Predicting Trends in Retinol and Beta-Carotene Plasma Levels Using Neural Networks

Khosrow Kaikhah, Ph.D.
Department of Computer Science
Texas State University
San Marcos, Texas 78666
kk02@TxState.edu

Abstract: — A novel knowledge discovery and prediction technique using neural networks is presented. A neural network is trained to learn the correlations among physical and dietary characteristics of several hundred people to their Retinol and Beta-Carotene Plasma Levels. The neural network is then pruned and modified to generalize the correlations and relationships existing in data. Finally, the neural network is used as a tool to discover and predict the hidden trends inherent in dataset.

Keywords: adaptive clustering, knowledge discovery, prediction, neural networks, pruning, training, retinol and beta-carotene plasma levels

1 Introduction

Large datasets usually contain hidden trends, which convey valuable knowledge about the dataset. The acquired knowledge is helpful in understanding the domain, which the data describes. The hidden trends, which can be expressed as rules or correlations, highlight the associations that exist in the data. Therefore, discovering these hidden trends, which are specific to the application, is extremely helpful and vital for analyzing the data.

We define a machine learning process that uses artificial neural networks to discover and predict trends in large datasets. A neural network is first trained to learn the inherent relationships among the data. The neural network is then modified via pruning and hidden layer activation clustering. Finally, the modified neural network is used as a tool to discover and predict hidden trends in the dataset. The extraction phase can be regulated through several control parameters, which determine the granularity of the trends.

2 Predicting Trends

We have developed a novel process, using neural networks, for predicting trends in large datasets, with *m* dimensional input space and *n* dimensional output space. Our process is independent of the application. The significance of our approach lies in using neural networks for discovering the hidden trends, with control parameters. The control parameters influence the discovery process in terms of importance and significance of the acquired trends. There are three phases in our approach: 1) neural network training, 2) pruning and clustering, and 3) trend discovery and prediction.

In phase one, the neural network is trained using a supervised learning method. The neural network learns the associations inherent in the dataset. In phase two, the neural network is pruned by removing all unnecessary connections and neurons, and the activation values of the hidden layer neurons are clustered using an adaptable clustering technique. In phase three, the hidden trends are discovered by using the recall mechanism of the modified neural network. These three phases are described in more detail in the next three sections.

2.1 Neural Network Training

Neural networks are able to solve highly complex problems due to the non-linear processing capabilities of their neurons. In addition, the inherent modularity of the neural networks structures makes them

adaptable to a wide range of applications. The neural network adjusts its parameters to accurately model the distribution of the provided dataset. Therefore, exploring the use of neural networks for discovering correlations and trends in data is prudent.

The input and output patterns may be real-valued or binary-valued. If the patterns are real-valued, each value is discretized and represented as a sequence of binary values, where each binary value represents a range of real values. For example, the age of a person can be discretized into 6 different intervals: (-20], (20-30],(30-40],(40,50],(50,65], and (65+]. Therefore, [0 0 0 1 0 0] would represent an age between 41 and 50. The number of neurons in the input and output layers are determined by the application, while the number of neurons in the hidden layer are dependent on the number of neurons in the input and output layers.

We use a gradient descent approach to train and update the connection strengths of the neural network. The gradient descent approach is an intelligent search for the global minima of the energy function. We use an energy function (Eq. 1), which is a combination of an error function and a penalty function.

$$\theta(w,v) = E(w,v) + P(w,v) \quad (1)$$

$$E(w,v) = \frac{1}{L}\sum_{l=i}^{L}\sum_{k=1}^{n}(o_{lk} - d_{lk})^2 \quad (2)$$

$$P(w,v) = \rho_{decay}\left(P_1(w,v) + P_2(w,v)\right) \quad (3)$$

$$P_1(w,v) = \varepsilon_1\left(\sum_{j=1}^{h}\sum_{i=1}^{m}\frac{\beta w_{ij}^2}{1+\beta w_{ij}^2} + \sum_{j=1}^{h}\sum_{k=1}^{n}\frac{\beta v_{jk}^2}{1+\beta v_{jk}^2}\right)$$

$$P_2(w,v) = \varepsilon_2\left(\sum_{j=1}^{h}\sum_{i=1}^{m} w_{ij}^2 + \sum_{j=1}^{h}\sum_{k=1}^{n} v_{jk}^2\right)$$

The error function (Eq. 2) computes the error of each neuron in the output layer, and the penalty function (Eq. 3) drives the connection strengths of unnecessary connections to very small values while strengthening the rest of the connections. The network is trained till it reaches a recall accuracy of 99% or higher.

