Computer aided identification of biological specimens using self-organizing maps

E. J. Dean¹, A. P. Engelbrecht² & A. Nicholas³

¹Dept of Information Technology, Durban Institute of Technology, SA
²Computer Science Department, University of Pretoria, SA
³Botany Department, University of Durban-Westville, SA

Abstract

It is often necessary or desirable that biological material be identified. However, given that there is an estimated 10 million living organisms on Earth, the identification of biological material can be problematic and consequently the services of a taxonomist specialist are often required. If such an expert is not readily available it is necessary to attempt an identification using an alternative method; but some of the alternative methods available are unsatisfactory or can lead to a wrong identification. One of the most common problems encountered when identifying specimens is that important diagnostic features are often not easily observed, or may even be completely absent. A number of techniques can be used to try to overcome this problem, one of which, the Self Organizing Map (or SOM), is a particularly appealing technique because of its ability to handle missing data. This paper explores the use of SOMs as a technique for the identification of indigenous trees of Acacia in KwaZulu-Natal, South Africa. The ability of the SOM technique to perform exploratory data analysis through data clustering is utilized and assessed, as is its usefulness for visualizing the results of the analysis of numerical, multivariate botanical datasets. The SOM’s ability to investigate, discover and interpret relationships within these datasets is examined, and the technique’s ability to identify tree species successfully is tested. The tests performed so far have provided promising results and suggest that the application of the SOM to the problem of identification could provide the breakthrough in computerized identification for which botanists have long been hoping.
1 Background

Information about plants and their properties may be accessed through their scientific names provided that these names are known. Therefore it is often necessary or desirable that biological material be identified to species level or below. Identification involves recognizing, selecting or associating closely the characteristics of an unknown object with another object of known identity. For instance the macroscopic identification of a tree is often done by taking material (leaf, bark, flowers, seed or root) from the unknown tree and comparing its characteristics with a list of characteristics of known trees. In addition, the location, climatic conditions and general form of a tree should be taken into consideration. As there is often a great deal of uncertainty regarding the recognition of identifying features in the sample material, the process of identification can be problematic. This is due to a number of factors, the most significant of these being the natural variation that is commonly exhibited by all living organisms, and the fact that critical diagnostic features are often absent or not easily observed. Owing to these difficulties the process of identification usually requires the services of a taxonomic expert. However, if such an expert is not readily available, it might be necessary to attempt an identification using an alternative method.

One possible method could involve the utilization of SOMs.

2 Current solutions

Historically, most identification of a botanical nature has been done by means of sequential single-entry keys (e.g. dichotomous keys) in books and articles; such as is found in Leistner [1]. These keys are made up of contrasting character statements that require the identifier to make a comparison with the specimen being keyed and then choose the right statement Radford et al [2]. The strict sequential nature of these keys, working on a strict order of character elimination, does not allow for easy backtracking or lateral progression. With these types of key, options for identifying an unknown specimen are limited; and the keys make no allowance for ambiguous or atypical data, or for the absence of characters in the material being identified.

Although these keys may provide an answer to an expert, a layman cannot easily or conveniently apply the key successfully, and, on occasions, even trained botanists fail. Many keys only start to ‘work properly’ when the user begins to understand the way in which the key’s compiler has defined any subjective characters involved. For example, “leaf shape obovate” can mean different things to different users and can lead to the choice of the wrong option in a key. It is because of this, and the use of specialized terminology, that the use of sequential keys is one of the main difficulties experienced by students and amateur botanists alike Stace [3]. If an incorrect step is taken along the path through the key (whether by error, misinterpretation, ambiguity or aberration of the specimen itself) the identification process is likely to fail. The chance of going wrong also increases as the number of steps in the key increases.
In order to increase the chance of successful identification some sequential keys employ reticulations. Attempts have been made to produce multiple-entry access, order-free keys or polyclaves (e.g. punch or clip card keys and tabular keys). Although these keys increase the chance of a successful identification they are not infallible and can only handle small datasets. They do have a major advantage in that they allow the user freedom to choose any character in any sequence, thus avoiding the rigid format of sequential keys Jones and Luchsinger [4] and allowing for the possible identification of incomplete specimens.

