Application of Self-Organizing Feature Maps to Water Resources Projects

B. Hsieh¹, and J. Ratcliff¹

¹US Army Engineers, Research and Development Center, Vicksburg, Mississippi, USA

Abstract – The self-organizing feature maps (SOFM) is a kind of ANNs (Artificial Neural Networks) method which is capable of clustering, classification, estimation, prediction, and data mining in a wide-spread range of disciplines. Two types of complex water resources related applications, namely, watershed hydrology and coastal storm surge, are demonstrated here. The former application is to find the best match watershed from a large knowledge base of over one thousand quantifying watersheds and to determine the reliability of “transplant” watershed information during the clustering processes while the later application uses SOFM adequately characterize the storm surge response, and provide a means for reliably estimating surge response for storms not simulated with a selected physics-based surge model. The computational procedures and results are presented.

Keywords: SOFM, Neural Networks, watershed similarity, storm surge response, water resources projects.

1 Introduction

Unsupervised training means the networks learn from their own classification of the training data, without external help. It is assumed that class membership is broadly defined by the input patterns that share common features, and that the network will be able to identify those common features across the range of input patterns. Self-organizing Feature Maps (SOFM) is a special kind of neural network that can be used for clustering tasks. Only one map node (winner) at a time is activated corresponding to each input. The location of the responses in the array tends to become ordered in the learning process as if some meaningful nonlinear coordinate system for the different input features were being created over the network. This illustrates an important and attractive feature of SOFM applications, in that a multi-dimensional input ensemble is mapped into a (one or) two-dimensional space, preserving the topological structure as much as possible. The SOFM is a set of artificial neurons, which are ordered in N space. A two dimensional array (n=2) is the most common map and is used to map an input signal in Rm (m >n) space onto the two-dimensional space. Basically, an SOFM typically consists of two layers (Figure 1). One is an input layer into which input feature vectors will be fed and other layer is a two-dimensional competitive layer, which orders the neuron’s responses spatially. Neurons can be arranged on a rectangular map so that they can be implemented using a simple 2-D data array. A hexagonally arranged neuron map is, however, often used because it has the advantage of the Euclidean distance (equi-distance) between adjacent neurons (Kohonen [6]). The detailed theoretical development for the algorithms for both supervised and unsupervised ANNs can be found in the most textbooks.

Visualization techniques to depict the data structure of the feature space in the form of clustering of neurons in the 2-D SOFM have been developed (Ultsch [10]). This visualization typically uses the grayscale to illustrate the distance between connection weights. The light shading typically represents a small distance and the dark shading represents a large distance.

This type of visualization is useful as long as relatively clear cluster boundaries exist or the granularity of the distance differences is large. When the cluster boundaries get fuzzy or the granularity of the distances become too small to represent with the grayscale, it becomes increasingly difficult to identify fuzzy cluster landscapes. Moreover, since all distance values are normalized, only relative (qualitative) analysis is allowed. Subsequently, this “grayscale distance map” cannot be used to compare different SOFM mapping results.

When the SOFM is used to discover some structure of the given samples in the feature space, it is often useful to visualize the finding in the form of clustering formations. Visualization techniques to depict the data structure of the feature space in the form of clustering of neurons in the 2-D SOFM have been developed (Ultsch [10]). This visualization
typically uses the grayscale to illustrate the distance between connection weights. The light shading typically represents a small distance and the dark shading represents a large distance.

NeuroDimensions [7] developed a visual version for the Kohonen topological feature maps to check the performance of SOFM. Three basic windows used for evaluating the clustering are quantization metric, united distance, and frequency.

Quantization Metric: It produces the average quantization error, which measures the goodness of fit of a clustering algorithm. It is the average distance between each input and the winning process element (PE). If the quantization error is large, then the winning PE is not a good representation of the input. If it is small, then the input is very close to the winning PE. The quantization error is best for comparing the clustering capabilities between multiple trainings of the same SOFM on the same point.

Unified Distance: This is the distance between PE clustering centers. The weights from the input to each PE cluster centers of the SOFM. Inside a cluster of inputs, SOFM PEs will be close to each other.

Kalteh, Hjorth, and Berndtsson [5] reviewed of the SOFM approach in water resources: analysis, modeling and application. They reviewed several applications in water resources processes and systems including riverflow and rainfall-runoff, precipitation, surface water quality, climate change, environment, and ecology subjects. Several recent studies (Chen, Chen, and Lin [1], Srinivas, etc. [8], and Gomez-Carracedo, etc. [2]) also successfully addressed for water resources and environmental related projects using SOFM.

