
http://dx.doi.org/10.1109/ICONIP.1999.845687

http://researchrepository.murdoch.edu.au/21846/

Copyright © 1999 IEEE

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.
AN INTELLIGENT DATA ANALYSIS APPROACH USING SELF-ORGANISING-MAPS

Chun Che Fung*, Kok Wai Wong** and Doug Myers*

* School of Electrical and Computer Engineering
  Curtin University of Technology
  Bentley, Western Australia
  Phone (618) 9266 2575 Fax (618) 9266 2584
  E-mail: tfungcc@cc.curtin.edu.au

** School of Information Technology
  South East Metropolitan College
  Thornlie, Western Australia
  Phone: (618) 9267 7682 Fax: (618) 9267 7683
  E-mail: wongk@thornlietafe.wa.edu.au

Abstract: A neural-network-based data-analysis model for the prediction and classification of field data has many attractions. However, there are problems in ensuring the generalisation capability of the data analysis model, in measuring the similarity between the original training data and the new unknown data and in processing large data volumes. This paper reports the use of self-organising maps (SOM) to overcome these difficulties and illustrates the utilisation of this approach though applications in the agricultural, resource exploration and mineral processing areas.

1. INTRODUCTION

Self-Organising Maps (SOM) [1,2] have been recognised as an important tool for information processing and data analysis. Most SOM applications focus on their self-organising and clustering capabilities as SOM has the ability to organise the input vectors in an unsupervised learning mode.

Intelligent data analysis based on artificial neural networks has gained popularity in an increasing number of application areas, however, there are problems to be resolved. Part of the problem is due to the lack of correlation between the training data and the testing data. These problems may be overcome by applying clustering technique to ensure the similarity of the testing and training data prior to further analysis process. It is suggested that SOM offers that capability and so allow inferential data analysis procedures to be enhanced.

Data analysis is usually performed on a sample set of observations taken from some population [3]. In most practical situations, a sample is all that is available and it may provide incomplete information on the population. The objective of data analysis is nonetheless to extract maximum information from that sample, to exhibit reasonable interpolation capability and to be used for extrapolation purposes.

In this paper, SOM is used to examine three aspects of this data analysis problem. First, the problem of ensuring the generalisation capability of the data analysis model is investigated. SOM data splitting validation is proposed to solve this. After the interpretation model is established, SOM is then used to provide measurement of the similarity between the training data and the new unknown data. However, in cases where the available data is large, it is always safer to assume that the underlying function that the interpretation model needs to learn is difficult to realise. SOM can then be used in establishing modular models for overcoming this problem.

The proposed SOM solutions to the intelligent data analysis problem have been applied to problems relating to local industrial problems here in Western Australia. The proposed approach has been applied in the areas of agriculture, resource exploration and mineral processing. In particular, the SOM approach is used in classifying Australian wheat varieties [4,5], to aid a Backpropagation Neural Network (BPNN) in providing better and more accurate well log analysis [6] and in assisting a BPNN in providing parameter identification in hydrocyclone data analysis [7].

2. SOM DATA SPLITTING

Split-sample validation is the most commonly used method for estimating the generalisation capability of a BPNN using the early-stopping approach [8]. Here, a set of validation data that is not used in the training process is used to calculate the validation error. The stopping point in this method is suggested to be the point when the validation error starts to rise. This point also indicates that the generalisation ability of the network starts to degrade. When training starts, the errors for both data sets will normally reduce. After many training iterations, the validation error normally starts to rise although the training error may continue to fall. The BPNN...
The training process can be stopped at this point, as further training will result in overfitting.

As the generalisation ability of the BPNN is highly dependent on the validation data set, the splitting method used is important. However, there are no rules to suggest the best method. Nevertheless, the validation data set should demonstrate two characteristics: (1) the validation set should be statistically close to the training set, and (2) the validation error should indicate the generalisation ability of the final BPNN. It should be used as the stopping criteria during the training process.

To demonstrate the idea, if \( U \) is the universal sample space of all the cases of data to be processed by the network, then the training set \( TR \) should be statistically equal to \( U \):

\[
s(\text{TR}) = s(U) \quad (1)
\]

where \( s() \) indicates the statistical characteristics of a data set.