More recently computerized keys have been produced which allow for identification and retrieval of information, but again existing systems tend to be successful only in cases that provide a clear either/or alternative and do not cater for overlapping or vague information. Few computer systems cater for cases that are not an exact fit to the stored data, i.e. fuzzy cases that do not fit exactly in a defined category. The major problem with all of these identification programs is that they do not utilize the full potential of the computer and in some cases are little more than a computerized polyclave.

An improvement in information retrieval systems that has been used for plant identification involves the employment of an Expert System (ES). Some existing systems, such as LucID (http://www.lucidcentral.com/), do utilize an ES. However, none of the systems that have been investigated Dean [5] takes advantage of the full power that modern computer hardware, software and AI techniques could offer. A problem with ESs is that they are static and not data driven, and are consequently only as good as the rules that are obtained from a human expert and then stored in the ES’s knowledge base. It is because of the need to allow for fuzzy and overlapping biological datasets that advanced computer techniques would greatly enhance the identification process. In particular, the ability of computers to mimic the human thought processes closely (i.e. artificial intelligence) needs to be fully utilized.

3 Introduction to the SOM

The SOM is a robust form of an unsupervised Neural Network (NN), which was first introduced by Professor Teuvo Kohonen [6]. Kohonen’s ideas were motivated by the study of the self-organization characteristics of the human cerebral cortex, which showed that topologically ordered maps could be used to represent patterns of relationship induced from or sensed within the inputted data or sensory input signals Engelbrecht [7]. The SOM is a non-parametric regression technique that converts multi-dimensional input data into a lower dimensional output abstraction. The principal goal of the SOM is to transform the pattern of an incoming signal of arbitrary (usually higher) dimension into a discrete 1-, 2- or 3-dimensional map. It performs this data transformation while preserving the topological order and correlated integrity of the input data Haykin [8].

The creation of a SOM requires two layers of processing nodes [6, 7, 8, 9]: the first is an input layer containing processing nodes for each component in the input vector; the second is an output layer of processing nodes that is associated
with those of the input layer. The number of processing nodes in the output layer is determined by the programmer and is based on the envisaged shape and size of the map, and on the number of independent inputs. There are also algorithms that can automatically grow a map to an “optimal size”. In a SOM network there are no hidden layers or hidden processing nodes.

The SOM is a competitive NN. Thus when the input is presented to the network each output unit competes to match the input pattern. The output that is closest to the input pattern is declared the winner. The weights of the winning unit are then adjusted, i.e. moved in the direction of the input pattern by a factor determined by the learning rate. This is the basic nature typical of competitive NNs.

The SOM differs from other unsupervised algorithms in that it creates a topological map based on patterns of correlated similarity. This map is dynamic and continually adjusts itself not only according to the winner’s weights, but also the weights of the output nodes in the neighbourhood of the winning node. Thus the output nodes that start with randomised weight values slowly align themselves with each other based on perceived patterns of similarity in the input data. When an input pattern is presented, a neighbourhood of nodes responds to the input pattern to see if it can match it.

4 SOM applications

Because of its ability to establish correlated patterns the SOM has diverse applications. Its use embraces many fields from physics and chemistry to biomedicine and psychology. More details and many references for these applications can be found in Kohonen [10]. In addition SOM techniques have been used to study traditional AI problems, such as the travelling salesman problem, and have been combined with other techniques including fuzzy logic, genetic algorithms and principal component analysis. Financial applications of SOMs are numerous and include: financial projections, estimation of market values, rating of financial instruments, selection of investments, risk and portfolio management, identification of market potentials, customer tracking and determination of consumer preferences Deboeck and Kohonen [9]. The SOM has been used extensively in engineering solutions, such as telecommunication problems, signal processing, radar measurements, process control and electronic-circuit design [10].

A major application of SOMs has been for data processing and analysis [9]. This study utilises this aspect of SOM application: in particular the SOM’s ability to compress data, retrieve information and extract knowledge in order to identify and classify plant species is utilized.

