Visual Data Mining: Integrating Machine Learning with Information Visualization

Dharmesh M. Maniyar and Ian T. Nabney
NCRG, Aston University
Birmingham B4 7ET, United Kingdom
maniyard@aston.ac.uk, nabneyit@aston.ac.uk

ABSTRACT

Today, the data available to tackle many scientific challenges is vast in quantity and diverse in nature. The exploration of heterogeneous information spaces requires suitable mining algorithms as well as effective visual interfaces. Most existing systems concentrate either on mining algorithms or on visualization techniques. Though visual methods developed in information visualization have been helpful, for improved understanding of a complex large high-dimensional dataset, there is a need for an effective projection of such a dataset onto a lower-dimension (2D or 3D) manifold. This paper introduces a flexible visual data mining framework which combines advanced projection algorithms developed in the machine learning domain and visual techniques developed in the information visualization domain. The framework follows Shneiderman's mantra to provide an effective user interface. The advantage of such an interface is that the user is directly involved in the data mining process. We integrate principled projection methods, such as Generative Topographic Mapping (GTM) and Hierarchical GTM (HGTM), with powerful visual techniques, such as magnification factors, directional curvatures, parallel coordinates, billboarding, and user interaction facilities, to provide an integrated visual data mining framework. Results on a real life high-dimensional dataset from the chemoinformatics domain are also reported and discussed. Projection results of GTM are analytically compared with the projection results from other traditional projection methods, and it is also shown that the HGTM algorithm provides additional value for large datasets. The computational complexity of these algorithms is discussed to demonstrate their suitability for the visual data mining framework.

Keywords

Visual Data Mining, Data Visualization, Principled Projection Algorithms, Information Visualization Techniques.

1. INTRODUCTION

The wide availability of ever-growing data sets from different domains has created a need for effective knowledge discovery and data mining. For data mining to be effective, it is important to include the domain expert in the data exploration process and combine the flexibility, creativity, and general knowledge of the domain expert with automated machine learning algorithms to obtain useful results [17]. The principal purpose of visual data exploration is to present the data in a visual form provided with interactive exploration facilities, allowing the domain expert to get insight into the data, draw conclusions, and understand the structure of the data. Exploration of complex information spaces is an important research topic in many fields, including computer graphics, data mining, pattern recognition, and other areas of statistics, as well as database management and data warehousing.

Visual techniques on their own cannot entirely replace analytic nonvisual mining algorithms to represent a large high-dimensional dataset in a meaningful way. Rather, it is useful to combine multiple methods from different domains for effective data exploration [20] [41] [14]. Recently, the core research in visual data mining has focused on combination of visual and nonvisual techniques as well as on integrating the user in the exploration process. Integrating visual and nonvisual methods in order to support a variety of exploration tasks, such as identifying patterns in large unstructured heterogeneous information or identifying clusters or studying data in different clusters in detail etc., requires sophisticated machine learning algorithms, visual methods, and interaction techniques.

Ankerst [1] classified visual data mining approaches into three categories. Methods of the first group apply visual methods independently of data mining algorithms. The second group uses visual methods in order to represent patterns and results from mining algorithms graphically. The third category tightly integrates both mining algorithms and visual methods in such a way that intermediate steps of the mining algorithms can be visualized and further guided by the domain expert. This tight integration allows users to control and steer the mining process directly based on the given visual feedback. The approach we present here, belongs to the third category where we introduce tight integration between principled projection algorithms and powerful information visualization techniques.

Shneiderman's mantra of "Overview first, zoom and filter, details-on-demand" [30] nicely summarizes the design philosophy of modern information visualization systems. First,

MDM '2006 Philadelphia, USA

the user needs to get an overview of the data. In the second stage, the user identifies interesting patterns and focuses on one or more of them. Finally, to analyze patterns in detail, the user needs to drill down and access details of the data. Information visualization technology may be used for all three steps of the data exploration process [20].

For a complex large high-dimensional dataset, where clear clustering is difficult or inappropriate, grouping of data points in soft clusters and then using visual aids to explore it further can reveal insight that may prove useful in data mining [36]. Principled projection of high-dimensional data on a lower-dimension space is an important step to obtain an effective grouping and clustering of a complex high-dimensional dataset [10]. Here, we use the term *projection* to mean any method of mapping data into a lower-dimensional space in such a way that the projected data keeps most of the topographic properties (i.e. 'structure') and makes it easier for the users to interpret the data to gain useful information from it. As mentioned by Ferreira et. al. [10], traditional projection methods such as Principle Component Analysis (PCA) [4], Factor Analysis [13], Multidimensional Scaling [42], Sammon's mapping [29], Self-Organizing Map (SOM) [18], and FastMap [12] are all used in the knowledge discovery and data mining domain [15] [37] [38] [19].

