Technical Disclosure Commons

Defensive Publications Series

February 26, 2018

A methodology for automated feature engineering in training of shallow neural networks for prediction of multi-linear growth failures ("Nova Algorithm")

Dermot Cochran
Hewlett Packard Enterprise

Martin Roe
Hewlett Packard Enterprise

CV Goudar
Hewlett Packard Enterprise

Breda Martyn
Hewlett Packard Enterprise

Mainak Das
Hewlett Packard Enterprise

See next page for additional authors

Inventor(s)
Dermot Cochran, Martin Roe, CV Goudar, Breda Martyn, Mainak Das, Pratush Uppuluri, Sean Lennon, Basudev Panda, Sarah King, Liam Dolan, and James Marshall Hughes

A methodology for automated feature engineering in training of shallow neural networks for prediction of multi-linear growth failures ("Nova Algorithm")

Prediction of outages due to multi-linear growth patterns using a binary classification neural net with automated pre-processing to derive training set, and post-processing to achieve ensemble predictions.

1. Problem(s) solved

Data center outages can be prevented or resolved with careful analysis of telemetry logs by experienced engineers with domain specific knowledge about server, storage and networking products. This process of prediction and prevention can be automated using available data sources. However, these data sources were too sparse to use deep learning for prediction of outages. There is a need for engineering of the feature vectors to allow the usage of a shallow neural net.

This invention disclosure describes how to make predictions using smaller data sets, allowing the use of shallow neural networks instead of deep learning.

This methodology for prediction and prevention of data center outages was applied to prediction of Common Provisioning Group (CPG) growth failures in 3PAR storage devices, using a trained neural net model with additional derived attributes, pre-processing rules and post-processing steps, applied to 90 days history of logical disk free space for each CPG in the storage device.

CPGs are collections of logical disk volumes, which can be dynamically allocated. There are hidden patterns in the way that logical disk free space is allocated and freed, which can lead to growth failures, causing an outage. These patterns are non-linear, and cannot be found using only a time series but can be found by training a neural net, in addition to time series analysis (e.g. ARIMA).

CPG growth failure outages can be predicted between 2-4 months in advance using a feature vector containing the most recent 91 days of normalized logical disk free space (reservoir level) plus 8 derived attributes, and a shallow neural network consisting of one "hidden" (inner) layer with 62 nodes and 500 iterations.

A decision tree methodology is needed to automate the process of feature engineering, that is, to identify the optimal subset of derived attributes, discover pre-filtering rules and discover the optimal thresholds for post-processing rules.

2. Prior Solutions

There are a number of alternative approaches for the combination of decision trees with shallow neural networks, but none use a decision tree to create and refine the training set for the neural net.

Sethi, Ishwar. (1990). Entropy nets: From decision trees to neural networks. Proceedings of the IEEE. 78. 1605 - 1613. 10.1109/5.58346.

Disadvantages of entropy nets: Entropy nets are the conversion of a decision tree into a neural net. This does not meet our requirement for automated feature engineering.

Soltan, A & Mohammadi, M. (2012). A hybrid model using decision tree and neural network for credit scoring problem. Management Science Letters, 2(5), 1683-1688.

The Soltan-Mohammadi model applies to decision tree analysis to the output of the neural net, rather than as a pre-processing layer. It does not solve our need for automated feature engineering.

Hinton's capsule methodology (https://research.google.com/pubs/pub46351.html) would provide a viable alternative but the details are different; our solution uses a small ensemble set of neural network models (see Figure 1) instead of a complex nested hierarchy of neural net classifiers.

Our approach has similarities to the Curriculum Learning Strategy (http://ronan.collobert.com/pub/matos/2009_curriculum_icml.pdf) but with the additional pre-processing and post-processing steps e.g. C5.0 decision tree and GARCH model for the time series.