2 Watershed similarity analysis

The ability to predict watershed hydrologic conditions and the associated potential for flooding to occur plays a significant role in planning and operational activities. To make highly accurate hydrologic predictions, either physically-based or system-based, the system parameters and prediction variables are sometimes unavailable or even totally missing. This certainly curtails the capability of prediction, particularly for operations where very little time is available to conduct the analysis. Very often, the information for a particular watershed may be entirely unavailable; this situation could be resolved by the similarity concept.

The purpose of this approach is to find the best match watershed from a large knowledge base and to determine the reliability of “transplant” watershed information such as hydrologic and climatic parameters (Hsieh and Jourdan [3]). The degree of similarity is based on inter and intra relationships among many geologic, soil, hydrologic and climatic factors. Various methods have been employed to analyze the similarity between two objects.

2.1 Geospatial knowledge base development

GIS data often includes satellite and other remotely sensed imagery. An example of the analysis of imagery involves either supervised or unsupervised classification. Unsupervised classification of imagery involves the analysis of color or black and white pixels of the image for the purposes of classifying image objects and entities, where, tone, texture and hue are used. Supervised classification of imagery involves referencing the pixels to actual field or site conditions and color balancing of the image for similar classification purposes. ANNs are increasingly being used for the purpose of determining spatial patterns. In the area of landscape ecology, the landscape pattern is an important factor enabling classification. Indeed, more recent developments in the area of remote sensing analysis involve ANNs for the analysis of images for the purposes of classifying objects.

Geospatial data of geographic locations and characteristic natural and constructed features were gathered for the database development. GIS databases were utilized for this endeavor; specifically the EPA’s Better Assessment Science Integration Point & Non Point Sources (BASINS) system provided the 300-meter USGS Digital Elevation Model (DEM), Land Use/Land Cover, Soils (STATSGO), and watershed gauge locations within the conterminous United States of America (Figure 2). These gauge locations were selected with the criteria of 100 percent complete dataset for medium- to moderately large- sized basins, 6 to 7900 km².

Figure 2. Watersheds within the conterminous United States

Watershed development was conducted with the Environmental Systems Research Institute’s ArcGis/ArcView and the Department of Defense’s Watershed Modeling System (WMS). From the GIS databases, data was extracted, projected and shaped into Arc/Info griddled ASCII data as input into the WMS interface where basin delineation and parameter estimations were conducted. Watershed parameters such as Drainage Area, Basin Slope, Basin Length, Basin Perimeter, etc. were among the variables derived for the ANN’s analyses. Watersheds selected were within a 10 percent margin of error when the areas were compared with recorded drainage areas from BASINS.

From these selected watersheds, mean daily flow data for their respective periods of record were compiled for the ANN’s verification process. In addition, thirty-year mean monthly and annual precipitation, as well as temperature data,
were derived from PRISM (Parameter-elevation Regressions on Independent Slopes Model) and presented as GIS coverages. Subsequent GIS analyses produced mean monthly and annual, precipitation and temperature data, for all selected basins. The final knowledge base has the dimension of a 1064 watersheds x 70 variables matrix with final relevant parameters listed as follows.

**Geometric Parameters:**
- Basin Area, Basin Slope, Basin Average Elevation,
- Basin Shape Factor, Basin Sinuosity Factor,
- Average Overland Flow Distance, Maximum Flow Distance, Maximum Flow Slope, Maximum Stream Slope, Centroid to Nearest Point of MaxFlowDist

**Land Use/Land Cover Parameters:**
- Residential/Industrial, Agricultural Land, Rangeland,
- Forest Land, Open Water, Wetlands, Exposed Rock,
- Tundra, Glaciers

**Soil Type Parameters:**
- Sands and Gravel, Silts, Sandy Loam, Clays

**Hydrologic Parameters:**
- Seasonal and Annual Precipitation, Seasonal and Annual Temperature

A data-driven computational procedure including knowledge base and two components of ANNs (clustering and classification) and prediction (verification) was developed. Takatsuka [9] applied SOFM and interactive 3-D visualization to geospatial data. Ultsch, Korus, and Kleine [10] developed the integration of neural networks and knowledge-based systems in medicine.

