Topographic Mappings and Feed-Forward Neural Networks

MICHAEL E. TIPPING

Doctor Of Philosophy

THE UNIVERSITY OF ASTON IN BIRMINGHAM

February 1996

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without proper acknowledgement.

THE UNIVERSITY OF ASTON IN BIRMINGHAM

Topographic Mappings and Feed-Forward Neural Networks

MICHAEL E. TIPPING

Doctor Of Philosophy, 1996

Thesis Summary

This thesis is a study of the generation of topographic mappings — dimension reducing transformations of data that preserve some element of geometric structure — with feed-forward neural networks.

As an alternative to established methods, a transformational variant of Sammon's method is proposed, where the projection is effected by a radial basis function neural network. This approach is related to the statistical field of multidimensional scaling, and from that the concept of a 'subjective metric' is defined, which permits the exploitation of additional prior knowledge concerning the data in the mapping process. This then enables the generation of more appropriate feature spaces for the purposes of enhanced visualisation or subsequent classification.

A comparison with established methods for feature extraction is given for data taken from the 1992 Research Assessment Exercise for higher educational institutions in the United Kingdom. This is a difficult high-dimensional dataset, and illustrates well the benefit of the new topographic technique.

A generalisation of the proposed model is considered for implementation of the classical multidimensional scaling (CMDS) routine. This is related to Oja's principal subspace neural network, whose learning rule is shown to descend the error surface of the proposed CMDS model.

Some of the technical issues concerning the design and training of topographic neural networks are investigated. It is shown that neural network models can be less sensitive to entrapment in the sub-optimal global minima that badly affect the standard Sammon algorithm, and tend to exhibit good generalisation as a result of implicit weight decay in the training process. It is further argued that for ideal structure retention, the network transformation should be perfectly smooth for all inter-data directions in input space.

Finally, there is a critique of optimisation techniques for topographic mappings, and a new training algorithm is proposed. A convergence proof is given, and the method is shown to produce lower-error mappings more rapidly than previous algorithms.

Keywords: Information Processing, Feature Extraction, Sammon Mapping, Multidimensional Scaling, Research Assessment Exercise

Contents

List of Figures

Chapter 1

Introduction

> Where is the knowledge we have lost in information?
>
> **T.S. Eliot —** *The Rock* (1934).

It is often said that we are living in the *information age*. The technological revolution of the latter half of the twentieth century has placed previously undreamt-of quantities of information at our fingertips. The maturity of the digital computer with its continual exponential growth in both power and storage capacity, allied with the emergence of multi-media and the dramatic recent expansion of global connectivity known rather grandiloquently as the 'digital information super-highway', offers unprecedented access to vast amounts of data, all over the world, for millions of users.

However, as the ease of access to information increases, so, inevitably, do the accompanying difficulties in its interpretation and understanding. It is very easy to become overwhelmed by the sheer volume available. One particular on-line information resource is the data collected for the 1992 Research Assessment Exercise for higher educational institutions in the United Kingdom. Even the small fraction of this large dataset that is studied later in this thesis contains over thirty-two thousand numbers and there is clearly little to be gained by study of the naked data alone; the knowledge remains locked away, impenetrably it may sometimes seem, behind the anonymous digits.

The key, therefore, lies in *information processing*. Whether for the purposes of visualisation, exploratory analysis or for subsequent computation, it is essential that the information be manipulated into a form which facilitates its ultimate use. The emphasis has thus shifted from the problem of the acquisition of information, to that of its exploitation for the purposes of deriving useful knowledge.

The type of information, or data, that will be considered in this thesis is that in *numeric* form. Data will, characteristically, be comprised of a set of measurements concerning a corresponding set of objects. For example, Fisher's familiar 'Iris' dataset contains measurements of sepal length, sepal width, petal length and petal width for fifty samples of each of three different varieties of iris flower. The previously mentioned Research Assessment data comprises nearly one-hundred-and-fifty different variables for over four thousand invidual departments from every university in the United Kingdom — variables which describe such quantities as the number of staff, the number of Ph.D. students and the number and value of research grants. This numeric form lends itself naturally to a vector-space interpretation, such that in the Iris dataset, each set of four sample measurements can be considered a distinct vector in four-dimensional space. In general, then, for all such datasets with p different fields, the data may be considered as a collection of similar point vectors in a p-dimensional space.

