Improving Healthcare Predictive Modeling using NeuroSequences

Hung-Han Chen
8787 Southside Blvd, Suite 503
Jacksonville, FL 32256
hunghchen@gmail.com

Michael T. Manry
Dept. of Electrical Engineering
University of Texas at Arlington
Arlington, Texas 76019
manry@uta.edu

Abstract:

This paper presents a practical method to improve the performance of MLP neural networks on healthcare predictive modeling. By adding a layer of data exploration using SOM, the data is clustered, at the same time the topological property is preserved, and then the task of predictive modeling is transformed to less complexity with the concept of "divide and conquer". The impacts of this new method are discussed and models of 3-month inpatient risk for 2.4 million insured members, an extreme unbalanced data, are compared to the result of a leading commercial risk score software.

1. Introduction

Predictive modeling for healthcare industry is "a set of tools used to stratify a population according to its risk of nearly any outcome"[3]. One of the goals of member profiling is to identify opportunities for intervention before the occurrence of adverse outcomes that result in high medical costs. It often involves data with unknown characteristics. If the targeted outcome is a rare event from the population, then it is usually very difficult to make prediction with such unbalanced data distribution.

Neural network has been one of the important methods for problem solving based upon the concept of artificial intelligence. The easy-to-use supervised learning rule, Backpropagation, has made Multi-Layer Perceptrons (MLP) popular for solving pattern recognition problems. But there also have been some critics for MLP neural networks regarding different aspects from many intelligent researchers since the day one. Unfortunately, most of them still are the challenges that neural networks need to face today.

Beside the claim that MLP neural networks may be trapped in local minima instead of finding the global solution, one of the major obstacles for neural networks becoming a real solution for practical problems is that the MLP neural networks have problems of scaling [1]. This issue of what and how Perceptron network will function when increasing the size and complexity of problems is often overlooked. There are also other concerns on MLP neural networks as described in [2]: it is not integrated with cost function; it needs long time to train; it may be over-fitting if training too long; it has catastrophic unlearning phenomenon; and it is mysticism to most people.

Section 2 describes the data representation and feature selection from healthcare raw data. Section 3 revisits the two types of conventional neural networks, MLP and Self-Organized Map (SOM). A practical method, NeuroSequences was proposed to improve MLP neural networks for predictive modeling in section 4, followed by discussion of how this new method can help resolving the drawbacks of MLP neural networks in section 5. Section 6 presents the result of this method applied to a 2.4 million-member population from a health insurance company to assess their inpatient risk for the next 3 months. Comparison with leading commercial rule-based software with clinical and treatment episode is also included. The conclusions and further discussion are given in section 7.

2. Data Representation and Feature Selection

Pre-processing is needed to convert raw data to input features for predictive models. Medical and pharmacy claims history within a certain period of time can be summarized, grouped and aggregated by ICD-9 diagnostic codes, CPT-4 procedure codes and NDC pharmacy codes into a set of input features. This set of input features can also include the utilization and grouping for major disease categories, like CAD, CHF, and Diabetic etc.

Let $X(m)$ be the vector of input features for member *m* from the insured population.

$$X(m) = (x_1(m), x_2(m), \ldots, x_N(m)) \quad (1)$$

Where *N* is the total number of features.

This set of input features can be used to model different healthcare related outcomes. Depending on what target the model is predicting, the most relevant features to the targeted outcome can be selected if their R-square from logistic regression are greater then a chosen minimum criterion, σ.

$$x_i \to x_j^{'} \text{ if } R_i^2 > \sigma \quad (2)$$

The new input vector for a designed outcome is then

$$X^{'}(m) = \left(x_1^{'}(m), x_2^{'}(m), \ldots, x_P^{'}(m)\right) \quad (3)$$

Where *P* is the number of selected features.

3. Conventional Neural Networks

3.1 Multi-Layer Perceptions (MLP)

The MLP neural networks with Backpropagation learning algorithm may have several drawbacks described in section 1; however, they do, in principal, offer all the potential of universal computing devices. They were intuitively appealing to many researchers because of their intrinsic nonlinearity, computational simplicity and resemblance to the behavior of neurons [1].