### 2.2 Demonstration example

From the knowledge base, all the geometric parameters, the land use/land cover, the soil types and the seasonal and annual mean values of both precipitation and temperature were used to test this calculation procedure. In order to test the reliability of the system development, three sizes of watershed are selected to examine the performance. The detailed search process is only presented in the first example.

The goal of this test is to use a known watershed (gage number 4288000) to search for the best similar watershed. This part of study is divided into two portions. While the clustering analysis is used to identify the similarity between the watersheds, the classification analysis is used to verify the clustering performance. To check the reliability of the prediction, time series hydrographs are used to compare the resulting search pattern. In this procedure, the hydrograph of gage 4288000 is hidden purposely in order to check the performance of the system once the best similar watershed is found.

During the clustering computation, a 5 x 5 matrix of SOFM is initially selected. Through repeated iterations (usually 200) of the examination of frequency, unified distance, and quantization of the unsupervised synapse, an optimal clustering set to distribute the winner for each watershed is obtained (Figure 1). The numbers in this 5 X 5 matrix show the most similar watershed within the same group (there are 25 groups in this case).

For classification, the problem was trained with (Multi-layered Feed Forward neural Networks (MLP) ANNs and the outcome provided the confidence level of the clustering analysis, which resulted in a successful classification rate of about 91 percent meeting the target. This result indicates that watershed 4288000 belongs to the group with 103 (group 7) most similar watersheds from the original 1063 possible candidates. This clustering-classification process is repeated until the final target watershed is found. This process is called search iteration. Figure 3-4 show the classification verification during the first and second search iterations respectively. It is noted that the size of clustering for this iteration has been reduced to a 3 X 3 matrix.

![Classification Verification](attachment:classification_verification.png)

**Figure 3.** Classification verification for the first system iteration with 1064 watersheds (group – X-axis; number of assigned watershed – Y-axis)

![Classification Verification](attachment:classification_verification_2.png)

**Figure 4.** Classification verification for the second system iteration with 103 watersheds (group – X-axis; number of assigned watershed – Y-axis)

The final candidate for this search is the watershed number 01144000. This implies that the flow patterns from station 01144000 will be most similar to those of station 04288000. Flow hydrograph comparisons between these two stations
during the period 1999-2001 are shown in Figure 5. Although the flow pattern, particularly, the phase matches very well, the performance of the amplitude representations is dissatisfactory.

![Daily Flow (cms) starts from 09/30/01](image1)

**Figure 5.** Most similar flow (cms) (01144000 - pink) vs. observed flow (04288000 - blue) flow – X-axis; days – Y-axis

When examining the involved parameters between these two watersheds, the area and maximum flow distance showed a significant difference. The estimated hydrograph was adjusted by taking the area ratio of station 04288000 and station 01144000 (Figure 6).

![Flow (cms) starts from 09/30/01](image2)

**Figure 6.** Flow (cms) estimation (e04288000 - pink) vs. observed flow (04288000 - blue) after basin area ratio adjustment for 33 inputs approach (flow – X-axis; days – Y- axis)

The major element in making this integration system a success is to tune the clustering group as well as rechecking the performance of the classification process. But the identification of the reliability for application also requires data on how well the “transplant” performs. Therefore, a series of combinations including the features of input parameters is adopted. Table 1 summarizes the performance due to the selection of input parameters. This indicates that the important group parameters are hydrologic, geometry, soil type, and land use. The performance difference between geometry and hydrologic groups is quite small. The magnitude of hydrographs could not be adjusted by ratios obtained using hydrologic, soil type, and land use groups since they do not contain the basin area factor after the best candidate is found.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Candidate Watershed</th>
<th>Correlation Coefficient</th>
<th>Mean Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Groups</td>
<td>0114000</td>
<td>0.92</td>
<td>0.17</td>
</tr>
<tr>
<td>Geometry</td>
<td>4282000</td>
<td>0.82</td>
<td>-7.80</td>
</tr>
<tr>
<td>Hydrologic</td>
<td>4288000</td>
<td>0.83</td>
<td>2.99</td>
</tr>
<tr>
<td>Land Use</td>
<td>2472000</td>
<td>0.10</td>
<td>-28.34</td>
</tr>
<tr>
<td>Soil Type</td>
<td>1170100</td>
<td>0.67</td>
<td>-4.25</td>
</tr>
</tbody>
</table>

2.3 Summary of watershed similarity analysis

An integration of database and ANNs learning was used to identify a very complex nonlinear watershed similarity analysis for military hydrology applications. While the unsupervised ANNs, such as SOFM, were used to perform the clustering of watershed characteristics, the supervised ANNs were used to identify the best match candidate watershed for classification analysis. The search procedure requires several iterations of the clustering-classification loop. The current knowledge base consists of 67 geometric, hydrological, land use, and soil type factors for 1064 selected watersheds. After removing the internal dependency and examining the annual and season representation, 33 factors were selected for final analysis.