5 Why SOMs are applicable to biological identification

One problem that is encountered in biosystematics is the large amount of data available for constructing classifications. This consists of data, some of which are pertinent for establishing relationships based on inherited similarity, and
some of which are irrelevant, or, in the case of homoplasy, even misleading. A spreadsheet would suffice for simple comparisons and overviews, and for the elucidation of patterns of similarity and dissimilarity of a limited number of taxa and for a limited number of characteristics. However, as the number of taxa or complexity of the information increases the use of a more proficient and accurate tool becomes essential for clustering and visualization. Due to its multifarious nature, the relationships and structures/patterns of datasets are often hidden and are therefore not explicit. As a result, they may defy exposition by classical statistical methods based on linear principles and *a priori* assumptions. A different approach is needed for identification of biological material that normally requires the services of a taxonomic expert able to comprehend the complex relationships of organisms in their domain area. NNs in the form of SOMs provide a suitable tool for this, and the technique analyses the different characteristics of the input data and groups together specimens with similar characteristics. SOM technology is data driven rather than knowledge-driven and provides a model-independent alternative to both traditional methods (which have the limitations mentioned above), and to other methods, such as other types NNs (which are also data driven) but which impose high demands with respect to tuning the algorithm. Being a data driven model the SOM also avoids some of the problems associated with finding the appropriate underlying reasoning of an expert.

Unlike traditional methods, which are largely based on the subjective assessment of relationships, the SOM is a numerical method. As a result, it is able to treat data statistically and can represent graded relationships objectively. Unlike orthodox taxonomy that makes *a priori* assumptions about the importance of certain characters (which are then said to be weighted) in order to establish relationships, SOMs make no assumptions about the distribution or importance of the data beforehand. As a result, it is possible that the SOM may uncover quite unexpected patterns of relationship between the selected input data Kaski [11].

One of the SOM’s greatest strengths is that it can still obtain results when there are missing data values. If the number of characteristics taken into account is fairly large then an appreciable number of data values may be missing without making the results meaningless. This is particularly important with respect to the identification of biological material, as there is often a considerable number of missing diagnostic characters.

A SOM has the capability to generalize. This means that the network can characterize input data even if it has never encountered them before. A new input is assimilated with the map unit to which it is most closely mapped or correlated. In addition, based on a trained map, even input vectors with missing data can be used to predict the values of the missing data [9].

A SOM can be used to investigate the differences and the degree of difference between clusters of taxon-related-data and, in the process, it is possible that some new feature that distinguishes between taxa might be discovered. The SOM’s purpose is to distinguish objects and this is what the process of identification particularly requires.
6 How the SOM works

The SOM is a form of NN that can provide an objective way of grouping data using self-organizing networks of artificial neurons. The neurons are arranged in a rectangular matrix with each neuron connected to its four nearest neighbours on the grid. The number of neurons in the matrix is chosen by the programmer but should be less than the number of training patterns and preferably also less than or equal to the number of independent training patterns. Using the input training data, which is stored in the form of vectors, the SOM creates a two-dimensional matrix (which resembles a landscape) onto which statistically correlated patterns of similarity will be mapped. To achieve this, each neuron in the network is associated with a different weight vector. These weight vectors may be initialised in various ways [7]. One of these methods of initialisation needs to be selected and implemented. The effect of initialisation is that each neuron on the map is assigned and stores a set or range of weights, each one of which corresponds to one of the values in the input training data vector. Each input vector in turn is then presented to each neuron in the network. The neuron associated with the weight vector closest to the input vector is found and is called the winner. Two of the common ways of determining the winning neuron are to calculate either the Euclidean norm or the Dot Product norm. The Dot Product norm is often used with binary data but in this study multivariate data is used and so the former norm is employed. The Euclidean norm is calculated by computing the mathematical distance between the given input vector and each weight vector of a neuron. The neuron associated with the weight vector that is most similar (i.e. has the minimum distance) to the input vector is the declared the winning neuron.