Given this interpretation, information processing can often be intuitively posited as a *dimensionality* *reduction* problem. For visualisation, human perception is attuned to two- or three-dimensional images, and real-world numeric data, which is generally of naturally high dimension, must be processed into more readable forms without loss of salient detail. In data-modelling applications, the sizable number of variables implied by high-dimensional data can be seriously disadvantageous. Sensible pre-processing of the data before the model-building stage can help alleviate these problems.

This thesis concerns one particular approach to extracting knowledge that is concealed within information. It is an investigation into the use of feed-forward neural networks to effect a particular class of dimension-reducing information-processing strategies — *topographic mappings*. Exactly what a topographic mapping is, why a neural network should be used to produce one and what is the relevant contribution of this thesis, are questions considered during the remainder of this introduction.

1.1 What is a Topographic Mapping?

Topographic mappings are a class of data-processing mechanisms which seek to preserve some notion of the *geometric structure* of the data within the reduced-dimensional representation. The term 'geometric structure' will be used in this thesis in the sense that *distance* relationships are important, so that points that lie close together in the data space will appear similarly close together in the map[1], and equally, under certain interpretations, points that are more distant in data space will, after mapping, remain likewise separated.

This latter question of interpretation exemplifies that, in practice, there may be alternative emphases placed on the nature of the structure preservation. One emphasis is that *all* distance relationships between data points are important, which implies a desire for global *isometry* between the data space and the map space. Alternatively, it may only be considered important that *neighbourhood relationships* are maintained, such that points that originally lie close together are likewise preserved in the map, and this is referred to as *topological ordering*.

While the word 'topological' is often used in certain contexts as a substitute for 'topographic', it is important to make the distinction between the distance-based criteria considered in this thesis and the notion of *topological invariance* in its strictly mathematical sense. Indeed, "spaces which appear quite different — geometrically for instance — may still be topologically equivalent." [Gamelin and Greene 1983]. In this thesis, 'topographic' will be considered synonymous with 'geometric', in that it is desired that all distance relationships be preserved in the mapping.

Perhaps the most intuitive, and certainly the most literal, example that may be given of a topographic map is that of the projection of the naturally spherical surface of the Earth down onto a two dimensional plane. Such a projection is shown in figure 1.1 below.

This simple illustration also serves to demonstrate an important principle — that when data undergoes a reduction in dimension, some structure is inevitably lost. In practical applications this is an important point, as for high-dimensional datasets the map will normally be of a much lower dimension compared to the original data, and this dimensional imbalance tends to accentuate that problem. In order to represent the topography of the surface of a three-dimensional globe on a two-dimensional plane in figure 1.1, it is necessary to introduce some distortion. While much structure is still retained — consider the interior geography of Europe, for example — at extremes of latitude distances in the map are considerably exaggerated, and even more severely, the left and right longitudinal edges of the map have been drastically separated. Hence the development of various alternatives to Mercator's technique within the field of cartography, such as the Cylindrical Equal Area and Peters' projections, with each introducing its own particular class of distortion.

[1]Throughout this thesis, the word "map" will be used in its intuitive visual sense to refer to the image of the mapping process, rather than in its mathematical sense, as a synonym for transformation.