In training iteration *t* for MLP, the batch mode Backpropagation learning algorithm propagates the error term from output layer back to hidden layers, and updates the weight vector of neuron *v*, $W_v(t)$, using *gradient descent* method:

$$W_v(t+1) = W_v(t) + \alpha(t)\frac{-\partial E_v(t)}{\partial W_v(t)} \quad (4)$$

where $E_v(t)$ the error term propagated back to the neuron *v* and $\alpha(t)$ is the learning factor. The adaptive mechanism for learning factor can be easily achieved by:

$$\alpha(t+1) = \begin{cases} \alpha(t) * 2 & \text{if Error decreases} \\ \alpha(t)/2 & \text{if Error increase} \end{cases} \quad (5)$$

3.2 Self-Organized Map (SOM)

Kohonen's Self-Organizing Maps (SOM) algorithm is considered as one of artificial neural models for the brain, especially the experimentally found "ordered maps" in the cortex layers. Some researchers are able to produce simulation solutions to the cortical mapping problem by using SOM [7].

A SOM consists of a single-layer feedforward network that is utilizing unsupervised competitive learning to produce low-dimensional representation of the training sample while preserving the topological properties of the input space [8].

SOM often is trained by updating the weights for each input vector, the update formula for neuron *v* with weight vector $W_v(t)$ at time *t* is

$$W_v(t+1) = W_v(t) + \Theta(v,t)\alpha(t)(X_t' - W_v(t)) \quad (6)$$

where $\alpha(t)$ is a monotonically decreasing learning coefficient and X_t' is one of $X^{'}(m)$ to be the input of SOM at time *t*. The neighbourhood function $\Theta(v,t)$ depends on the lattice distance between the best matching neuron for X_t' and neuron *v*.

While the SOM algorithm may differ from traditional clustering analysis by adding the element of neighborhood function, the end result of SOM is not so different from clustering analysis in the sense of input-output relationship: there will be one single *winning* neuron, whose weight vector lies closest to the input vector X_t'.

Even though SOM algorithm inherits the capabilities of unsupervised learning and clustering analysis, the one-layer ordered map is simply not enough when a hierarchical structure is required, as the anatomical finding of cortex suggests.

4. Method of NeuroSequences

4.1 Memory and Learning

The architecture of neural networks is loosely based on the structure of human nervous systems. When believing brain is a network of many neurons, researchers in 1960s were interested in modeling brain by grouping a bunch of neurons together. With the learning capability improved by Backpropagation algorithm in the late 1980s, the neural network's

knowledge and memories can then be distributed throughout its connectivity, just like a real brain.

However, the structure of neural networks is way too simple compared with the physical architecture of the brain, the neocortex, and the performance of neural networks is still far from satisfaction. On the other hand, neuroscientists have long ago discovered cortical columns [4, 5] in the human 6-layered cerebral cortex in the sense of functional and/or anatomical features. With the advance of modern neuroscience from past decades, researchers have summarized four attributes of neocortical memory [6]:

1. The neocortex stores sequences of patterns.
2. The neocortex recalls patterns auto-associatively.
3. The neocortex stores patterns in an invariant form.
4. The neocortex stores patterns in a hierarchy.

To improve the performance of MLP neural networks closer to a human brain, the structure of MLP neural networks needs to adapt the concept of cortical columns. The focus here is on adding the functionality of auto-association with unsupervised learning, invariant form with clustering analysis, and more layers to the hierarchical structure of the traditional MLP neural networks.

As neural network researchers finding ways to improve MLP structure, some have developed Neural Network Tree (NNTree) [9, 10] to integrate the advantages of decision tree and neural networks. A typical NNTree can have up to 6 levels, or the depth of the tree is 6. However, there are issues surrounding the efficiency and effectiveness for its implementation and the splitting criterion for the non-terminal nodes of the tree [11, 12].

There is another type of research that fits a local model from the winning neuron and a set of neighbors of the SOM map by using a set of single layer neural networks [13]. The training of the system consists of two phases: first, the SOM is trained with the input data set; second, all the single layer neural networks are trained using the weights of SOM. The goal here is to obtain a finite set of local models that represents the global dynamics of the data. However, this method uses alternative cost function with only first order approximation.

4.2 Proposed Training Algorithm

With the needs of unsupervised learning and clustering analysis to be incorporated with neural networks using supervised learning, the proposed method in this paper suggests using one level of SOM and one level with several Backpropagation MLP neural networks; therefore there is no need to eliminate non-terminal nodes as in the NNTree. Then by controlling the data sequences parsed through the subgroups of SOM neurons to MLP neural networks, local dynamic modeling can be achieved under the topological space created by SOM.