A general algorithm for the SOM method can be summarized as:

```plaintext
standardize input data:
initialise weight vectors;
for each iteration (t = 0, 1, 2, ...)
    for each input vector (i = 0, 1, 2, ...)
        for each neuron weight vector (j = 0, 1, 2, ...)
            for each vector component (k = 0, 1, 2, ...)
                Dist_{i,j} = \sqrt{\sum_k (input_{i,k} - weight_{j,k})^2}
next k;
next j;
find winner;
adjust winner's neighbourhood;
next i;
next t;
display output;
```
For each sample input vector, \( \mathbf{x} \), the winning neuron weight vector, \( \mathbf{m}_c \), is selected such that:

\[
\| \mathbf{x}(t) - \mathbf{m}_c(t) \| = \min_j \| \mathbf{x}(t) - \mathbf{m}_j(t) \|, \quad \forall j
\]

The neuron associated with the weight of the vector closest to the input vector is the winner. The weights of the winner and the winner’s neighbours are then adjusted so that they are closer to the input vector. The formula by which the neighbourhood weights are adjusted is:

\[
\mathbf{m}_j(t + 1) = \mathbf{m}_j(t) + \alpha(t) \{ h_{c,j}(t) (\mathbf{x}(t) - \mathbf{m}_j(t)) \}
\]

where \( \alpha(t) \) is the learning rate factor which decreases steadily for each iteration and \( h_{c,j}(t) \) is the neighbourhood function, which is a decreasing function of the distance between the \( j \)th and \( c \)th neurons on the neural grid [9], [10] and [12].

The size of the winner’s neighbourhood varies throughout the training process. Initially all of the neurons in the network may be included in the neighbourhood of the winner, but as training progresses the size of the neighbourhood is decreased. After each presentation of the complete training dataset the neighbourhood is decreased until it includes only the winning neuron. The amount by which the neurons in the winner’s neighbourhood are allowed to adjust their weights is also reduced throughout the training period. Eventually, the vectors surrounding the winning neuron will, to varying degrees, also closely represent the input vector. The learning process continues (at a decreasing rate) until the map reaches an accurate result (i.e. an accurate representation of the pattern exhibited by the training data). The degree of error of the map can be mathematically calculated and this can then be used as an indication of the degree of its accuracy. Training can stop when this error is statistically insignificant or, alternatively, after a maximum number of iterations is exceeded. In the end each neuron will have a unique weight range and be surrounded by a neighbourhood of neurons with similar weights.

Once the SOM is produced the topological structure of the input matrix will be maintained such that if two input vectors are close to each other in the input space, the corresponding weight vectors will be close to each other on the map. Because of this clustering of vectors and neurons (which are often separated from each other by obvious discontinuities) it becomes possible to identify boundaries on the map. These boundaries define different vector clusters nested around a neuron. These clusters, as defined by their calculated vectors, consist of input variables with similar characteristics. In this paper these clusters correlate to Acacia species and in turn clusters representing similar Acacia species are situated close to each other on the output map.

Once trained in this way, and a more or less stable but optimal landscape of similarity established, the SOM is now ready to use as a deductive or predictive tool and the test input vectors can be presented to each neuron in the network.
7 The SOM tool used

The SOM tool used in this study is the state-of-the-art Eudaptics’ Viscovery SOMine®, version 4.0. Viscovery is a data mining product for the visual analysis and exploration of numerical, multivariate datasets. It can be used to discover, analyze and interpret relationships within the data and gives a non-linear representation of the implicit relationships.

Viscovery uses a batch version of the SOM training rule where the weight values are updated only after all patterns have been presented, rather than updating weights after each pattern presentation. Using the Batch-SOM no learning rate is required and use of a growing SOM grid makes the training phase faster [9].

The version of software used allows for datasets consisting of up to 50 variables (characteristics) of a maximum of 10000 records. For this investigation 6 records on each of 20 species (giving 120 records with 49 variables in total) were used for training the network. 100 records were used for testing the network (five on each of 20 species). The training data and the test data were taken from reliable sources of expertise [13, 14, 15, 16].

