies in the ability of the latter to better adapt to experience, apprehending new knowledge and accommodating to new perspectives, regardless of the unfamiliarity of the environment from which it learns. Although computers do crunch numbers and perform computations on command with much greater speed, the unparalleled capacity of the brain's neural structure to instinctively re-organize according to novel information enables the brain to perform considerably complex operations more efficiently than a computer would, since machines are programmed only to memorize and carry out the specific *set of instructions* humans provide to them, thereby limiting their problem-solving potential to problems we already understand ourselves. To overcome this impediment, ANNs are designed to perform well on new information in order to be flexibly applicable to and functional on new problems. The quality of the network's performance on such cases depends on its generalization ability, or the ability to recognize trends from the training data and employ what it has learned to make predictions on new test data -- much like how the human brain learns from an initial experience.

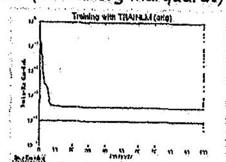
Nonetheless, ANNs often perform poorly when applied to new cases dissimilar to those they have encountered, a flaw possibly attributed to data anomalies that adversely affect the training process. The primary goal of training is to fit the network to the "signal" of the training data such that an adequate generalization of the regularities appearing in the training set can be extracted; it follows that the network may inadvertently memorize the training data with the aim of generating a low training error if the training set contains outlier points that cannot be generalized, consequently skewing the emulated trend. The fallacy in adapting to idiosyncrasies so precisely, however, is that the network will not be flexible enough to generalize new cases arising from a test set that will most certainly deviate from the training set. Further, if the training data is not sufficiently representative of the test data, or if the test data itself contains outliers, a network that has overfitted the training examples will fail to adapt to and process the novel data satisfactorily. Thus, it is important to develop methods to *improve a neural network's generalization* ability, since the quality of future predictions on a comprehensive set of all possible data -- not only the data on which the network has already trained -- is the ultimate (and more realistic) determinant of a network's proficiency.

In this study, in order to ascertain whether the network's performance error was distributed throughout the data or concentrated in a single area, we isolate inputs and distinguish combinations of independent inputs that most significantly contribute to the overall error. *We find* that the isolation of inputs u and g and the combination of inputs u and r produce considerably low errors as compared to the error generated by the inclusion of all nine inputs. We also investigate the *consequences of training* a network with a data set containing outliers and the effects of eliminating outliers on the error value generated by the testing process. *We find* that removing anomalies from the training set actually results in a greater error, since the network has no means by which to develop a good response to untypical data when given a limited training set that is unrepresentative of a test set containing outliers. *Moreover*, we examine the relationship between the size of a voronoi cell and the magnitude of the disparity between the correct output and predicted output of outliers. Given that a substantial correlation exists, we can utilize voronoi tessellation to "warn" a network to detect outliers based on the high volume and low distribution of their voronoi cells and to train carefully on such points such that the true generalization of the data is maintained.

Methodology

Error Back-Propagation Training Algorithm (Levenberg-Marquardt)

- *The network begins with arbitrarily weighted parameters*
- A vector of a galaxy property input is fed into the network, which performs calculations to project an estimated value for the distance to the galaxy as the output
- The output vector is compared against the measured output distance; the mean squared difference between the two is noted as the error
- Using the gradient descent method, the network adjusts its parameter weights according to the direction of the steepest decreasing error
- After all galaxy property inputs are iterated through the network, the parameters will have adjusted in such a way that the network's outputted distance predictions become increasingly accurate of the correct distances and the performance error is minimized (see figure above)



Network training toward low error through gradient descent approach

Input Selection

- To determine where the network is making the mistakes, isolate each property input (i.e. u, g, r, i, z, petro50, petro90, qr, and fracdev) and run them separately through the network
- Input properties generating errors closest to the error produced by the inclusion of all nine inputs serve as the most significant contributors to the overall error
- the training algorithm is applied to all possible combinations of two property inputs, as opposed to individual inputs, *in order to better approach* the network's actual training potential,

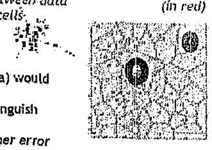
Outlier Removal

- Eliminate outliers residing far from the rest of the data as well as the test data
- Idea: If the training data is reduced to a more general representation of the test set, the network will more easily recognize patterns in the data *in order to extract a trend applicable to the test data*

Similar points reside in high density area (in blue)
Outlier points reside in low density area (in red)

Voronoi Tessellation

Perpendicular bisectors (in red) between data points (in blue) designate voronoi cells



- Voronoi cells define the space around data point closer to that point than to any other plotted point
- An outlier point far from the rest of the data set (in a low density area) would reside in a large cell
- If large voronoi cells correspond to poor network fittings, we can distinguish outliers by voronoi cell volume
- The network can then train to accommodate the likelihood of the higher error caused by an outlier