Three demonstration examples, including random selection, average size, and median size watersheds were used as the target to search for the best match corresponding candidate. The first example obtained a good correlation coefficient (0.92) for hydrograph prediction (2 years daily flow). It is found that the basin area ratio provides a reasonable factor for making the adjustment for hydrograph prediction. The preliminary sensitivity tests indicate that the hydrologic factors are the best factors in producing a fitness for transplant. In general, monthly hydrograph comparisons have better agreement than the daily hydrographs for both average and median size watershed examples. The most significant reliability is obtained when many watershed.
patterns are included in the knowledge base. Development of an automated search procedure for a unique solution is the direction proposed for further research.

3 Storm surge runs analysis

Numerical simulation is the most accurate and efficient modern technique to calculate the surge response during a storm/hurricane event. Recently, a very successful interagency team effort has been made to model the storm surge response for hurricane events along the northern coast of the Gulf of Mexico. The dependencies between surge and waves are treated through coupled models, and a probabilistic approach has been adopted for calculating inundation levels and their associated probabilities. However, numerous model runs are required to cover a wide range of possible hurricane scenarios to meet the management and project design needs. This newly established modeling framework requires a significant amount of resources including personnel and computer resources, as well as contract labor and other factors which raise project costs and completion time requirements. Good planning with an alternative technology path that can reduce costs for projects is highly desirable. Simulation using ANN techniques was examined as a possible tool for reducing the resources required to make storm surge estimates for design purposes. Often in design, a large set of storm surge simulations must be made for each of a series of different project alternatives (such as different levee alignments). Increasing numbers of alternatives dramatically escalate the computational requirements for a detailed modeling approach.

3.1 LACRP storm surge runs example

A computational procedure (Hsieh and Ratcliff [4]) to perform this effort is shown in Figure 7. The first step is to identify the significant storm parameters and use the resulting surge responses to build the ANNs model. The performance of this ANNs model is critical to assure the right input parameters are selected. The second step of the approach is clustering analysis, using SOFM to separate the similar storm patterns from the knowledge base (input parameters only), and to form a number of subgroups. The third step is to split each subgroup into two components: training and testing storm sets. The ensemble training component from all subgroups along with corresponding surges is the knowledge base which is assumed to represent the required ADCIRC runs, while the ensemble testing component from all subgroups along with corresponding surges are considered to be the unnecessary ADCIRC runs. More ADCIRC runs in the final ensemble testing group means a higher percentage of runs saving that can be obtained under the good performance of testing group from ANNs modeling. Although there is no particular rule to follow how to separate the training and testing group, but at least two numerical model runs from each subgroup need to be selected to be the training group if this subgroup contains more than or equal to two numerical model runs. For a large number of numerical model runs, the decision is based on either second level of clustering or the variation for the most sensitive storm parameters.

4 Figure 7. Computational procedure to determine necessary surge model scenario run.

They are 5 storm input parameters (CpLand (central pressure at landfall), VelAvg (forward translational speed of the storm center, which was assumed to be constant along the specified track), Rm (radius to the maximum winds), MaxWind (maximum wind speed at landfall), Distance (distance from the storm center at landfall to the location of interest), and Angle (angle between the storm center at landfall and the location of interest). The corresponding surge produced by the storm at the location of interest is the output. Since CpLand and Distance are negatively correlated to the surge response, a negative sign for both inputs is taken. The Angle for each storm is further decomposed into cosine (Angle-x) and sine (Angle-y) components. The first 4 input parameters are considered as global storm parameters while the remaining input parameters (Distance, Angle-x, and Angle-y) are treated as local, or positional, parameters.