The input datasets were systematically converted to a numerical representation. Some of the data was represented in binary format, with ‘1’ indicating that a characteristic was present and a ‘0’ indicating that the characteristic is never present. Botanically such characteristics are said to be ‘all or nothing’. Multivariate data that formed a morphocline (grading of one character state to another) lent itself to grading. For example, where a particular specimen could be different colours or shades of colour, a number from 0.1 – 0.9 was used to represent the different information with 0.1 for a light colour or shade and 0.9 for a dark colour or shade. This grading helped to reduce the number of characteristics that would otherwise have been necessary. It also helped to scale the input data so that large values did not unduly influence the output of the map Kaski [12].

8 Experimental results

For this study the species of the genus Acacia found in the province of KwaZulu-Natal South Africa were used to test the ability of the SOM to identify down to species level. The genus Acacia, with circa 900 species distributed mainly in Africa and Australasia Germishuizen [18], is biologically complex and identification can be problematic. As many species are used for timber, gum, dyeing and tanning, perfumes and horticulture their correct identification is important at both societal and economic levels Willis [19]. Acacia species are also an important constituent of the savannas. The province of KwaZulu-Natal with over 750 tree species is home to almost two thirds of the total number of trees found in southern Africa, and can be considered remarkable at the sub-continental level because of this. This province contains 20 indigenous Acacia species Pooley [14].

Combined datasets (thorns, pods and flowers) were used for the training of the network. Experimental training cycles were performed with 20, 50, 80 and...
100 nodes; with 50 nodes promising results were obtained. The input data consisted of 20 independent species and a total of 120 datasets. Therefore, it would have been better if a maximum of 20 nodes had been used. A hexagonal lattice map was used for ease of visualization of the outputted results and a rectangular frame was used to achieve a stable orientation for the data space. The actual size of the map used was 9 x 7 neurons and a neighbourhood of 7 neurons was specified. The results obtained using the combined input data for thorns, flowers and pods are shown in Figure 1.

Figure 1: SOM Training Map showing species clustered according to their overall similarity, as based on the inputted training data.

In Figure 1 the clusters are numbered and the *Acacia* species found at these nodes are shown below.

<table>
<thead>
<tr>
<th>Number</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>A. caffra</em></td>
</tr>
<tr>
<td>2</td>
<td><em>A. xanthophloea</em></td>
</tr>
<tr>
<td>3,7</td>
<td><em>A. robusta</em></td>
</tr>
<tr>
<td>4</td>
<td><em>A. luederitzii</em></td>
</tr>
<tr>
<td>5</td>
<td><em>A. karroo</em></td>
</tr>
<tr>
<td>6</td>
<td><em>A. sieberana</em></td>
</tr>
<tr>
<td>7,3</td>
<td><em>A. robusta</em></td>
</tr>
<tr>
<td>8</td>
<td><em>A. ataxacantha</em></td>
</tr>
<tr>
<td>9</td>
<td><em>A. burkei</em></td>
</tr>
<tr>
<td>10,14</td>
<td><em>A. davyi</em></td>
</tr>
<tr>
<td>11,18</td>
<td><em>A. kraussiana</em></td>
</tr>
</tbody>
</table>

Once the NN had been trained the test data was run to see if the network could correctly identify 100 specimens in which not all the diagnostic features were present. The results are shown in Figure 2.
In Figure 2 specimens of A. ataxacantha, A. borleae, A. brevispica, A. burkei, A. caffra, A. davyi, A. karroo, A. nigrescens, A. nilotica, A. robusta, A. sieberana, A. tortilis and A. xanthophloea have all been mapped to the right clusters resulting in 100% correct identification. The identification of specimens of A. luederitzii, A. senegal and A. swazica was 80% accurate. For A. gerrardii, A. grandicornuta and A. Schweinfurthii the correct identification was occurred in 60% of the tests. This means that out of 100 test specimen 91 were correctly identified.