In other words, instead of using SOM to transform input data into low-dimensional data projection, the method proposed in this paper uses SOM as a sequence parser, passing groups of input vectors through the SOM neurons to the next layer of MLP networks. Figure 1 illustrates the concept of NeuroSequences.

The detail of parsing the input vectors through SOM is described as the follows. After the ordered map is created from SOM algorithm, subgroups of neurons can be formed according to its property of topological preservation. There are two steps in the process to form SOM subgroups, as described in the following:

1. Rank each SOM neuron with a measurement.
2. Form subgroups based upon that measurement with desirable sizes.

Assuming two subgroups on the SOM plane were formed, the input vectors X_t', which are associated to those neurons in one of the two subgroups, will be then identified and labeled as two partitions are created within the input space.

$$X_t' \rightarrow \begin{cases} G_{s1} & \text{if } \textit{neuron} \text{ in subgroup } 1 \textit{ wins} \\ G_{s2} & \text{if } \textit{neuron} \text{ in subgroup } 2 \textit{ wins} \end{cases} \quad (6)$$

The subgroups of input space, G_{s1} and G_{s2}, then are sent to MLP network 1 and MLP network 2 for processing, respectively.

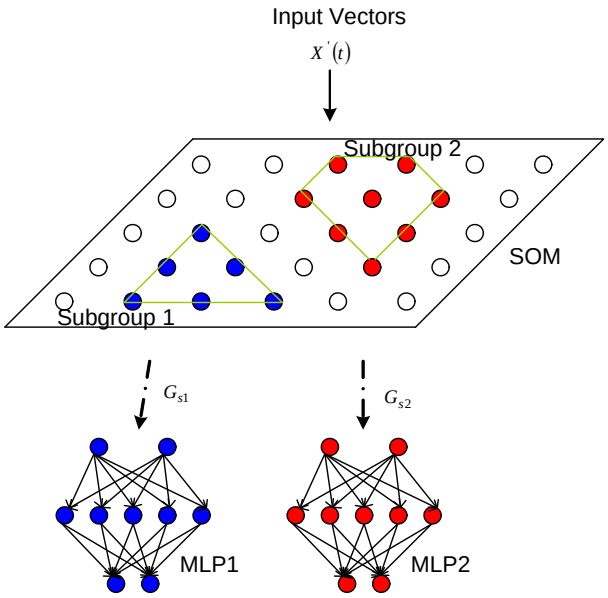

Figure 1. Concept of NeuroSequences

5 Impact of NeuroSequences

The architecture of the proposed method has produced the effects on minimizing the drawbacks of MLP neural networks with two schemes: One is the data exploration with SOM, and the other is controlling data flows. The impact for resampling unbalanced data is also discussed in this section.

5.1 Data Exploration and Visualization

It would be interesting to see how human brain solves complex problems. However, there is no reason to expect that a person can solve a complex problem by using same amount of time, same amount of resource, achieving same level of performance as solving a simple problem. It is almost safe to say that everyone will face difficulties for up scale problems. Still, complex problems need to be solved in the real world.

To conquer a difficult problem, the first thing one can do is to understand the data. Often the process of understanding data is achieved by plotting graphs and charts, and based upon that, our brains can then make decisions to solve the problem. In other words, data visualization has been considered an important way to understand data characteristics. If there is a mechanism for visualization of high-dimensional data, as SOM algorithm does, then we can apply the strategy of divide and conquer to solve complex problems when there is no known good method. This process can also be viewed in the concept of NeuroSequences of Figure 1

As we know that clustering analysis partitions a data set into subsets, as clusters, so that the data in each subset share some common characteristics, SOM can further form subgroups to preserve the input data topology by preserving the neighborhood relationships from the projection [14]. Therefore, the relationships between the SOM neurons can now be very useful for data exploration if the topographic error is minimized.

With this proposed method, the layer of data exploration using SOM is added to reduce the complexity of problem by performing clustering analysis and input space transformation. And since the complexity has been reduced, MLP neural networks can achieve the desired result with less training time, overfitting caused by training too long can be avoided here.