If \( s(\text{TR}) \) covers the complete sample space, the validation set \( (VA) \) and testing set \( (TE) \) should be statistically similar to the training set. That is:

\[
s(\text{VA}) \subseteq s(\text{TR}) \quad (2)
\]

\[
s(\text{TE}) \subseteq s(\text{TR}) \quad (3)
\]

with the condition: \( VA \cap TE = \emptyset \).

However, if a random approach of data splitting is used, this may result in a worst-case situation as illustrated by the following equations.

\[
s(\text{TR}) \subseteq s(U) \quad (4)
\]

\[
s(\text{VA}) \subseteq s(U) \quad (5)
\]

\[
s(\text{TE}) \subseteq s(U) \quad (6)
\]

and conditions:

\[
\text{TR} \cap VA \cap TE = \emptyset
\]

\[
s(\text{TR}) \neq s(\text{VA}) \neq s(\text{TE})
\]

In this case, the statistical characteristics of the three data sets are all mutually exclusive with the effect that the training set does not cover all the sample space. Subsequently, the validation and testing sets will not be able to give a fair indication of the generalisation ability of the network.

In the proposed SOM data-splitting technique [9,10], the available data are first classified into different clusters using unsupervised learning. If \( U \) is classified into \( C_1 \) to \( C_n \) clusters, then \( U \) can be written as:

\[
U = \{ C_1 + C_2 + C_3 + \ldots + C_n \} \quad (7)
\]

If the training data set is selected from each one of the \( n \) clusters and the rest are left for testing and validation, then the conditions on equation (2) and (3) are satisfied. In this case, the training set will cover all the desired underlying cases. The validation set and testing set are subsets from the clusters from which the training set is selected.

From the above, an important and crucial task is splitting the available data into training and validation sets with similar statistical characteristics. The training set should include information on what the BPNN should learn, and the validation set should act as a teacher to guide the BPNN such that the network will learn the correct function. As the BPNN is based on a training set to obtain the underlying knowledge, therefore understandably, it should contain more data than the validation set.

When obtaining the training set, there will be some environmental factors that affect the measurements. As a result, it is not possible to have an exact function that describes the relationship between \( X \) and \( Y \). However, a probabilistic relationship governed by a joint probability law \( P(\nu) \) can be used to describe the relative frequency of occurrence of vector pair \( (X_n, Y_n) \) for an \( n \) training set. The joint probability function \( P(\nu) \) can be further separated into an environmental probability function \( P(\mu) \) and a conditional probability function \( P(\gamma) \). Thus the probability function may be expressed as:

\[
P(\nu) = P(\mu)P(\gamma) \quad (8)
\]

The environmental probability function \( P(\mu) \) describes the occurrence of the input \( X \). The conditional probability function \( P(\gamma) \) describes the occurrence of the output \( Y \) based on the given input \( X \). A vector pair \( (X, Y) \) is considered as noise if \( X \) does not follow the environmental probability function \( P(\mu) \), or the output \( Y \) based on the given \( X \) does not follow the conditional probability function \( P(\gamma) \).

The rule of splitting the available data into training and validation sets is that the training set should be statistically similar to the whole sample space. The validation set should also be statistically similar to
The training set, as it has to act as a teacher. This rule suggests deploying the SOM algorithm. SOM can be used as a nonlinear probability density function projection on the two-dimensional map. Therefore, in each node, the probability density function of the input vectors being mapped onto it should have a similar probability density function. This also implies that the input vectors that are mapped onto the same node should have similar relative occurrences as denoted by $P(X)$. This $P(X)$ is similar to the environmental probability function $P(Y)$ in equation (8). From the analysis of equation (8), the role of training the BPNN can be said to be a search for the conditional probability law $P(Y|X)$. The formulation of the $P(X)$ here has to be extended. Instead of mapping just the input vector $X$, both the input vector $X$ and target vector $Y$ are used in the learning of the SOM. A joint probability between $X$ and $Y$ is assumed and is denoted as $P(X,Y)$. It can be further expressed as:

$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$ (9)

As equation (9) is similar to equation (8), it implies that the joint probability function density of a SOM is directly related to the joint probability function. With this, it can also be realised that the joint vectors of $X$ and $Y$ falling in the same node should have very similar statistical characteristics.