![Figure 2: SOM showing test results.](image)

In Figure 2 the following abbreviations have been used for the test specimens:

- atx A. ataxacantha
- lutx A. luederitzii
- botx A. borleae
- ngtx A. nigrescens
- btxt A. brevispica
- nltx A. nilotica
- butx A. burkei
- rotx A. robusta
- catx A. caffra
- swtx A. Schweinfurthii
- datx A. davyi
- setx A. Senegal
- getx A. gerrardii
- sitx A. sieberana
- gtxt A. grandicornuta
- sztx A. swazica
- ktxt A. karroo
- totx A. tortilis
- krtx A. kraussiana
- xatx A. xanthophloea

where x is the number of the specimen.
For the purposes of this experiment it was preferred that the software determine the order of importance of the variables, and, therefore, the results of identification were obtained without the programmer trying to influence the output by assigning priorities to characteristics. That is, the results were obtained without prior human intervention and assumptions. At this stage no attempt has been made to look for or correct the causes of misidentification. This will be done in future studies. Table 1 presents a summary of the classification errors.

<table>
<thead>
<tr>
<th>Cluster No</th>
<th>Cluster</th>
<th>Error</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Top left</td>
<td><em>A. schweinfurthii</em> (swt17 &amp; swt9) identified as <em>A. kraussiana.</em></td>
<td>2/5</td>
</tr>
<tr>
<td>4</td>
<td>Top left</td>
<td><em>A. senegal</em> (set21) identified as <em>A. luederitzii.</em></td>
<td>1/5</td>
</tr>
<tr>
<td>7</td>
<td>Top center</td>
<td><em>A. grandicornuta</em> (grt11, grt13); <em>A. gerrardii</em> (get14, get15) identified as <em>A. robusta.</em></td>
<td>2/5</td>
</tr>
<tr>
<td>5</td>
<td>Bottom right</td>
<td><em>A. swazica</em> (szt9) identified as <em>A. karroo</em></td>
<td>1/5</td>
</tr>
<tr>
<td>8</td>
<td>Left center</td>
<td><em>A. luederitzii</em> (lut18) identified as <em>A. ataxacantha.</em></td>
<td>1/5</td>
</tr>
</tbody>
</table>

Experimental training cycles were also performed with separate datasets, i.e. the network was first trained with the thorn data separately and then with the flower and pod data in that order. On the resulting maps there were several species aligned to the same neuron. The test data thus resulted in more identification errors than were obtained when the combined datasets were used for training. These results are not illustrated or tabled here. Using the combined dataset to train the network had the advantage that only one map needed to be analysed and interpreted, and different clusters did not appear on each of the maps.

9 Discussion

The datasets used for training the SOM in this study had some missing values (approximately 47%) and the testing dataset have a considerable number of missing values (approximately 65%), as would be consistent with what often happens in reality when trying to identify biological specimens. For example, both pods and flowers are important characteristics for identifying trees, but they are seldom found on a tree at the same time. When the percentage of missing data is taken into consideration the results of these preliminary tests are very encouraging. However, it is expected that when more input data is collected, and the size of the training dataset is increased, the results will improve. Also if principal component analysis was performed in order to determine the important
identifying characteristics, and only those characteristics were selected for the training set, a wider range of other characteristics could be included (for example, leaf and bark characteristics which so far have been ignored because of the restriction to 50 characteristics imposed by the version of Viscovery software used) and this should improve the successful recognition of the different species.

It was hoped that the results would have improved when the network was trained with separate datasets, interestingly this was not the case.

Living organisms, especially higher organisms, are incredibly complex and defined by thousands of macro- and micro-morphological or anatomical characters, each of which can be translated into numerical vectors. In this study the restrictions, imposed by the Viscovery software version used, resulted in only 50 variables being available to represent each species. This limited the efficacy and usefulness of the SOM technique in this application. This can, to a certain extent, be overcome by using methods that select the principal components or diagnostic characters for identification. In other words a priori selection is abandoned in favour of the selection of characteristics known to be of diagnostic or evolutionary importance. In taxonomy this is known as weighting characters.

It is planned that the next stage in this research will involve investigating and improving the representation of the data; and the use of principal component analysis to select the most significant diagnostic variables used in the identification process.

Acknowledgements

This research is made possible by the support received by the National Research Foundation of South Africa.

References