It has been an ultimate goal for many researchers to solve the problem of global minimum with advanced technologies. However, global minimum is not often seen in complex problems. Especially when population is changing, coding system is changing, and fee for service is changing everyday, it certainly is impossible to measure a global minimum without accounting for all the factors. For those problems that global minimum may exist, an advanced search method [17] can help Backpropagation getting out of most local minima and eventually it may reach the lowest point on the error surface. Before that happens, this proposed method could still be useful when solving problems that have more than one solution, i.e. constrain satisfaction. And most of operation research problems, including pattern recognition and member profiling, fit into this category.

5.2 Controlling Data Flows

If divide and conquer can help reduce the complexity of the problem, then controlling the data flow would help driving the solution to a specific direction. This direction here can be viewed as an operational point, which plays a similar role of a cost function. In fact, risk identification often requires careful balancing of sensitivity and specificity.

When we are choosing a point of operation, often we need to consider the resource we have, the penalty for misclassification, or even a desired goal set by management team. Therefore, even though we may show all the relationship between the rate of true positives and the rate of false positives in a ROC curve (Receiver Operating Characteristics) from a risk score system, practically we can only choose one or a few points for operation with the intention of meeting all the criteria I just described above.

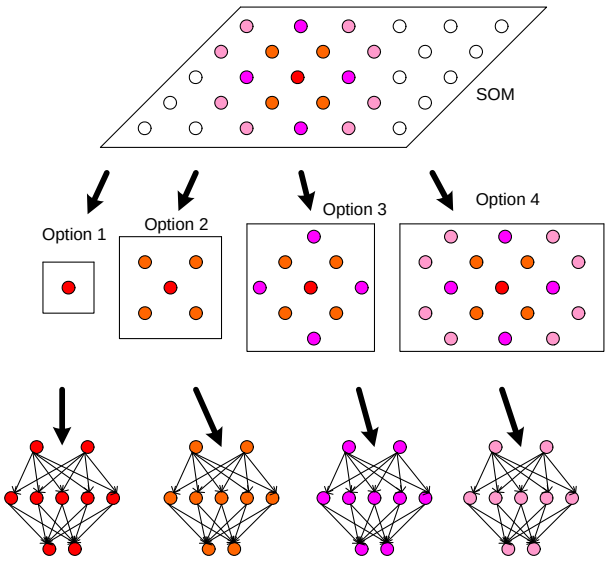

Figure 2. Examples of Controlling Data Flow

The method presented in this paper also includes a flexible mechanism to control the sizes of topological subgroups generated in SOM layer. Assuming the accuracy of a SOM neuron can be painted with 4 different colors. We can use RED color to identify the accuracy of a SOM neuron for 40% or above. ORANGE for 30%, PURPLE for 20%, PINK for 10%, and the rest is uncolored. In order to achieve the goal of operational point, there will be at least 4 choices to setup the SOM subgroup so that appropriate amount of members will be sent to the following MLP network. The examples of controlling data flows are illustrated in Figure 2. With this mechanism, we can approximate the result model to the neighborhood of specific target. Figure 3 shows three models of Neurosequences selection are located in different desired target ROC regions.

5.3 Unbalanced Data

Many traditional approaches to machine learning classification problem assume the target classes sharing the similar prior probabilities. With extreme unbalanced data, those approaches, including MLP neural networks, will certainly fail to create a predictive model if no remedy has been applied. There are several types of techniques that have been studied for this purpose.

Cost sensitive learning is one type of remedies that are used to solve the issue of unbalanced data. This method is intuitively to modify the classifier so that the learning takes place proportionally to the distribution of the classes. Furthermore, the cost or penalty for misclassification could be included in the cost sensitive learning. However, the cost is often difficult to be quantified and the classifier modified with input distribution will not be as useful in other domains.