3. Description

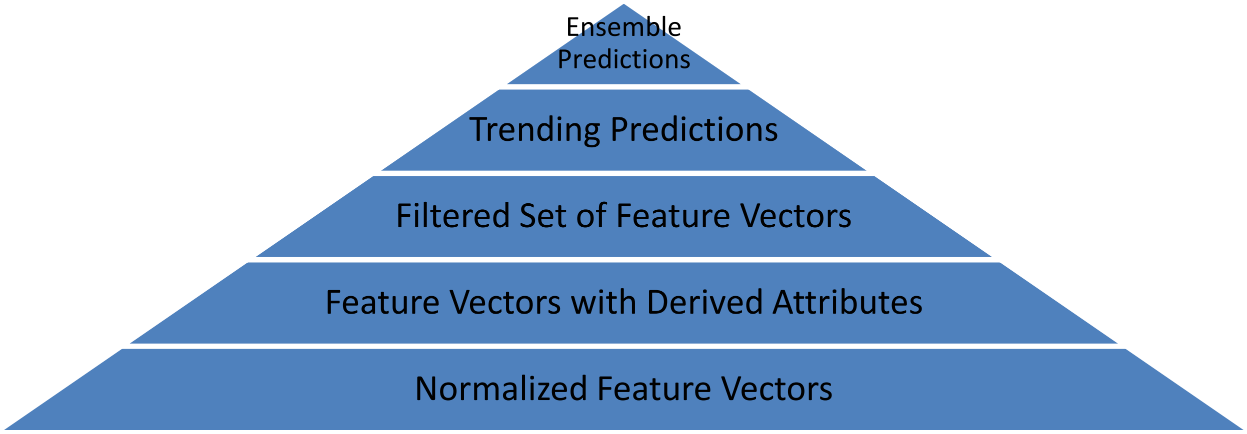

Figure 1 - Refinement of Feature Vectors and Predictions

Step	**Name**	**Explanation**	**Example**	**Fully or Partly Implemented in Pilot?**
1	Historical data collection	For the resource in question we need 2-3 years of historical data with samples taken at regular intervals	Daily logical disk free space level for each CPG on a storage device	Full
2	Prediction window	Assemble base feature vectors for a defined period	91 days of logical disk free space	Full

3	Normalization	(a) Add an attribute for the maximum capacity of the reservoir; convert this attribute to a normalization scale (b) Normalize the other data points relative to the maximum capacity	(a) A logarithmic scale for free space capacity (b) Percentage free space over last 91 days	Partly – maximum capacity attribute was excluded from pilot

4	Derivation of Additional Attributes	(a) Calculate maximum decline over 3,5,7 days to size of the time window (b) Calculate other derived attributes	The first class of candidate attributes are average slope over the last 7, 15, 30 and 45 days Some other possible derived attributes are (i) absolute (or scaled) value of free space on last day, (ii) standard deviation of daily percentage free space, (iii) highest daily decline, (iv) highest daily increase in % free space, (v) number of days with a decrease in space., (vi) number of days with an increase in space and (vii) number of days with a change of direction from an increase to a decrease or the reverse	Partly – only a limited subset of derived attributes were used

5	Selection of best derived attributes	Select the derived attribute that led to most reduction in false positives and add to feature vector, then find the next most significant derived attribute, until the improvement is marginal The neural net is trained with different permutations of derived attributes and we select the subset which yields the best results; thus the derived attributes are not static; they can change after re-training with new data	The top 8 derived attributes were: Derived Feature = Standard deviation of daily percentage change Derived Feature = Maximum contiguous decline of percentage of free space over used space Derived Feature = Maximum contiguous decline of percentage of free space over used space divided by the value of the percentage of free space over used space on the last day. Derived Feature = LD Space remaining on last day divided by the growth increment. This data was scaled using the formula (Xi-mean/standard deviation). Derived Feature = Negative Change Count Derived Feature = Differential Sign Change Count Derived Feature = Positive Change Count	Partly – this step was hard-coded in pilot

			Derived Feature = Average slope over the last 15 days Derived Feature = Maximum negative change over one day	
6	Derivation of pre-filtering rules	Decision tree analysis (e.g. C5.0 decision tree) is used to derive pre-processing rules based on the extended feature vector	If attribute-A less than threshold-A and attribute-B greater than threshold-B, then exclude that feature vector from the training set	Partly – this step was hard-coded in pilot