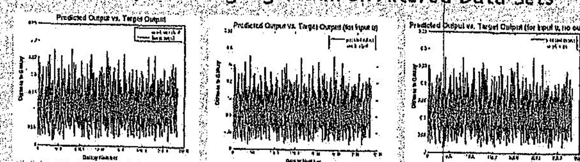
Data

- The data used in this study was obtained from the Sloan Digital Sky Survey (SDSS), which provides astronomers with what is currently the most extensive mapping of the universe, covering 25% of the sky and cataloging the spectral properties (e.g. luminosity, color, surface temperature) of over 100 million celestial objects
- Images generated by the SDSS are collected through five filters named u, g, r, i, and z that have respective wavelengths of 3540, 4760, 6222, 7632, and 9049 Å.
- By measuring the photometric redshifts of the aforementioned wavelengths of a galaxy, astronomers can ascertain the extent to which the galaxy is receding from which the distance to the galaxy can be calculated
- Data collected for a small select group of galaxies (approximately 30,000) contains accurate measurements of the galaxies' redshifts, in addition to measurements of their spectral properties
- The above data set containing both redshift measurements as well as spectral properties of the selected galaxies served as the training set for the purposes of this study; the data set containing only the spectral properties of a separate group of galaxies served as the test set

Fig. 1 The redshifts of the galaxies photographed above indicate that other objects with similar red-shifts are located within their vicinity (www.sdss.org).

Figures

Application of Training Algorithm on Altered Data Sets



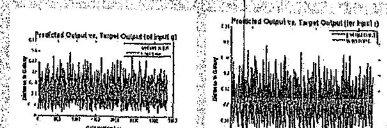


Fig. 2 Network performance on all property inputs, input u, input u without outliers, input g, and input r.

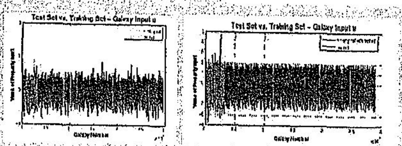


Fig. 3 Example of training set before and after outlier elimination. Data points non-representative of the test data and the majority of the training data were removed; consequently, the extremities existing in the test set are not encompassed by the reduced training set.

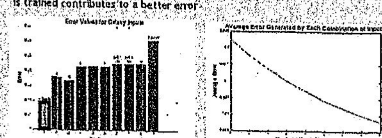


Fig. 4 Comparison of overall error to errors produced by individual inputs (left), as well as to average errors produced by each possible combination of inputs (right). Results indicate that the input isolation of u and g yield acceptably low errors and that each additional property on which the network is trained contributes to a better error.

[PARALLEL COORDINATES]

Fig. 5 Independent inputs form a curved surface, while dependent inputs form a spherical surface. Results verify the independent relationship between inputs u and r.

Voronoi Tessellation of Error Generated by Inputs u & r

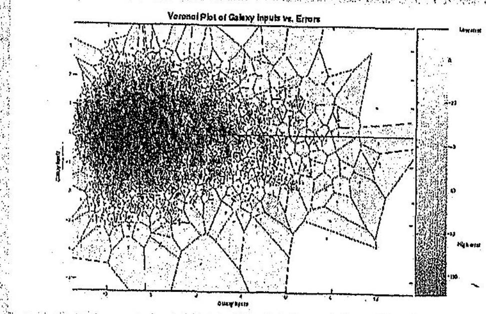
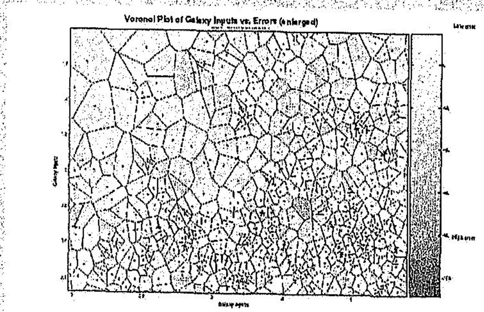


Fig. 6 Many outliers residing in large voronoi cells exhibit surprisingly low errors, while several specific voronoi cells distributed throughout regions of high density exhibit particularly high errors and distinguish the areas of the training data in which the network was making mistakes. Note that the voronoi cells corresponding to high errors appear to be randomly dispersed throughout the tessellation.



Conclusions

Although the network's best performance resulted from the inputting of all galaxy properties, the combination of galaxy property inputs u and r produced a notably low error that was considerably close to the overall error.

The network generated a lower error after training on data containing outliers, since the wide-ranging test set encompassed data anomalies to which the network's learned trend would not have otherwise been applicable. By training the network on several outliers, the network grew more robust and improved its recognition of patterns appearing in an intricate training set that was difficult to generalize, thus precluding the risk of overfitting that would have arose from training on an overly general or condensed data set.

Voronoi tessellation reveals that outliers with high voronoi cell volumes are associated with low- to mid-range errors, which reinforces the notion that the network had already adjusted itself to place less importance on outermost points when generalizing the overall trend of the data. Detecting outliers by voronoi volume therefore proves to be an effective method for the network to further improve generalization ability.

For future research, the impact of excluding unusually high or low errors as the network learns the correlation between performance errors and voronoi cells of outliers should be studied, since the voronoi volumes corresponding to extreme error values may impact the training process in the same way outlying data points do.

References

[illegible]
[illegible]
[illegible]
[illegible]
[illegible]