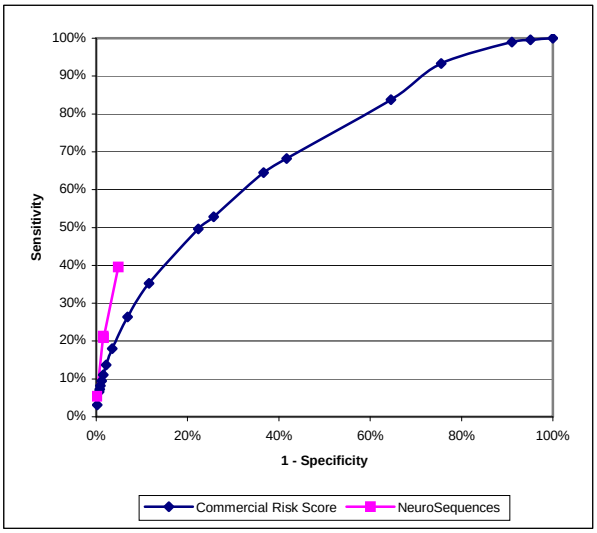

Figure 3. Different Operation Points in ROC curve for NeuroSequences

Resampling is another type of remedies that is not to modify the learning behavior of the classifier but to modify the distribution of the data. While oversampling is to increase the number of records of minority class, undersampling is to decrease the number of records of majority class. If training time and memory complexity are not the issues, oversampling has the advantage over undersampling because no information from data has been lost during the oversampling.

There are more techniques that could be added to resampling methods in a meaningful manner. Therefore, these techniques can normally perform better than random resampling. One of the techniques is to resample the data based on the number of records per cluster rather than just the number of records per class [15]. This guided resampling technique has been proved to perform better than blind resampling, and certain knowledge about the subcomponents for each class is required.

There are two advantages in the method presented in this paper regarding the need of resampling for extreme unbalanced data. First, with property of probability density matching in SOM, the feature map tends to overrepresent regions of low input density and underrepresent regions of high input density [16]. Second, when MLPs resample the data passed from SOM's neurons, it has similar benefit of resampling from clusters, with property of topology preserving in SOM.

6 Model Results for 3-Month Inpatient Risk

A population of 2.4-million Florida members is assessed to predict their future inpatient risk within the next three months. The outcome is designed to be binary. The analysis periods of claim history are set up as Figure 4. The prevalence for validation data is about 1.3%. There is no easy solution for such extreme unbalanced dataset. Even using resampling techniques, traditional MLP neural networks still cannot create a robust and reliable model.

A model created by leading commercial health risk assessment software is used for comparison. This software uses Episode Treatment Groups (ETGs) as the fundamental building block for illness classification. However, the risk score from this model leads to too many false positives, as shown in Table 2.

With NeuroSequences, medical and pharmacy claims history of one year are summarized and grouped, according to ICD-9 diagnostic codes, CPT-4 procedure codes and NDC pharmacy codes, into a set of features. The standard inputs consist of 77 features. The most relevant features to the designed outcome are selected with a minimum criterion of R-square. For this dataset, 53 features are selected based upon equation (2).

After the number of input features is determined, the SOM map used in the simulations is then constructed with the dimensions of 8 columns by 12 rows, as shown in Figure 5. The result of SOM training can be painted with different colors on this two dimensional map. For examples, Red paint indicates the SOM neurons with accuracy of 15% or above. Purple for 10%, and Pink for 6%. For the purpose of controlling data flow, it is flexible to define the number and size of subgroups based upon the SOM topology. Figure 5 also illustrates the subgroups from the 15k model.

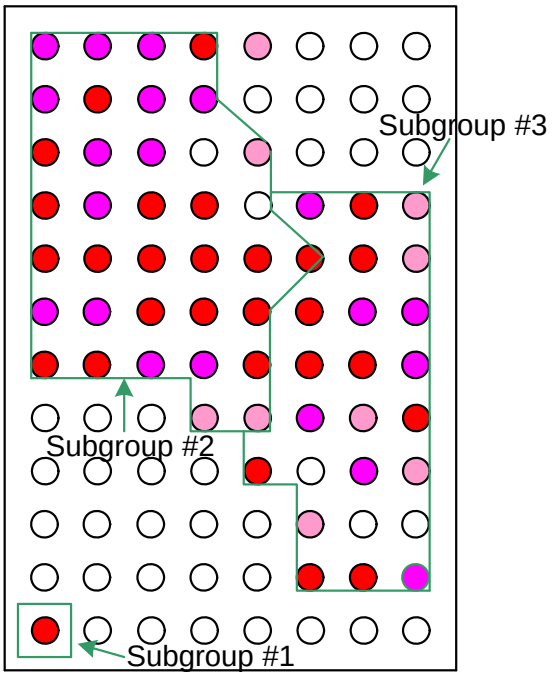

Figure 5. Example of SOM Topology

MLP networks are then trained with the subsets of the input data, as described in Figure 1 and equation (5). If the input data are excluded from the designed subsets, as associated with the white-color neurons or outside of all subgroups, then they will be filtered to be the default class. The MLP networks used in the simulations are Backpropagation MLP networks with adaptive learning factor. The number of hidden neurons is set to be 10, and the number of iterations is 500.