For many real life large high-dimensional datasets, the Generative Topographic Mapping (GTM) [8], a principled projection algorithm, provides better projections than those obtained from traditional methods, such as PCA, Sammon's mapping, and SOM [5] [26]. Moreover, since the GTM provides a probabilistic representation of the projection manifold, it is possible to analytically describe (local) geometric properties anywhere on the manifold. For example, we can calculate the local magnification factors [6], which describe how small regions in the visualization space are stretched or compressed when mapped to the data space. Note that it is not possible to obtain magnification factors for PCA and Sammon's mapping. For the SOM, the magnification factors can only be approximated [7]. It is also possible in the GTM to calculate analytically the local directional curvatures of the projection manifold to provide the user with a facility for monitoring the amount of folding and neighborhood preservation in the projection manifold [34]. The details of how these geometric properties of manifold can be used during visual data mining are presented in Section 3.1.

Moreover, it has been argued that a single two-dimensional projection, even if it is non-linear, is not usually sufficient to capture all of the interesting aspects of a large high-dimensional datasets. Hierarchical extensions of visualization methods allow the user to "drill down" into the data; each plot covers a smaller region and it is therefore easier to discern the structure of the data. Hierarchical GTM (HGTM) is a hierarchical visualization system which allows the user to explore interesting regions in more detail [35].

The work presented here describes a flexible framework for visual data mining which combines principled projection algorithms developed in the machine learning domain and visual techniques developed in the information visualization domain to achieve a better understanding of the information (data) space. The framework follows Shneiderman's mantra to provide an effective user interface (a software tool). The advantage of such an interface is that the user is directly involved in the data mining process taking advantage of powerful and principled machine learning algorithms.

The results presented, on a real life dataset from the chemoinformatics domain, clearly show that this interface provides a useful platform for visual data mining of large high-dimensional datasets. Projection results of GTM are analytically compared with projection results from other methods traditionally used in the visual data mining domain. Using the hierarchical data visualization output, the tool also supports the development of new mixture of local experts models [25].

The remainder of this paper is organized as follows. In Section 2, we provide an overview of GTM and HGTM. The main information visualization and interaction techniques applied are described in Section 3. The integrated visual data mining framework we propose is discussed in Section 4. The experiments are presented in Section 5. In Section 6, we discuss computational costs for the projection algorithms. Finally, we draw the main conclusions in Section 7.

2. PROJECTION ALGORITHMS

This section provides a short overview of the projection algorithms we use.

2.1 Generative Topographic Mapping (GTM)

The GTM models a probability distribution in the (observable) high-dimensional data space, $\mathcal{D} = \Re^D$, by means of low-dimensional latent, or hidden, variables [8]. The data is visualized in the latent space, $\mathcal{H} \subset \Re^L$.

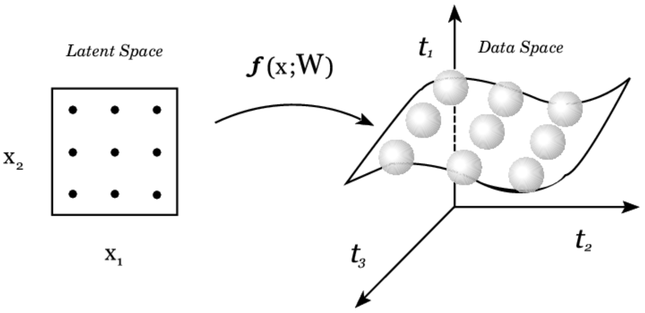


Figure 1: Schematic representation of the GTM model.

As demonstrated in Figure 1 (adapted from [8]), we cover the latent space, $\mathcal{H}$, with an array of K latent space centers, $\mathbf{x}_i \in \mathcal{H}, i = 1, 2, ..., K$. The non-linear GTM transformation, $f : \mathcal{H} \Rightarrow \mathcal{D}$, from the latent space to the data space is defined using an RBF network. To this end, we cover the latent space with a set of $M - 1$ fixed non-linear basis functions (here we use Gaussian functions of the same width σ), $\phi : \mathcal{H} \Rightarrow \Re, j = 1, 2, ..., M - 1$, centred on a regular grid in the latent space. Given a point $\mathbf{x} \in \mathcal{H}$ in the latent space, its image under the map f is

$$f(\mathbf{x}) = \mathbf{W}\phi(\mathbf{x}), \tag{1}$$

where $\mathbf{W}$ is a $D \times M$ matrix of weight parameters and $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), ..., \phi_M(\mathbf{x}))^T$.

The GTM creates a generative probabilistic model in the data space by placing a radially symmetric Gaussian with zero mean and inverse variance β around images, under f, of the latent space centres $\mathbf{x}_i \in \mathcal{H}, i = 1, 2, ..., K$. We refer to the Gaussian density associated with the center $\mathbf{x}_i$

by $P(\mathbf{t} \mid \mathbf{x}_i, \mathbf{W}, \beta)$. Defining a uniform prior over $\mathbf{x}_i$, the density model in the data space provided by the GTM is $P(\mathbf{t} \mid \mathbf{W}, \beta) = \frac{1}{K}\sum_{i=1}^{K} P(\mathbf{t} \mid \mathbf{x}_i, \mathbf{W}, \beta)$.