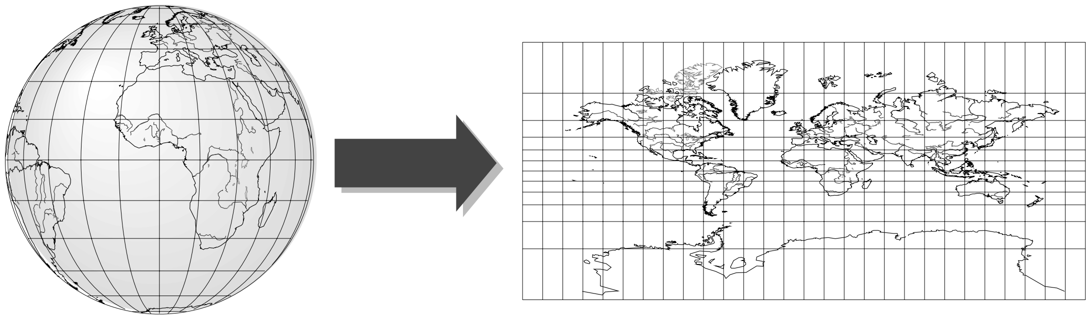

Figure 1.1: A Mercator's projection of the spherical Earth down to a two-dimensional map.

As is evident from the geographical example above, topographic maps can be highly valuable as tools for *visualisation* and *data analysis*. Structure-retaining maps can generally be interpreted quite intuitively, and, as will be seen later, often much more so than other reduced-dimension representations. Many important relationships between the data points can be inferred by viewing the map — notably the detection of *clusters*, or sets of points closely grouped in the data space and which should be similarly adjacent in the projection. However, as will be discussed later in this thesis, under certain conditions, apparent structure exhibited in a map may in fact be *artefactual*, and not be representative of the true geometry in data space. The potential for such phenomena should always be borne in mind when interpreting topographic mappings.

1.2 Why Use a Feed-Forward Neural Network?

There are already some well-established methods for topographic mapping. From the domain of engineering, there is the *Sammon mapping*, or *Nonlinear Mapping*, [Sammon 1969] which is closely related to some of the techniques from the statistical field of *multidimensional scaling* [Davison 1983]. While still in popular use, both approaches possess several inherent disadvantages, the most significant being that when a map has been generated, it effectively acts as a *look-up table* such that there is no potential for projecting new, previously unseen, data. Importantly, this implies that there is no facility for *generalisation*, a principal feature of neural networks and one which, after a given network has been trained, enables prospective inferences to be drawn and predictions to be made concerning new data.

There is also an existing neural network architecture designed specifically for topographic mapping, and that is Kohonen's ubiquitous *self-organising feature map* [Kohonen 1995], which exploits implicit lateral connectivity in the output layer of neurons. This neuro-biologically inspired scheme, however, also exhibits several disadvantages and this thesis will propose an alternative paradigm which exploits the standard feed-forward network architectures.

Feed-forward neural networks are now well established as tools for many information-processing tasks — regression, function approximation, time series prediction, nonlinear dimension-reduction, clustering and classification are examples of the diverse range of applications. (See [Haykin 1994] for a comprehensive coverage.) Divorced from their neuro-biological foundation, the major attraction of neural network models is that certain classes thereof have been shown to be *universal function approximators*, such that they are capable of modelling any continuous function over a bounded domain, given sufficient network complexity. This property implies that, given appropriate design and training, neural networks can be employed as *semi-parametric* models, and thus require fewer prior assumptions about the underlying relationships in the data.

It would be attractive, then, to generate topographic mappings using such architectures. That is, the function that *transforms* the vectors in the data space to a corresponding set of image vectors in the map will be effected by a feed-forward neural network. This concept is illustrated in figure 1.2.