Table 1. The example of confusion matrix

		Target	
		True	*False*
Prediction	*True*	TP	FP
	False	FN	TN

The performance for a binary output is normally presented with a confusion matrix, as shown in Table 1. True positive (TP), false positive (FP), false negative (FN), and true negative (TN) are counted from the prediction. With this matrix, we can then calculate the sensitivity and positive predictive value (PPV) with the following formula:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$PPV = \frac{TP}{TP + FP}$$

A predictive model normally wants to maximize these two measures. As described in Figure 3, NeuroSequences can create different models when

targeting different operational points. The 5k model created by NeuroSequences in Table 2 is targeting to select 5,000 members who are likely to have an inpatient stay in the next three months. Table 2 also lists a 10k model and a 15k model. Figure 6 shows the bar chart for TP and FP from risk scores and simulations.

7. Conclusions and Further Discussion

This paper proposes a practical method to improve the performance of MLP neural networks on healthcare predictive modeling. By adding a layer of data exploration using SOM, the task of predictive modeling was transformed to less complexity with the concept of "divide and conquer". The impacts of this new method were discussed in Section 3 regarding to those drawbacks of traditional MLP neural networks.

Table 2. Comparison between NeuroSequences and Commercial Risk Score

Commercial Risk Score	**True Positives**	**False Positives**	**Total**	**Sensitivity**	**PPV**
> 13	1,748	9,099	10,847	5.31%	16.12%
> 14	1,619	8,124	9,743	4.92%	16.62%
> 15	1,531	7,346	8,877	4.65%	17.25%
> 16	1,416	6,679	8,095	4.30%	17.49%
> 17	1,302	6,121	7,423	3.96%	17.54%
> 18	1,213	5,582	6,795	3.69%	17.85%
> 19	1,143	5,116	6,259	3.47%	18.26%
> 20	1,081	4,695	5,776	3.29%	18.72%
> 21	1,019	4,356	5,375	3.10%	18.96%
> 22	973	4,042	5,015	2.96%	19.40%
Chen's model					
5k model	1,778	2,708	4,486	5.40%	39.63%
10k model	2,412	5,913	8,325	7.33%	28.97%
15k model	3,004	10,336	13,340	9.13%	22.52%

Besides the flexibility of approaching to a designed operational point, this method also maintains the property of computational simplicity from SOM and MLP. Also an advanced search method can help Backpropagation and MLP getting out of most local minima.

However, there are downsides, too. This method needs more training data to take the advantage of data exploration and followed with multiple MLP neural networks. This may seem to be critical to some applications, but often not an issue for healthcare predictive modeling and other complex problems.

Another issue is that global minimum may still be out of reach when the data can be trained with limited time only. But nevertheless, this proposed method could still be used for solving problems that can often allow more than one solution, i.e. constrain satisfaction. Fortunately, problems of pattern recognition and classification fit into this category.

Currently, SOM and MLPs are trained separately in this proposed method. And their performances are quite good compared to the current method used. But it would become better if further researches can focus on how to train SOM and MLPs at the same time.

References:

[1] M. Minsky and S. Papert (1988), "Epilog: the new connectionism", Perceptrons, 3rd ed., Cambridge: MIT Press, pp. 247-280.

[2] An Introduction to Neural Networks, James A. Anderson, pp. 275-277. Cambridge: MIT Press

[3] MS Cousins, LM Shickle, JA Bander, "An Introduction to Predictive Modeling for Disease Management Risk Stratification", Disease Management, 2002; 5: 157-167.

[4] Vernon Mountcastle (1978), "An Organizing Principle for Cerebral Function: The Unit Model and the Distributed System", *The Mindful Brain* (Gerald M. Edelman and Vernon B. Mountcastle, eds.) Cambridge, MA: MIT Press.

[5] Hubel**,** D. H. & Wiesel, T. N. (1962) *J. Physiol.* 160, 106-154.