For the purpose of data visualization, we use Bayes' theorem to invert the transformation f from the latent space $\mathcal{H}$ to the data space $\mathcal{D}$. The posterior distribution on $\mathcal{H}$, given a data point $\mathbf{t}_n \in \mathcal{D}$, is a sum of delta functions centred at centres $\mathbf{x}_i$, with coefficients equal to the posterior probability $R_{i,n}$ that the i-th Gaussian (corresponding to the latent space center $\mathbf{x}_i$) generated $\mathbf{t}_n$ [8],

$$R_{i,n} = \frac{P(\mathbf{t}_n \mid \mathbf{x}_i, \mathbf{W}, \beta)}{\sum_{j=1}^{K} P(\mathbf{t}_n \mid \mathbf{x}_j, \mathbf{W}, \beta)}. \qquad (2)$$

The latent space representation of the point $\mathbf{t}_n$, i.e. *the projection of* $\mathbf{t}_n$, is taken to be the mean, $\sum_{i=1}^{K} R_{in}\mathbf{x}_i$ of the posterior distribution on $\mathcal{H}$. The parameters of the GTM (weights $\mathbf{W}$ and inverse variance β) are learned from data using an Expectation Maximization (EM) algorithm [4].

The f–image of the latent space $\mathcal{H}, \Omega = f(\mathcal{H}) = \{f(\mathbf{x}) \in \Re^D \mid \mathbf{x} \in \mathcal{H}\}$, forms a smooth L-dimensional manifold in the data space. We refer to the manifold Ω as the *projection manifold* of the GTM.

2.2 Hierarchical GTM (HGTM)

The hierarchical GTM (HGTM) arranges a set of GTMs and their corresponding plots in a tree structure $\mathcal{T}$ [33]. An example HGTM structure is shown in Figure 2.

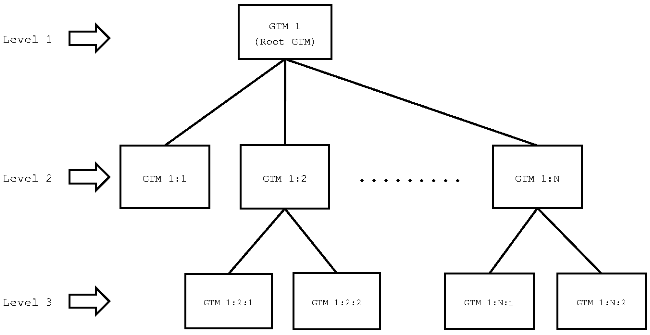


Figure 2: An example structure for HGTM.

The *Root* of the hierarchy is at level 1, i.e. *Level(Root)* = 1. Children of a model $\mathcal{N}$ with *Level*($\mathcal{N}$) $= \ell$ are at level $\ell + 1$, i.e. *Level*($\mathcal{M}$) $= \ell + 1$, for all $\mathcal{M} \in$ *Children*($\mathcal{N}$). Each model $\mathcal{M}$ in the hierarchy, except for *Root*, has an associated parent-conditional mixture coefficient, or prior $\pi(\mathcal{M} \mid \mathit{Parent}(\mathcal{M}))$. The priors are non-negative and satisfy the consistency condition: $\sum_{\mathcal{M} \in \mathit{Children}(\mathcal{N})} \pi(\mathcal{M} \mid \mathcal{N}) = 1$. Unconditional priors for the models are recursively calculated as follows: $\pi(\mathit{Root}) = 1$, and for all other models

$$\pi(\mathcal{M}) = \prod_{i=2}^{\mathit{Level}(\mathcal{M})} \pi(\mathit{Path}(\mathcal{M})_i \mid \mathit{Path}(\mathcal{M})_{i-1}), \qquad (3)$$

where *Path*($\mathcal{M}$) $= (\mathit{Root}, ..., \mathcal{M})$ is the N-tuple ($N = \mathit{Level}(\mathcal{M})$) of nodes defining the path in $\mathcal{T}$ from *Root* to $\mathcal{M}$.

The distribution given by the hierarchical model is a mixture of leaf models of $\mathcal{T}$,

$$P(t \mid \mathcal{T}) = \sum_{\mathcal{M} \in \mathit{Leaves}(\mathcal{T})} \pi(\mathcal{M}) P(t \mid \mathcal{M}). \qquad (4)$$

We obtain a *soft* segmentation of the input space from the HGTM model.

Non-leaf models not only play a role in the process of creating the hierarchical model, but in the context of data visualization can be useful for determining the relationship between related subplots in the hierarchy.