2.2 Pruning and Clustering

Phase two involves two steps: 1) pruning the neural network, and 2) clustering the hidden layer activation values. The penalty function, used during the training phase, drives the strengths of unnecessary connections to approach zero very quickly. The insignificant connections having very small values can safely be removed without considerable impact on the performance of the network. For each input to hidden layer connection(w_{ij}),if $\max_k |v_{jk} w_{ij}| < 0.1$ remove w_{ij}, and for each hidden to output layer connection (v_{jk}), if $|v_{jk}| \le 0.1$ remove v_{jk}.

After removing all weak connections, any input layer neuron having no outgoing connections can be pruned. In addition, any hidden layer neuron having no incoming or outgoing connections can safely be pruned. Finally, any output layer neuron having no incoming connections can be pruned. Removal of input layer neurons corresponds to having irrelevant inputs in the data model; removal of hidden layer neurons reduces the complexity of the network and the clustering step; and removal of the output layer neurons corresponds to having irrelevant outputs in the data model. Pruning the neural network results in a less complex network while improving its generalization.

Once the pruning step is complete, the network is trained with the same dataset in phase one to ensure that the recall accuracy of the network has not diminished significantly. If the recall accuracy of the network drops by more than 2%, the pruned connections and neurons are restored and a stepwise pruning approach is pursued. In the stepwise pruning approach, the weak incoming and outgoing connections of the hidden layer neurons are pruned and the network is re-trained and tested for recall accuracy, one hidden neuron at a time.

After completing the pruning step, the activation values of each hidden layer neuron are dynamically clustered and re-clustered with a cluster radius and confidence radius, respectively. The clustering algorithm is adaptable, that is, the clusters are created dynamically as activation values are added into the clusterspace. Therefore, the number of clusters and the number of activation values in each cluster are not known *a priori*.

For each hidden layer neuron *j*:

1.Calculate its activation value y_j^p, for each input pattern p

2.If $min\left\|G_c - y_j^p\right\| \le r_{cluster}$ for all clusters

y_j^p belongs to G_c

else

y_j^p creates a new cluster

The centroid of a cluster, G_c, represents the mean of the activation values in the cluster and can be used as the representative value of the cluster, while the frequency of each cluster $freq_{Cluster}$ represents the number of activation values in that cluster. By using the centroids of the clusters, each hidden layer neuron has a minimal set of activations. This helps with getting generalized outputs at the output layer. The centroid is adjusted dynamically as new elements y_j^p are added to the cluster.

$$G_c^{new} = \frac{(G_c^{old} \cdot freq_{Cluster}) + y_j^p}{freq_{Cluster} + 1} \quad (4)$$

Since dynamic clustering is order sensitive, once the clusters are dynamically created with a cluster radius that is less than a predetermined upper bound, all elements will be re-clustered with a confidence radius of one-half the cluster radius.

The upper bound for cluster radius defines a range for which the hidden layer neuron activation values can fluctuate without compromising the network performance. For maintaining the accuracy of the network, the following must hold,

$$\left|z_k - z_k^*\right| \le \rho \quad (5)$$

where z_k^* is the desired value of the output layer neuron, and ρ (the tolerance factor) is typically set to a small value less than 1. The upper bound for the cluster radius that maintains the network accuracy is:

$$\left|r_c\right| \le \frac{\ln\left(\frac{1}{\rho} - 1\right)}{\max_k \left|\sum_{j=1}^{m} v_{jk}\right|} \quad (6)$$

where v_{jk} is the instar for each output layer neuron.

The benefits of re-clustering are twofold: 1) Due to order sensitivity of dynamic clustering, some of the activation values may be misclassified. Re-clustering alleviates this deficiency by classifying the activation values in appropriate clusters. 2) Re-clustering with a different radius (confidence radius) eliminates any possible overlaps among clusters. In addition during re-clustering, the frequency of each confidence cluster is calculated, which will be utilized in the trend discovery phase.

2.3 Trend Discovery and Prediction

In the final phase of the process, the knowledge acquired by the trained and modified neural network is discovered in the form of rules. This is done by utilizing the generalization of the hidden layer neuron activation values as well as the control parameters. The novelty of the extraction process is the use of the hidden layer as a filter by performing vigilant tests on the clusters. Clusters identify common regions of activations along with the frequency of such activities. In addition, clusters provide representative values (the mean of the clusters) that can be used to retrieve generalized outputs.

The control parameters for the extraction process include: a) cluster radius, b) confidence frequency, and c) hidden layer activation level. The cluster radius determines the coarseness of the clusters. The confidence radius is usually set to one-half of the cluster radius to remove any possible overlaps among clusters. The confidence frequency defines the minimum acceptable rate of commonality among data patterns. The hidden layer activation level defines the maximum level of tolerance for inactive hidden layer neurons.