[6] Jeff Hawkins, (2004), "Chapter 4: Memory", On Intelligence, Henry Holt And Company, pp. 65 - 84

[7] Nicholas V. Swindale (2000), "How Many Maps are there in Visual Cortex?" Cerebral Cortex, Vol. 10, No. 7, 633-643, July 2000.

[8] Teuvo Kohonen (1982), "Self-Organized Formation of Topologically Correct Feature Maps", Biological Cybernetics, 43, 59-69, Springer-Verlag.

[9] H. Guo and S. B. Gelfand, "Classification trees with neural network feature extraction," IEEE Trans. On Neural Networks, Vol. 3, No. 6, pp. 923-933, Nov. 1992.

[10] Q. F. Zhao, "Evolutionary design of neural network tree - integration of decision tree, neural network and GA," Proc. IEEE Congress on Evolutionary Computation, pp. 240-244, Seoul, 2001.

[11] T. Takeda and Q. F. Zhao, "Growing Neural Network Trees Efficiently and Effectively," Proc. International Conference on Hybrid Intelligent Systems (HIS'03), pp. 107-115, Dec. 2003.

[12] Pradipta Maji, Efficient Design of Neural Network Tree Using A New Splitting Criterion, Neurocomputing, Elsevier (Article in Press).

[13] O. Fontenla-Romero, A. Alonso-Betanzos, E. Castillo, J. C. Principe, B. Guijarro-Berdiñas, "Local Modeling Using Self-Organizing Maps and Single Layer Neural Networks", ICANN 2002, pp. 945-950

[14] E. Ursuaga and F. Martin, "Topology Preservation in SOM", International Journal of Applied Mathematics and Computer Science, Volume 1, Number 1, 2004, pp 19 - 22.

[15] Nickerson, A., Japkowicz, N. and Milios, E. "Using Unsupervised Learning to Guide Resampling in Imbalanced Data Sets," Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, 2001.

[16] S. Haykin, Neural Networks, IEEE Press, Macmillan College Publishing Company, Inc. pp. 422.

[17] Hung-Han Chen, "The Turning Points on MLP's Error Surface", Accepted in Fifth International Symposium on Neural Networks, September 2008, Beijing, China.