The HGTM is trained using the EM algorithm to maximize its likelihood with respect to the data sample $\varsigma = \{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_N\}$. Training of a hierarchy of GTMs proceeds in a recursive fashion. Visualization and interaction is provided to the domain experts during the intermediate steps of training an HGTM model. A detailed description of user interaction and construction of HGTM models is provided in section 4.

3. VISUAL TECHNIQUES

There is a large number of visual techniques developed in the information visualization domain which can be used for visualizing data or results. In addition to standard 2D/3D graphs, there is a number of more sophisticated visualization techniques. Keim [17] provided an informative overview of different classes of information visualization techniques. In addition to the usual facilities such as zoom, rotate, etc., the following specialized information visualization aids support exploration of the projection obtained from the principled machine learning algorithms to create a powerful visual data mining framework.

3.1 Magnification Factors and Directional Curvatures

One of the main advantages of using GTM–based models is that it is possible to analytically calculate the Magnification Factors (MF) [6] and the Directional Curvature (DC) [34] of the GTM projection manifold. MFs of a GTM projection manifold, Ω, are calculated as the determinant of the Jacobian of the GTM map f [6]. Magnification factor plots are used to observe the amount of stretching in a GTM manifold at different parts of the latent space, which helps in understanding the data space, outlier detection, and cluster separation. Tiňo *et. al.* [34] derived a closed-form formula for directional curvatures of the GTM projection manifold, Ω, for a latent space point $\mathbf{x} \in \mathcal{H}$ and a directional vector $\mathbf{h} \in \mathcal{H}$. Directional curvature plots allow the user to observe the direction and amount of folding in the GTM manifold. This can help the user detect regions where the GTM manifold does not fit the data well. It is possible that groups of data points far apart when projected onto the projection manifold are close together in the data space due to high folding in the manifold. This neighborhood preservation in the data space can be spotted with a strong curvature band on the corresponding directional curvature plot.

The magnification factor is represented by color shading in the projection manifold (e.g., see Figure 4(c)). The lighter the color, the more stretch in the projection manifold. The direction of folding in the projection manifold plot is presented using a small line for each part of the projection manifold in the directional curvature plots (e.g., see Figure 4(d)). The length and the shade of the background color represents the magnitude of folding. The longer the line and the lighter the background color, higher the folding (curvature).

3.2 Local Parallel Coordinates

The *parallel coordinates technique* [16] maps the D dimensional data space onto two display dimensions by using D equidistant axes which are parallel to one of the display axes. It displays each multi-dimensional data point as a polygonal line which intersects the horizontal dimension axes at the position corresponding to the data value for the corresponding dimension.

Instead of displaying parallel coordinates for all the data points together, which is impractical for a large dataset, we provide an interactive facility to let the user select a point on the projection manifold and display parallel coordinates for the n nearest neighbors of that selected point. Figure 3(a) displays an example of parallel coordinates used on a GTM projection: when the user clicks on a point in the projection (upper plot), the data space visualization graph shows a color coded plot of normalized property values for a group of points close in the projection space. We call this a *local* parallel coordinates technique. This facility has proved very useful for the domain experts at Pfizer[1] to understand large high-dimensional datasets. Using this facility, the user can study properties of a high-dimensional data point in the data space while working with the lower-dimensional latent (projection) space. A detailed example discussing how local parallel coordinates are used to explore a projection manifold is presented in Section 5.2.

3.3 Billboarding

For many real-life datasets which have a natural representation, e.g. chemical compound structure, handwritten digit recognition, face recognition, galaxy classification, etc., using this natural representation of data points in the projection is more useful to understand the data compared with data represented by labelled and/or colored dots.

Here the term 'billboarding' means visualizing a natural representation of a data point in the form of an image, in such a way that the image always faces the viewer (even in 3D). A chemical compound structure or a hand written digit image is certainly more user-friendly than a dot.

Partiview [22] is an interactive 3D visualization tool supporting a billboarding facility, primarily created for astronomy-related applications. But recently it has been successfully used for visualizing the output of some machine learning algorithms [31].

The number of pictures that can be displayed at a time depends on how much graphics memory is present. Figure 3(b) presents a close up of the points visualized for the MNIST database [21] using Laplacian eigenmaps [3]. Billboarding presentation of images of the handwritten digits provides us an intuitive visualization and can help us to identify why certain data points are misclassified (e.g., notice that in Figure 3(b), images of 7s and 9s on the top left corner of the plot are quite similar). Partiview also provides many useful interaction facilities, such as 3D zooming and traversal, selective plotting of classes, properties displace, etc. [22].

4. THE INTEGRATED VISUAL DATA MINING FRAMEWORK

The integrated visual data mining framework combines principled projection algorithms, discussed in Section 2, and visual techniques, discussed in Section 3, to achieve a better understanding of the data space. It follows Shneiderman's mantra [30],"Overview first, zoom and filter, details on demand", to develop an effective interface.