The methodology for satisfying the splitting data rule has been formulated. The $n$ available data sets that consist of $X$ input vectors and $Y$ output vectors are first used to train the SOM. After the map has been trained and individual quantisation errors have been generated, selection can be made. A data set is selected as validation data if it has a small quantisation error as compared to the other data sets in the same node. This will ensure that the validation set is a sub-set of the training set. However, for cases where there is only one data set in that node, it will be left in the training set. This is to ensure that the training set covers the whole sample space of the available data and to ensure that the training set is always larger than the validation set. After all the available data has been split into training and validation sets, the BPNN can start to learn and the process is stopped by using the early stopping validation technique.

3. STATISTICAL COMPARISON

The issue of evaluating an indication of the confidence level for the predicted properties in unknown cases is considered. The objective is to provide an indication of the usability of the trained interpretation model when it is used for any new data that may be statistically different from the training data. In cases where the indication shows that the unknown new cases are very different from the trained cases, the predicted results cannot be totally trusted. This will be useful in providing a confidence indication to the analyst.

To perform the confidence level indication, a SOM is used to classify the training data to a pre-defined two-dimensional map. At the completion of this unsupervised learning stage, an average quantisation error is generated that gives a measure of the fitness of the training data in the resultant clusters. Any subsequent unknown input data to be applied to the prediction model are now mapped onto the trained SOM. An average quantisation error is generated that measures the statistical similarity between the unknown data set and the trained map. Comparing the average quantisation errors of the training data set and the unknown data set indicates the similarity. These values suggest to the users how similar or different are the trained and predicted data sets. It provides the user an assurance of the predicted output from the BPNN interpretation model.

4. SOM-BASED MODULAR NEURAL NETWORKS

When there is a large volume of available training data, the Modular Neural Network (MNN) is proposed for the purpose of analysis. The MNN is based on the Self-organising Map (SOM) [1,2]. Learning Vector Quantisation (LVQ) [11] and BPNN [12]. However, an MNN [13] can only be used when the available training data is large. As compared to the usual BPNN approach with its single network, the MNN employs a number of sub-networks. SOM and LVQ are used to classify the raw data. Several BPNNs corresponding to the number of classes obtained from the SOM are then trained for the purpose of function prediction. Since the number of data to be handled by each sub-network is relatively small, the training time is significantly shortened. As the data that falls into the same sub-network will have similar characteristics, this effectively reduces the complexity of the function that the ANN needs to learn.

An MNN is arranged into two major sections. The first focuses on classification. The second covers the prediction results of the MNN.

An ANN is capable of learning any non-linear function from the available training data. However,
if the available training data is large and complex like that of Figure 1, the underlying function may be too complex for a single ANN to cope with. This may be overcome by modularising the task as shown in Figure 2. If the data can be first classified before the ANN learning process, then the functions handled by each sub-section of a modular structure will be very much simpler compared to the whole training data set. Consequently, the function should be able to learn in a shorter time and better prediction results obtained.

There are several ways of performing classification. However, a technique that can be done automatically and transparently to a human analyst is most desirable. SOM is selected as the best classification approach in designing this MNN as it uses unsupervised learning. It has the ability to learn and organise information without being given correct outputs for its inputs. A SOM network consists of two layers of nodes. Each output node is computed with the dot product of its weight vector and the input vector. The result will reflect the similarity between the two vectors. At the end of the training, the SOM will make use of its learning ability to arrange the available training data into a different cluster. After the SOM classification of the training data, supervised learning in the form of LVQ is employed to fine-tune the classification process such that it could be used for any unknown input data. LVQ is closely related to SOM, but uses the given classification information to define the class regions in the input space. In this case, SOM and LVQ will learn from the data and perform their own classification process. This removes the need for any human intervention in sorting data.