7	Selection of best pre-filtering rules	We select the pre-processing rules that yield most improvement in precision and recall The new engineered training set includes feature vectors with derived attributes and filtered by pre-processing rules that will change if we re-train using new data	Rule 1 is defined as: (Day 1 > 77.17) & (Variability <= 0.5004117) Rule 2 is defined as: (Last Day > 85.09667) & (negative change count <=61) & (max positive <= 10.67667) & (differential sign change count <=16) & (last 15 day slope variance <= 0.2052487) Rule 3 is defined as: (variability <= 6.722171) & (mean slope of last 15 days <= 0.5142223) & (negative change count <= 63) & (differential sign change count<= 1) & (last 14 Days average > 60) & (standard deviation <= 3) Rule 4 is defined as:(variability <= 1.458835) & (negative change count <= 61) & (differential sign change count <= 15) & (last 14 Days average > 55) & (max positive > 0.1) Rule 5 is defined as: (Day 1 > 73.48333) & (coefficient of variation <= 6.158184) & (mean	Partly – this step was hard-coded in pilot

			slope of last 15 days <= 0.5546666) & (mean slope of last 15 days > 0.0) Rule 7 is defined as: (maximum decline over percentage left <= 0.3254047) & (average > 66.14458) & (coefficient of variation <=5) & (negative change count > 11) & (negative change count <=28) Rule 8 is defined as: (avg > 65.67363) & (lastnmeanslope15 <= 0.5546666) & (maxneg > -1.76) & (maxpos > 0.163334)	

8	Derivation of ensemble batch size (see Figure 2).	*(a)* Combine successive predictions over a defined period, *or alternatively train several different neural network models with the same data to reduce noise and over-fitting* (b) Choose the minimum number of trending predictions needed	(a) A wait period of 10 days (b) A minimum of 10 trending predictions with the last 10 days A wait period and voting threshold is calculated e.g. 9 out of 10 days would mean that the same prediction result must be achieved at least 9 times out of the last 10 days - this creates an ensemble prediction based on 10 feature vectors, staggered one day apart. Based on a defined training set we select the voting threshold and wait period that yields the best precision and recall, for an arbitrary neural net e.g. 56 nodes and 1 hidden layer.	Partly – this step was hard-coded in pilot; only one neural net model was trained, but we used trending predictions over a range of 10 days to reduce noise and over-fitting.
9	Derivation of quiet period	Exclude any trending predictions with a certain time frame of the last ensemble prediction	A quiet period of 15 days between prediction notifications	Partly – this step was hard-coded in pilot

10	Inclusion of secondary model	A GARCH or ARIMA time series model as a secondary filter to reduce false positives	Primary predictions suppressed based on confidence level and history of false positives	Not implemented in pilot

.

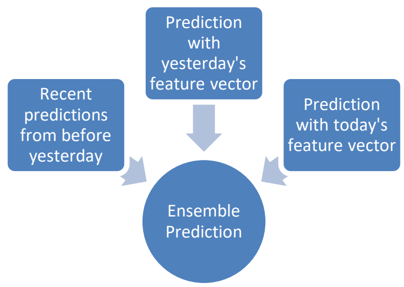

Figure 2 - Ensemble predictions

4. Advantages

The methodology provides a new method for prediction of resource consumption growth outages in data centers. It can be generalized to forecasting capacity for any resource that rises or falls (in a multi-linear pattern) on a daily basis. It also provides a method for automated feature engineering when training shallow neural networks using sparse data.

This is superior to simple capacity forecasting because we are learning from the data, and more versatile than deep learning because we can also learn from sparse data sets.

The unexpected benefit is that this solution could be applied to any reservoir like resource that rises and falls on a daily basis e.g. number of available parking slots.

5. Title

Automated feature engineering for shallow neural networks

6. Abstract

Deep learning requires a large data set. When working with smaller data sets we need an automated approach for feature learning as a pre-processing step to create the training set for a shallow neural network with one hidden layer. The outcome of the neural net classifier is filtered using post-processing rules over a defined wait and quiet period to create an ensemble

prediction with higher quality. This gives us the benefits of deep learning, but with a smaller data set.

A partial implementation of this methodology was used for a pilot release for prediction of Common Previsioning Group (CPG) growth failures in 3PAR storage devices, using a trained neural net model with additional derived attributes, pre-processing rules and post-processing steps, applied to 90 days history of logical disk free space for each CPG on the storage device.

Authors:

Dermot Cochran - Hewlett Packard Enterprise
Martin Roe - Hewlett Packard Enterprise
CV Goudar - Hewlett Packard Enterprise
Breda Martyn - Hewlett Packard Enterprise
Mainak Das - Hewlett Packard Enterprise
Pratush Uppuluri - Hewlett Packard Enterprise
Sean Lennon - Hewlett Packard Enterprise
Basudev Panda - Hewlett Packard Enterprise
Sarah King - Hewlett Packard Enterprise
Liam Dolan - Hewlett Packard Enterprise
James Marshall Hughes - Hewlett Packard Enterprise