Trend discovery and prediction is performed by presenting all permutations of input categories as input patterns to the trained and modified neural network along with the desired control parameters.

For each input pattern p:

1. Calculate the activation value y_j^p for each hidden layer neuron
2. For each hidden layer neuron j (in parallel)

If $(\left\|G_c - y_j^p\right\| \le r_{Conf}$ AND $freq_{cluster} \ge Freq_{Conf})$

$Y_j^{Act\ p} = G_c$

else

$Y_j^{Act\ p} = 0$

3. If the percentage of inactive hidden layer neurons does not exceed the desired hidden layer activation level, propagate the hidden layer values to the output layer.

The input patterns that satisfy the rigorous extraction phase requirements and produce an output pattern represent generalization and correlations that exist in the dataset. The level of generalization and correlation acceptance is regulated by the control parameters. This ensures that inconsistent patterns, which fall outside confidence regions of hidden layer activations, or fall within regions of low activity levels, are not considered. There may be many duplicates in these accepted input-output pairs. In addition, several input-output pairs may have the same input pattern or the same output pattern. Those pairs having the same input patterns will be combined, and, those pairs having the same output patterns will be combined. This post-processing is necessary to determine the minimal set of trends. Any input or output attribute not included in the discovered trend corresponds to irrelevant attributes in the dataset.

3 Predicting Trends in Retinol and Beta-Carotene Plasma Levels

The relationships that exist between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene, and other carotenoids are discovered by the neural network; and significant trends are predicted based on the existing correlations. The training dataset, consisting of 1315 subjects, is obtained from a study conducted by Nierenberg et al. in "Determinants of plasma levels of beta-carotene and retinol", American Journal of Epidemiology. There are twelve categories that describe the dietary and personal characteristics of subjects. These categories are used as inputs to the neural network. In addition, there are two categories of plasma levels which are used as outputs of the neural network. All input/output categories are described in Table 1.

Each category is discretized into several intervals to define the binary input/output patterns. Table 2 represents the discrete intervals for each category. We designed a neural network with 51 input layer neurons, which is the total number of discretized intervals for the twelve input categories, 40 hidden layer neurons, and 12 output layer neurons, which represent the total number of intervals for the two output categories. The neural network is trained to achieve a recall accuracy of 100%. After the training phase, the network was pruned and the hidden layer activation values were clustered. Although about 30% of connections as well as about 3% of hidden layer neurons were pruned, none of the input or output layer neurons were pruned. This reflects the importance of all dietary and physical categories we used for discovering trends in plasma levels. The pruned neural network maintains a recall accuracy rate of 100%.

The network is then used as a tool to predict trends. All possible permutations of input categories are presented to the input layer, one pattern at a time. Any input pattern that passes the rigorous vigilance test at the hidden layer and produces an output pattern constitute a discovered trend. There may be several duplicates in the discovered trends. In addition, several trends may have the same or similar input patterns, or several trends may have the same of similar output patterns. These trends are combined and duplicates are removed to discover the minimal set of trends.

3.1 The Predicted Trends

During the trend prediction phase, six different trends were discovered with the following control parameters: cluster radius set at 0.2 (the upper bound was calculated to be 0.4); the confidence frequency set at 30% of the total population; and the hidden layer activation level set at 5%. These predicted trends are based on the high level commonalty that exists in the training dataset. They represent the existing correlation between levels of Plasma Retinol and Beta-Carotene with respect to the dietary and physical characteristics. The discovered trends are represented in table 3.

4 Conclusions

We were able to discover the existing trends in plasma levels based on the dietary and physical characteristics of over thirteen hundred subjects. These trends represent the hidden knowledge inherent in the dataset and are based on the high level of commonalty that exist in the dataset. The desired level of commonalty can be regulated through the control parameters. Although, about 12% of the subjects were male, the process was not able to discover any trends for male subjects. This is

contributed to the low number of male subjects in the dataset. The medical experts who reviewed the results of our discovery process were impressed by the exactness and expressive power of the discovered trends. They indicated that this knowledge can be used to better monitor and predicate the plasma levels of their patients.

The knowledge discovery technique can be applied to any medical application domain that deals with vast amounts of data. The discovered trends can help the experts in the field to better understand the environment the data describes.