Since it is known that a relationship exists between the input vectors and the characteristics within the output data, an approach to determine such mapping has been formulated. SOM is first applied to classify the input and output data. The classes obtained are then used to label the input vectors. The input vectors coupled with the output class labels are then applied to the LVQ algorithm. A number of BPNN networks corresponding to the number of classes obtained from the SOM are trained. After the classification process, the data are then fed into different BPNN according to the characteristics of the data. In this way, training of the BPNN is expected to take shorter time.

The process is summarised in the following steps:

1. Normalise the input and output data.
2. Determine the number of classes required and apply the SOM algorithm to the input and output vectors.
3. Label the input vectors according to output classifications from Step 2.
4. Apply the LVQ algorithm to the normalised inputs and establish the network.
5. Prepare to train a few BPNNs, each network corresponds to a class from Step 2.
6. Train each BPNN using the SOM data splitting validation approach mentioned in the previous section.

Once the network is trained, new input data can be classified by applying the normalised data to the network.
5. APPLICATION EXAMPLES

The following application examples are used to illustrate the usefulness of the proposed SOM data analysis approach. The problems have been investigated by the authors over the past years. SOM has enhanced the performance of the data analysis model and improved the results in these cases. Further details and descriptions of the problems can be found in the referred papers.

5.1 Agriculture – Wheat Variety Recognition

There is an increasing need for marketing and other reasons to be able to identify grain varieties. This is particularly true in verifying the proportions in an admixture. With the increasing number of registered wheat varieties and their genetic similarity, it has become extremely difficult to classify Australian wheat varieties by visual inspection. Chemical analysis and statistical pattern recognition may be used but each method has its drawbacks [4,5,14].

SOM has been used to perform this classification task with some success [14]. The test outcomes indicate that SOM is able to classify up to two or three wheat varieties with a maximum accuracy of 96.5% and 88% respectively.

When SOM is used in classifying wheat varieties, the input vectors are sample sets shape features such as rays or aspect ratio measurements. After the network has been trained with a selective set of data, the output node which gives the highest response to a specific class of wheat variety within the input training data sets is labelled to that class of wheat. When an output node gives the same response to two or more wheat varieties, its neighbouring nodes are taken into account and the majority rule is applied to determine the labelling for the node.

5.2 Resource Exploration – Log Data Analysis

Two key issues in the reservoir evaluation of petroleum exploration using well log data are the characterisation of formation and the prediction of petrophysical properties such as porosity, permeability and volume of clay [6]. While a set core data gives an accurate picture of the petrophysical properties at specific depths, it is a lengthy process and great expense is incurred in obtaining such data. Hence only limited core data are available at selected wells and depths. The objective of well log data analysis is to therefore establish an accurate interpretation model for the prediction of the petrophysical properties for uncored depths and boreholes around that region.

An accurate prediction is essential to the ultimate determination of the economic viability of the exploration and the production capacity of the particular well or region.

BPNN is an emerging technology in this field. However, the raw application of the BPNN may not provide reliable well log analysis. The three problems raised in the beginning of this paper are the major concerns for the application of BPNN techniques in this field. However, with the application of SOMs for data analysis in the manner outlined, these concerns can be eliminated. Research results indicate this approach has led to an increase in the reliability of the prediction [13,15].

5.3 Mineral Processing – Hydrocyclone Parameters Identification

While hydrocyclones are used in the mineral processing industry for particle separation, an exact model of a hydrocyclone is difficult to derive due to their highly non-linear characteristics and the large number of parameters involved [7]. As efficient operation of a hydrocyclone is important in improving system performance, it is essential that the model should be able to provide non-linear mapping between the multi-dimensional system inputs and outputs. Although the collection of the data in this field is different compared to well log data analysis, they both fall into the same category of inferential data analysis problem. Therefore, the methodology used in the previous section can be duplicated and used in this field. Research results show that SOM methods can also increase the prediction reliability [16].

6. CONCLUSION

This use of SOMs offers advantages in framing the interpretation model in intelligent data analysis. SOM-based intelligent data analysis approach in three significant applications areas illustrated the value of the method. Together with other data-analysis tools, SOM can provide a useful approach to improve the performance of the data-analysis process.

7. REFERENCES