To support the 'overview first' stage of Shneiderman's mantra, output of the projection algorithms and basic visualization aids such as highlight, rotate, etc., are provided for exploring a large high-dimensional dataset. For the second stage, 'zoom and filter', visualization aids such as zooming, filtering interesting regions on the projection manifold with the use of magnification factor and directional curvatures plots, etc., are provided. This allows the user to identify and concentrate on interesting subsets of the projection we obtained in the first stage. The third stage, 'details-on-demand', is supported using local parallel coordinates and billboarding. Integration with other visualization tool is also possible at various stages.

Moreover, a single two-dimensional projection, even if it is non-linear, is not usually sufficient to capture all of the interesting aspects of a large high-dimensional data sets. A hierarchical system which allows user to interactively drill down in the projection can be useful.

Interactive visual methods support the construction of HGTM models and allow the user to explore interactively interesting regions in more detail. Visual aids described in Section 3 are provided at each stage of the HGTM model development. First, a base (*Root*) GTM is trained and used to visualize the data. Then the user identifies interesting regions on the visualization plot that they would like to explore in greater detail. In particular, the user chooses a collection of points, $c_i \in \mathcal{H}$, by clicking on the projection plot. The "regions of interest" given by these points (centres) are then transformed into the data space as Voronoi compartments [2] defined by the mapped points $f_{Root}(c_i) \in \mathcal{D}$, where f_{Root} is the map of the *Root* GTM. The child GTMs are initiated by local PCA in the corresponding Voronoi compartments.

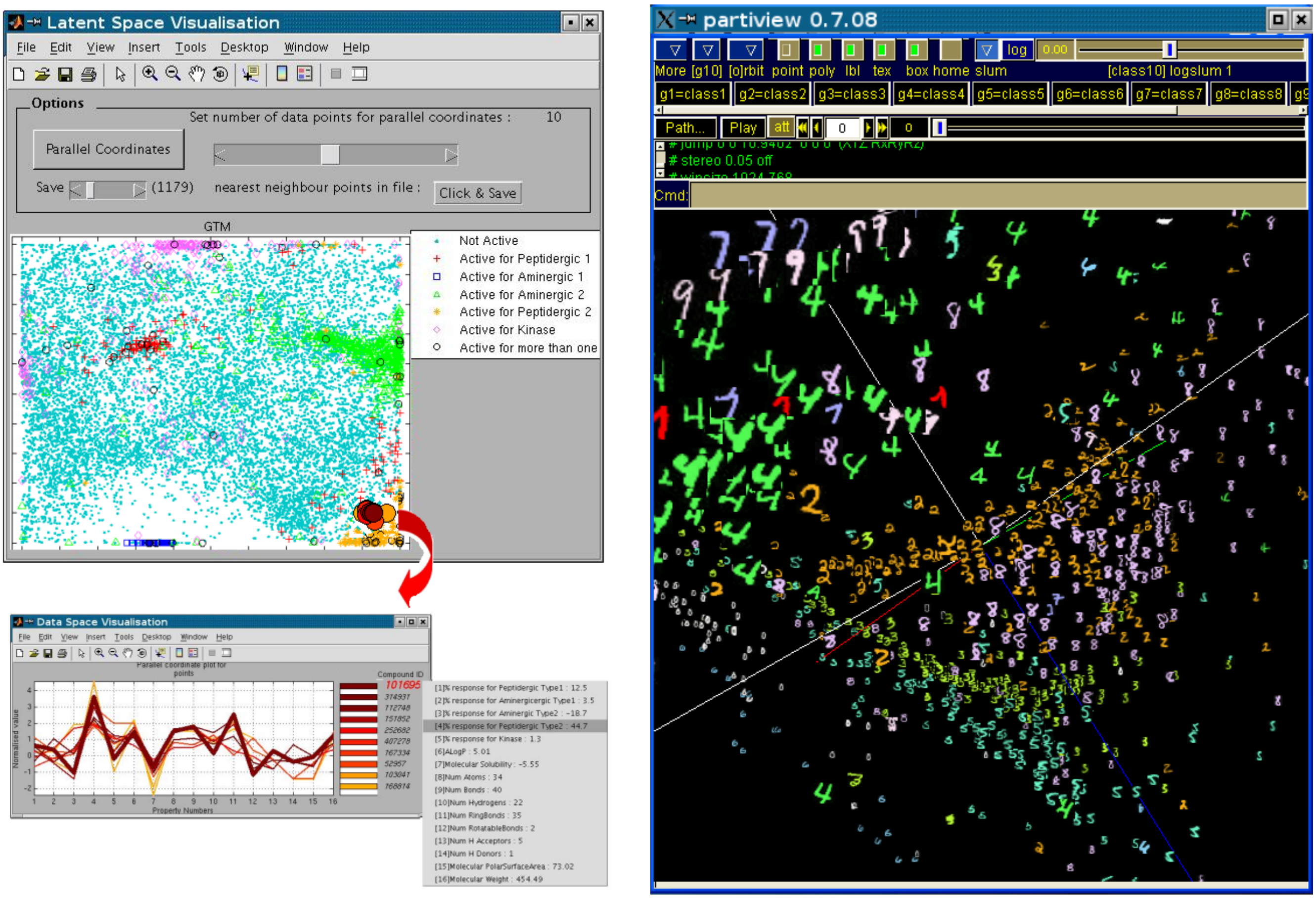
After training the child GTMs and seeing the lower level visualization plots, the user may decide to proceed further and model in greater detail some portions of the lower level plots, etc. Thus, HGTM allows domain experts to segment the input space interactively using data visualization.