6 References

[1] Robert Andrews, Joachim Diederich, and Alan Tickle, "A Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks", *Neurocomputing Research Center*, 1995

[2] Amit Gupta, Sang Park, and Siuva M. Lam, "Generalized Analytic Rule Extraction for Feedforward Neural Networks", *IEEE transactions on knowledge and data engineering,* 1999

[3] Rudy Setiono, "Extracting Rules from Pruned Neural Networks for Breast Cancer Diagnosis", *Artificial Intelligence in Medicine*, 1996

[4] Rudy Setiono and Huan Liu, "Effective Data Mining Using Neural Networks", *IEEE transactions on knowledge and data engineering,* 1996

[5] Tony Kai, Yun Chan, Eng Chong Tan, and Neeraj Haralalka, "A Novel Neural Network for Data Mining", *8th International Conference on Neural Information Processing Proceedings. Vol.2*, 2001

[6] Mark W. Craven and Jude W. Shavlik, "Using Neural Networks for Data Mining", *Future Generation Computer Systems special issue on Data Mining,* 1998

[7] Jason T. L. Wang, Qicheng Ma, Dennis Shasha, and Cathy H.Wu, "Application of Neural Networks to Biological Data Mining: A Case Study in Protein Sequence Classification", *The Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* August 20-23, 2000 Boston, MA, USA.

Table 1: Data Description

	Variable	*Description*
I_1	AGE	Age (years)
I_2	SEX	Sex (1=Male, 2=Female)
I_3	SMOKSTAT	Smoking status (1=Never, 2=Former, 3=Current Smoker)
I_4	QUETELET	(weight/(height^2))
I_5	VITAMIN USE	Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
I_6	CALORIES	Number of calories consumed per day
I_7	FAT	Grams of fat consumed per day
I_8	FIBER	Grams of fiber consumed per day
I_9	ALCOHOL	Number of alcoholic drinks consumed per week
I_{10}	CHOLESTEROL	Cholesterol consumed (mg per day)
I_{11}	BETA DIET	Dietary beta-carotene consumed (mcg per day)
I_{12}	RET. DIET	Dietary retinol consumed (mcg per day)
O_1	BETA PLASMA	Plasma beta-carotene (ng/ml)
O_2	RET. PLASMA	Plasma Retinol (ng/ml)

Table 2: Discrete Intervals

Variable	*No. of Neurons*	*Intervals*
AGE	6	[20-30],(30-40],(40-50],(50-60],(60-70],(70-80]
SEX	2	1=Male, 2=Female
SMOKSTAT	3	1=Never, 2=Former, 3=Current Smoker
QUETELET	6	[15-20],(20-25],(25-30],(30-35],(35-40],(40+]
VITAMIN USE	3	1=Yes, fairly often, 2=Yes, not often, 3=No
CALORIES	5	-1000,(1000-1500],(1500-2000],(2000-2500],(2500+]
FAT	4	[0-50],(50-100],(100-150],(150-200]
FIBER	4	[0-5],(5-10],(10-15],(15+]
ALCOHOL	6	[0-2],(2-4],(4-6],(6-10],(10-20],(20+]
CHOLESTEROL	4	[0-250],(250-500],(500-750],(750+]
BETA DIET	4	[0-2000],(2000-4000],(4000-6000],(6000-8000]
RET. DIET	4	[0-500],(500-1000],(1000-1500],(1500+]
BETA PLASMA	6	[0-100],(100-200],(200-300],(300-400],(400-800],(800+]
RET. PLASMA	6	[0-200],(200-400],(400-600],(600-800],(800-1000], (1000-1200]

Table 3: The Six Predicted Trends

	Retinol Concentration	**0-100**	**0-100**	**0-100**	**200-300**	**100-200**	**100-200**
	Beta-Carotene Concentration	**400-600**	**600-800**	**800-1000**	**400-600**	**400-600**	**600-800**
AGE		50-60	30-40	50-60	30-40	40-50	50-60
SEX		Female	Female	Female	Female	Female	Female
SMOKSTAT		Current	Never	Never	Never	Never	Former
QUETELET		15-20	35-40	25-30	20-25	20-25	25-30
VITUSE		No	Not Often	Often	Not Often	Often	No
CALORIES		1500-2000	1500-2000	1500-2000	2000-2500	1500-2000	1500-2000
FAT		50-100	0-50	50-100	50-100	50-100	50-100
FIBER		10-15	10-15	10-15	15+	10-15	15+
ALCHOHOL		0-2	0-2	0-2	0-2	0-2	10-20
CHOLESTROL		0-250	0-250	0-250	0-250	0-250	0-250
BETADIET		2000-4000	2000-4000	0-2000	0-2000	4000-6000	2000-4000
RETDIET		0-500	0-500	1500+	1000-1500	0-500	1500+