This demonstration shows results for a computational point (141, circled in Figure 8) from among many considered in the LACPR study (Figure 8) to illustrate the ANN application concept. Usually, a correlation coefficient analysis is conducted between all the inputs and corresponding output in order to check the sensitivity of the system. Figure 9 shows the most significant input parameter – angle for point 141 with X and Y components over 152 storm events. Figure 10 illustrates the results of ANNs modeling for point 141. MLP is the training algorithm and total iterations are 5000. With high statistical significance (high correlation coefficient and low mean absolute error, for example), the ANNs modeling proves to be a satisfactory tool to quantify the relationship between inputs and output. The blue lines shows the computed peak surge using the ADCIRC model for all storms while the red line shows the computed peak surge values.
using ANNs modeling. This is the first step to identify the applicability for selected ANNs technique.

A 5x5 SOFM clustering analysis was then applied to group the dimensional 152x7 information (7 input parameter factors for each of 152 storms). The choice of the size for SOFM process is based on how detail you would like to deal with the clustering from the system. While large matrix may break 152 storms into too many pieces, small matrix may require the second level of clustering.

The ratio of total storm numbers to total subgroup numbers estimates the proper matrix size is 5x5 or 4x4. The final destination for each represented storm after iteration process for position adjustment is called “Get Winner”. An optimal pattern distribution matrix for these 152 storms all reach “WINNER” is shown in Figure 11.

Figure 8. Surge response points from LACPR ADCIRC model near New Orleans area.

Figure 9. X and Y components for approach angle from point 141 (x-axis represents storm numbers and y-axis represents sine/cosine of the angle, between 1 and -1).

Figure 10. Comparison of results from ANNs- MLP training for point 141 and calculated surges from ADCIRC simulations of 152 storms based on 7 storm input parameters (pink represents ANNs simulation and blue shows ADCIRC results; x-axis represents storm number and y-axis represents surge response (ft)).

Figure 11. A 5x5 SOFM clustering analysis and its training and testing components fro152 LACPR ADCIRC runs based on 7 storm parameters (point 141).

The number for each grid cell of the matrix shows that similar patterns are found from 7 input parameters. It is noted that the “0” value for a particular grid cell in the matrix indicates that there is no storm falls that specific pattern. It usually happens when too large size of matrix is assigned or too little variation of pattern does exist. The splitting process is then applied to separate the storms within a grid cell into a training component and a testing component after an ensemble process is conducted by collecting the minimum required storm events into the training group and putting the remaining events into the testing group. Since the Angle-y was found to be the most significant parameter, it was used as a criterion, including extreme values and part of represented values from this parameter, to determine into which component each storm should go. The lower part of Figure 11 presents the final assignment of storms into the training and testing components. The ANNs training, using 85 selected storms with surge as output was applied to examine performance for the testing component. Figure 12 illustrates performance of the testing component, the 67 selected storms (a correlation coefficient 0.912 was achieved). From this analysis, it is possible to avoid 44 percent (67 out of 152) of ADCIRC simulations, if the storm simulations have to be repeated.

Results also suggest that it might be possible to intelligently reduce the number of storms considered in simulations to look at various alternatives. To compare the surge frequencies computed based upon these three series, the response surges for return periods up to 2000 years are computed. The maximum deviation is about 0.18 m (0.6 ft)
between original ADCIRC runs and this combination approach for point 141 (Figure 13).

3.2 Summary of storm surge runs analysis

This study uses unsupervised ANNs (SOFM) to cluster storm patterns. This is based on four global storm parameters and three local, or positional, parameters. The angle between the location of interest and the location of storm landfall was found to be the most sensitive input parameter, due in large part to the influence of the Mississippi River delta and levee system in dictating local surge conditions in southeastern Louisiana. The splitting process is able to separate all storm patterns into training and testing components. The number of storms in the testing component equals the number of numerical runs that can be potentially be reduced by simulating surge through the ANNs model using the training components along with their corresponding surge from actual numerical simulations of storm surge using the ADCIRC model. A demonstration project (LACPR), results for a single point in each case, show successful application of the developed computational procedures. Point 141 from LACPR project demonstrates reducing model runs about 40 percent of storm model runs. Results showed that the more storm patterns that are involved in the training component, the higher percentage in the reduction of numerical runs, which makes intuitive sense.

4 Acknowledgement

The Military Hydrology Research Program in the Coastal and Hydraulics Laboratory at the US Army Engineer Research and Development Center funded the watershed similarity study and USACE New Orleans District funded the storm surge runs analysis effort. Permission was granted by the Chief of Engineers to publish this information.

5 References