When the dataset is very large, the higher-level projection plots may be cluttered and confused (with densely clustered and overlapping projections). This makes it difficult for the user to select locations for submodels at the next level. In such cases, an alternative semi-automatic submodel initialization algorithm [28], based on minimum message length (MML) criteria, which decides both the number of submodels and their location can be used for higher-level projections of the visualization hierarchy and then the domain expert can take control to guide the lower-level projections.

We have developed an interactive software tool that supports this framework [24]. The interface is developed in MATLAB[2] using the NETLAB [27] toolbox. The tool supports other projection methods, such as PCA, Sammon's mapping, and SOM. The interface has proved useful for domain experts to understand and mine large high-dimensional

[1] Pfizer Global Research and Development, Kent, UK.

[2] ©The MathWorks Inc., http://www.mathworks.com/


Latent Space Visualisation
File Edit View Insert Tools Desktop Window Help
Options
Set number of data points for parallel coordinates : 10
Parallel Coordinates
Save (1179) nearest neighbour points in file : Click & Save
GTM
Not Active
Active for Peptidergic 1
Active for Aminergic 1
Active for Aminergic 2
Active for Peptidergic 2
Active for Kinase
Active for more than one
Data Space Visualisation
Parallel coordinate plot for points
Normalised value
Property Numbers
Compound ID
101695
314931
112748
151852
252682
407278
167334
52957
103041
168814
partiview 0.7.08
More [g10] [o]rbit point poly lbl tex box home slum [class10] logslum 1
g1=class1 g2=class2 g3=class3 g4=class4 g5=class5 g6=class6 g7=class7 g8=class8
Path... Play
# stereo 0.05 off
Cmd:

[1]% response for Peptidergic Type1 : 12.5
[2]% response for Aminergicergic Type1 : 3.5
[3]% response for Aminergic Type2 : -18.7
[4]% response for Peptidergic Type2 : 44.7
[5]% response for Kinase : 1.3
[6]ALogP : 5.01
[7]Molecular Solubility : -5.55
[8]Num Atoms : 34
[9]Num Bonds : 40
[10]Num Hydrogens : 22
[11]Num RingBonds : 35
[12]Num RotatableBonds : 2
[13]Num H Acceptors : 5
[14]Num H Donors : 1
[15]Molecular PolarSurfaceArea : 73.02
[16]Molecular Weight : 454.49

(a) The projection interface and the local parallel coordinates.

(b) Billboarding example.

Figure 3: Information visualization facilities.

datasets.

Visualization is a valuable tool for exploring and understanding data, but in many applications the fundamental task is one of prediction. It has been argued that a single global classification/regression model can rarely capture the full variability of a huge multi-dimensional dataset. Instead, local models, each focused on a separate area of input space (a cluster), often work better since the mapping in different areas may vary. The tool also supports the guided mixture of local experts model, which uses the soft segmentation obtained using probabilistic hierarchical visualization algorithms, such as HGTM, to formulate the guided local mixture of experts model [25]. Thus the visual mining framework is not just a visual exploration tool but also supports guided modelling where the domain expert is closely involved.

5. EXPERIMENTS

Experiments were carried out on a real-life dataset from the pharmaceutical domain. A typical challenge in the early stages of the drug discovery process is to understand and explore large datasets containing High Throughput Screening (HTS) results (biological activity) alongside some whole-molecular properties [11]. A chemist/biologist is interested in studying and exploring clusters of active compounds to understand the data and make informed decisions for future screens. The chemoinformatics dataset provided by the chemist at Pfizer is described in Section 5.1. The results are discussed in Section 5.2.

5.1 The Dataset

The chemoinformatics dataset, we used, is composed of 23,600 compounds having biological activity data for five different biological targets and 11 whole-molecular physicochemical properties. Thus, the dataset has, in total, 16 variables (dimensions) in the data space and we want to effectively visualize it on a 2-dimensional manifold.

Out of these five biological targets, two are peptidergic G-Protein coupled receptor (GPCR) targets, two are aminergic GPCR targets, and one is a kinase target. The four GPCR targets are of related receptor types whilst the kinase is a completely unrelated enzyme target class. Table 1 lists the label information and distribution of compounds in different labels.