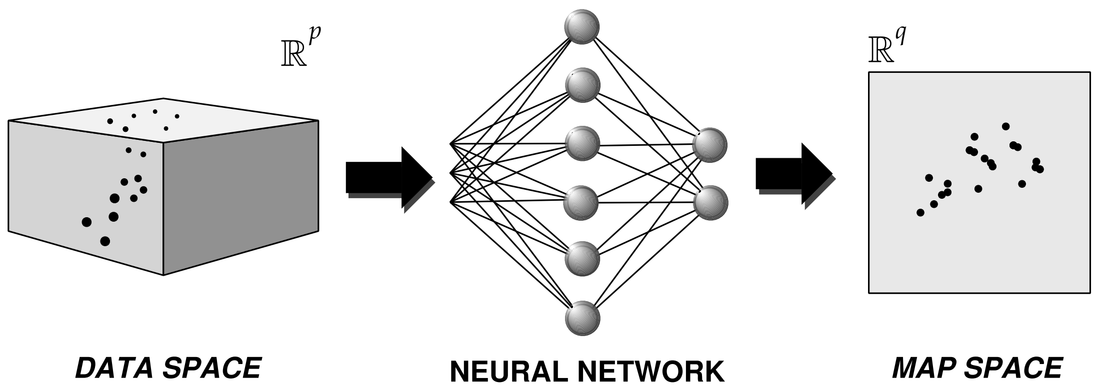

Figure 1.2: A neural network effecting a topographic transformation.

On initial consideration, the training of such a topographic transformation might appear problematic. In the majority of neural network applications, for example regression or classification, there are a set of *target* vectors, corresponding to the set of input vectors — effectively a set of *desired* outputs that the network is trained to reproduce. This scenario is a referred to as a *supervised* problem. In the *unsupervised* topographic case, for each input datum there is no such specific target information, and alternative training algorithms must be developed, based on structural (distance) constraints.

The specific neural network model introduced in this thesis, for reasons of textual brevity, will be known as 'NEUROSCALE', as it is a neural network 'scaling' procedure. NEUROSCALE utilises a *radial basis function neural network* (RBF) [Broomhead and Lowe 1988; Lowe 1995] to transform the p-dimensional input vector to the q-dimensional output vector, where, in general, $p > q$. An RBF comprises a single hidden layer of h neurons, as exhibited by the network in figure 1.2, which represents a set of *basis functions*, each of which has a *centre* located at some point in the input space. The number of such functions is generally chosen to be fewer than the number of data points, and their corresponding centres are initially distributed (and are generally fixed) amongst the data, such that their distribution approximates that of the data points themselves. The output of each hidden node for a given input vector is then calculated as some function (e.g. Gaussian) of the distance from the data point to the centre of the function. In this way the basis functions are *radially symmetric*. The output of the network is then calculated as a weighted, linear summation of the hidden nodes, which for supervised problems with sum-of-squares error functions, permits the weights to be trained by standard linear algebraic methods [Strang 1988]. So mathematically, for a p-dimensional input vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, the q-dimensional output vector $\mathbf{y} = (y_1, y_2, \ldots, y_q)$ is given by:

$$y_i = \sum_{j=1}^{h} w_{ij} \phi_j(\| \mathbf{x} - \boldsymbol{\mu}_j \|), \tag{1.1}$$

where $\phi_j(\cdot)$ is the j^{th} basis function with centre $\boldsymbol{\mu}_j$, and w_{ij} is the weight from that basis function to output node i. An important result concerning this particular type of network is that it is capable of *universal approximation* [Park and Sandberg 1991].

The training algorithm for the RBF constrains vectors in the output space to be located such that they preserve, as optimally as possible, the distance relationships between their corresponding vectors in the input space. This is in contrast to Kohonen's approach, in which the distribution of the output vectors is approximately representative of the data *density*. This can be one of the disadvantages of the latter approach, particularly in applications where global relationships are considered important.

A further important feature of the NEUROSCALE approach to topographic mapping is the inclusion of a unique mechanism for incorporating *preferential* information. This enables additional knowledge concerning the data (for example class labels or other relevant measurements) to be exploited for the purposes of enhancing clustering, improving group separation or even to impose some additional global ordering upon those groups. Such a facility can be considered as adding a *supervisory* component to the otherwise unsupervised feature extraction process, and this interpretation provides an appropriate basis for comparison with other established information-processing paradigms.