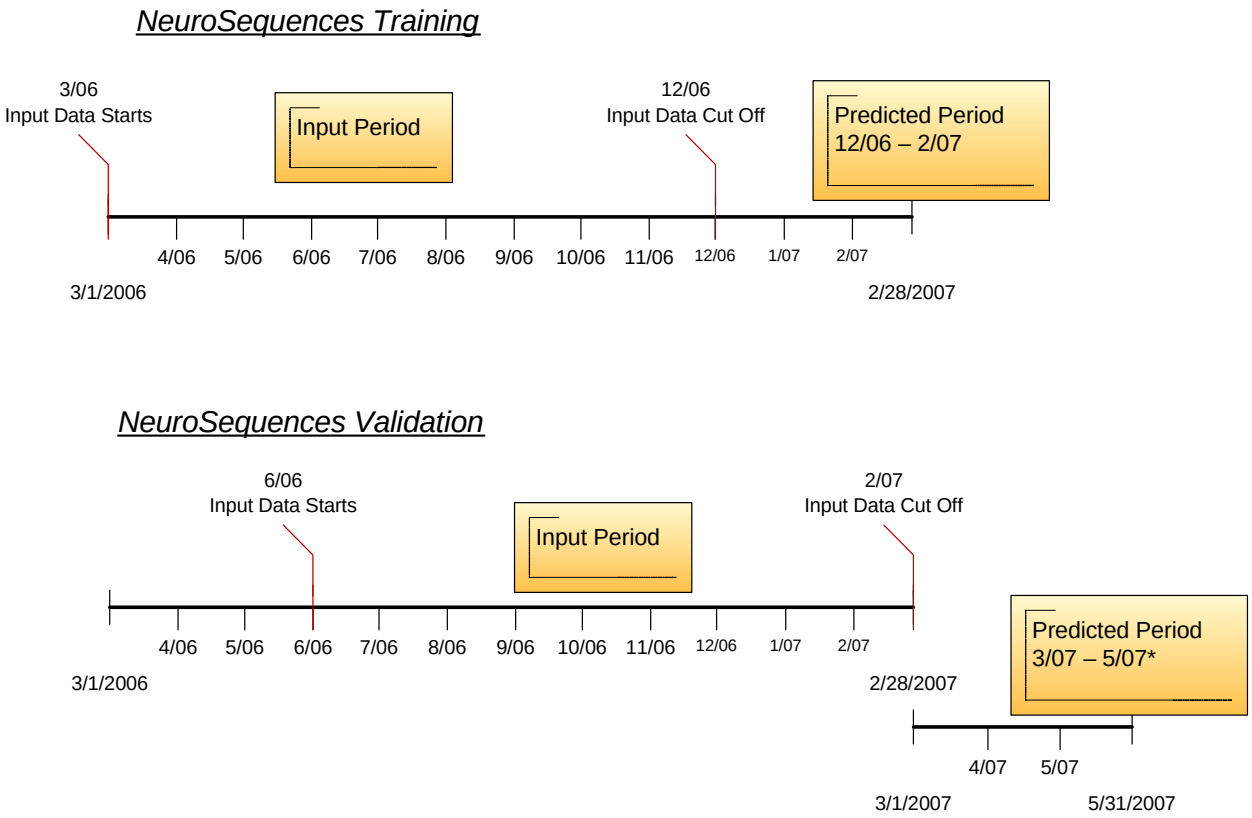

Figure 4. Timelines for Training and Validation

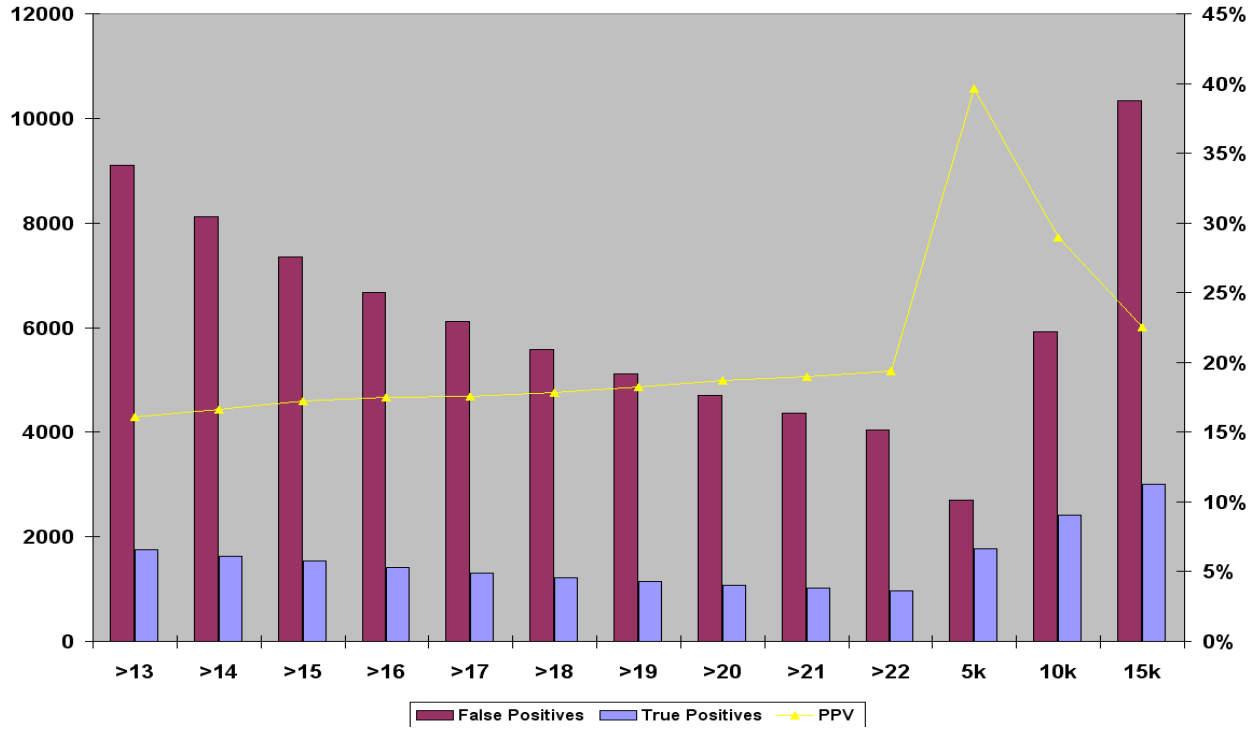

Figure 6. Bar Charts for TP and FP from Simulations