In addition to the biological activity values, 11 whole–

Table 1: Label information and compound distribution across labels.

Label Description	Marker	Compounds
`Not active in any screen`	●	21540
`Active for peptidergic type1`	+	236
`Active for peptidergic type2`	*	362
`Active for aminergic type1`	□	100
`Active for aminergic type2`	△	818
`Active for kinase`	◇	412
`Active for more than 1 screen`	○	132

Table 2: Molecular physicochemical properties.

`AlogP`
`Molecular solubility`
`Number of atoms`
`Number of bonds`
`Number of Hydrogens`
`Number of ring bonds`
`Number of rotatable ring bonds`
`Number of Hydrogen acceptors`
`Number of Hydrogen donors`
`Molecular polar surface area`
`Molecular weight`

molecule physiochemical properties were included for each compound in the dataset. Table 2 lists the physicochemical properties used.

Since different input variables in the dataset have different ranges, before the development of visualization models we apply a linear transformation (Z-score transformation) to have similar ranges for all variables. Each variable is treated independently and is rescaled as follows:

$$\mu_i = \frac{1}{N}\sum_{n=1}^{N} {x_i}^n \tag{5}$$

$$\sigma_i{}^2 = \frac{1}{N-1}\sum_{n=1}^{N}({x_i}^n - \mu_i)^2, \tag{6}$$

where $n = 1, ..., N$ indexes the patterns, and μ_i and $\sigma_i{}^2$ represent mean and variance of variable i respectively. Then the values are scaled by

$$\tilde{x}_i^n = \frac{{x_i}^n - \mu_i}{\sigma_i}. \tag{7}$$

where $\tilde{x}_i^n$ is the scaled value of variable i for the pattern n. Histograms of variables after the scaling show approximately normal distributions (the plot is not included in the paper due to the space constraint).

50% of the dataset was used as the training set and the remaining 50% was used as the test set.

5.2 Results

Figure 4(a) shows the projection using the Neuroscale [23], a novel neural network implementation of Sammon's mapping, algorithm. The projection in Figure 4(a) is like a blob and does not help us to understand the 'structure' of data in data space. GTM visualization results are shown in Figure 4(b). The GTM plot shows clear clusters for the compounds active for different targets. The GTM visualization is certainly more informative than the visualization results obtained from PCA, Sammon's mapping, and SOM (PCA and SOM projection results are similar to those for the Sammon's mapping. They are not shown here because of the space constraint).

Though visually we can easily observe the effectiveness of GTM projection on this dataset, it is useful to get an analytical measurement of the separation between different data classes in the projections. To obtain such a measurement, first we fit a Gaussian mixture model (GMM) [4] on each class in the projection space and then we calculate the Kullback-Leibler (KL) [9] divergence between the fitted GMMs as below:

$$KL(p_a, p_b) = \sum_x p_a(x)\log\frac{p_a(x)}{p_b(x)}, \tag{8}$$

where p_a and p_b are the GMMs for classes a and b respectively.