That this supervisory mechanism is of tangible benefit, and that NEUROSCALE in general is an effective tool for the exploratory analysis of data, will be shown in an application to one particularly complex dataset. The data in question is taken from the *1992 Research Assessment Exercise* for higher educational institutions in the United Kingdom and is typical of real-world datasets. The data itself is high-dimensional and polluted by noise, and there is additional information available in terms of a class label ("research rating") that is biased by the subjective opinion of an assessment panel. Nevertheless, this extra knowledge will be exploited to generate improved visualisation spaces which can be used as a basis for subsequent prediction of unclassified data.

The NEUROSCALE approach as detailed in this thesis is an incremental development of recent research effort directed at exploiting neural networks to perform structure-retaining mappings. The author is unaware of any significant theoretical investigation into the training and application of such models, and a considerable portion of this thesis is devoted to such detailed analysis.

1.3 Plan of This Thesis

Chapter 1 is this introduction.

Chapter 2 will describe standard approaches to topographic mapping — Kohonen's self-organising feature map, the Sammon mapping and multidimensional scaling — and consider the key distinctions between the three, along with their respective advantages and disadvantages.

Chapter 3 introduces the NEUROSCALE model and relates it to previous work, giving examples of its application to various datasets. These illustrate both the topographic property of the neural network transformation and the facility to exploit additional knowledge.

Chapter 4 is a detailed study of data taken from the 1992 Research Assessment Exercise. Data from the subject areas of physics, chemistry and biological sciences is analysed, both by NEUROSCALE and by other established feature extraction techniques. The emphasis of this chapter is on the visualisation and exploratory analysis of the high-dimensional data, but there are additional results presented for classification experiments, including the use of NEUROSCALE as a pre-processor in prediction models.

Chapter 5 describes a generalisation of the NEUROSCALE approach to classical multidimensional scaling. This is closely related to other neural networks specifically designed for generating principal component projections, notably Oja's principal subspace network, and the parallels are analysed.

Chapter 6 is a study of some of the underlying theoretical aspects of training neural networks to effect topographic mappings. The problem of local minima is considered, and the dynamics of the relative supervision learning algorithm investigated. Analysis is presented concerning the necessary form and smoothness for topographic transformations which is highly relevant to the question of generalisation.

Chapter 7 considers the optimisation of topographic transformations. Standard techniques are compared, and alternative heuristic strategies also reviewed. An efficient new training algorithm for networks linear in their weights is presented, and its properties studied.

Chapter 8 concludes the thesis with a summary of the significant results therein and suggests directions for future research.

The content of this thesis represents original research. The work within has not previously appeared elsewhere, with the exception of those research papers produced during the normal course of its preparation. Material from Chapters 3 and 4 has appeared in [Lowe and Tipping 1995; Lowe and Tipping 1996], while a paper based on Chapter 5 has been submitted for future publication [Tipping 1996].

1.4 Notation

In general, throughout this thesis, the notation below in table 1.1 will be adopted:

Symbol	Meaning
N	The number of data points
p	The dimension of input space
q	The dimension of the map, or feature, space
h	The number of hidden units in a neural network
$\mathbf{x}_i$	A point vector in the input space
$\mathbf{X}$	The matrix of row-vector input points, $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)^{\mathrm{T}}$
$\mathbf{y}_i$	A point vector in the feature space
$\mathbf{Y}$	The matrix of row-vector mapped points, $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)^{\mathrm{T}}$
$\mathbf{A}^{\mathrm{T}}$	The transpose of matrix (or vector) $\mathbf{A}$
$\mathrm{tr}\,[\mathbf{A}]$	The trace of matrix $\mathbf{A}$
$\lvert\mathbf{A}\rvert$	The determinant of matrix $\mathbf{A}$
$\Vert \mathbf{v} \Vert$	The (L_2) norm of vector $\mathbf{v}$
$\mathbf{u}_k$	The k-th eigenvector of some matrix
λ_k	The corresponding eigenvalue
①,②, . . .	A numbered list of items
❶,❷, . . .	A sequential algorithm

Table 1.1: Notation