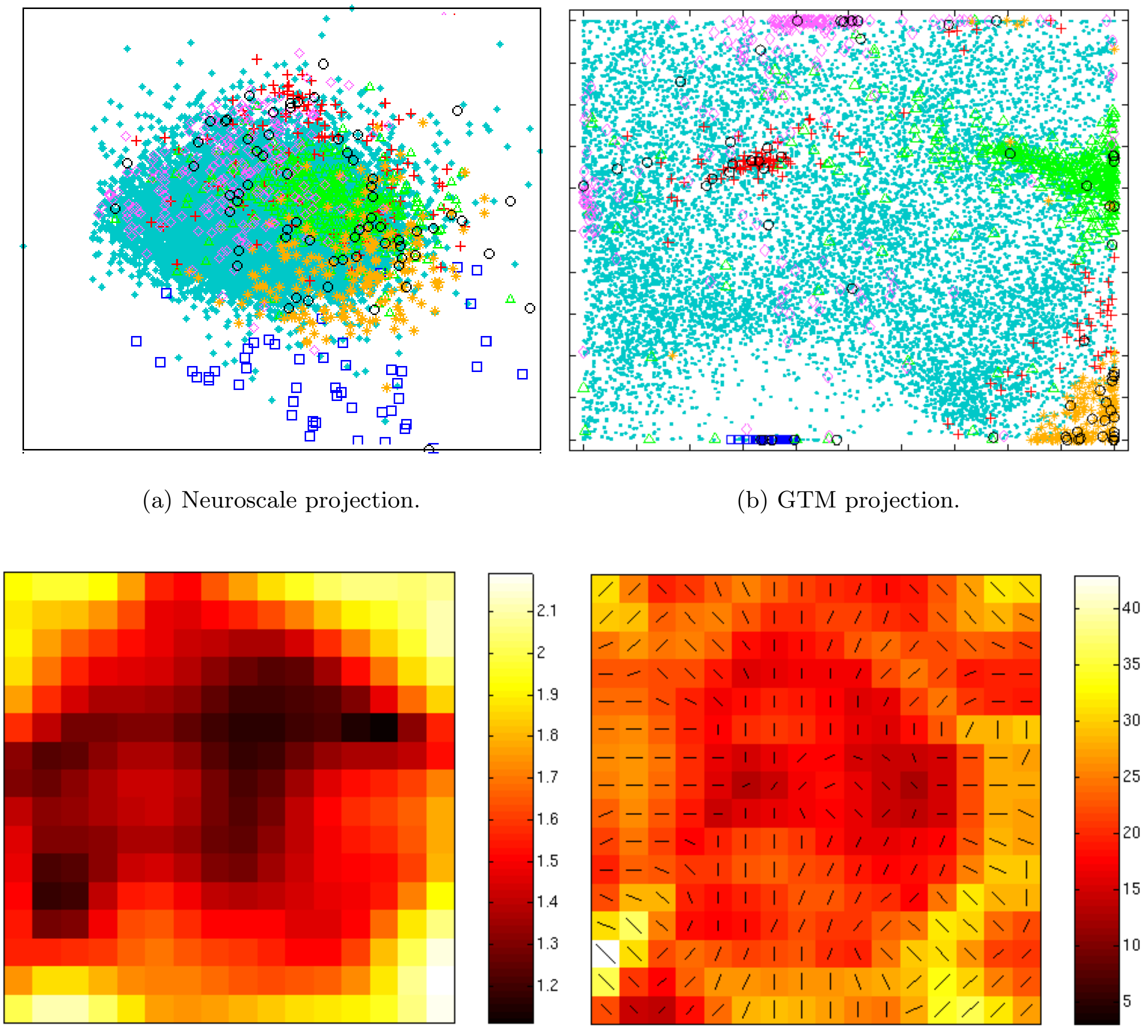
The higher the value of KL divergence, the greater the separation between classes. Table 3 presents the sum of the entries in the KL divergence matrix for different visualization model outputs. GTM projection has a much higher KL divergence matrix sum, which is evidence for better separation using GTM.

Table 3: KL divergence matrix sum for different projection models.

Visualization model	KL divergence matrix sum
PCA	33.63
Neuroscale	56.34
SOM	56.37
GTM	127.42

Once the GTM projection is obtained, the user can study the dataset in more detail using different interactive visual aids. The corresponding magnification factor (MF) and directional curvature (DC) plots of the GTM projection manifold (Figure 4(b)) are presented in Figure 4(c) and Figure 4(d) respectively. Using the MF plot displayed in Figure 4(c), we observe the stretching in the data space. For example, higher value of MF in the bottom-right corner of the MF plot shows that compounds active for peptidergic type 2 target (marked as *) are clustered far away from the bulk of the data. Visual inspection of the direction lines and background color in the DC plot presented in 4(d) can help the user to understand the curvature directions and magnitude of the folding in the projection manifold.

Magnification factor and directional curvature plots are also useful for making decisions about number and the positions of the centers for GTM subplots during the training of an HGTM model. For example, the lighter bands at the bottom right corner in the directional curvature plot (see 4(d)) reveals a huge folding in the projection manifold to cover the data space. This helps us to understand that there could be a cluster there even though the data points are not marked (labeled) differently. Magnification factors and directional curvatures plots are mainly used to understand projection space and data space in detail. But if the data are not colored (labeled) (for example if we do the analysis on a virtual compound library), magnification factor and directional cur-



(a) Neuroscale projection.

(b) GTM projection.

(c) Magnification Factors for the GTM projection.

(d) Directional Curvatures for the GTM projection.

Figure 4: Neuroscale and GTM projections for the chemoinformatics dataset with MF and DC plots for the GTM projection. Legend information for (a) and (b) is given in Table 1.

vature plots can be used to observe clusters in the projection and data space.

Local parallel coordinate plots help us to observe variations in the patterns in different regions of a projection plot. Figure 5 shows how patterns of physicochemical properties vary in different regions of the GTM projection. A careful study with the parallel coordinate technique reveals interesting structures in the projection space. It can be observed that the active compounds for different targets are nicely clustered. Close study with the help of local parallel coordinate technique reveals that the compounds marked as '∘', present in the clusters for peptidergic type 1 and peptidergic type 2, are the active compounds for both of the peptidergic targets. That is in line with the fact that some compounds are active for both of the peptidergic targets.

Billboarding is an effective way to display chemical compound structures in the projection manifold. Because of the confidentiality agreement with Pfizer, we are not able produce billboarding results on the chemoinformatics dataset.

The HGTM visualization results are presented in Figure 6. Active compounds can be seen in different clusters in the root GTM. The deeper level plots clearly separate interesting local regions. At each level, the magnification factor plots, directional curvature plots, parallel coordinates, and the basic visual exploration facilities are used to make decisions about where to place the center of a subplot for the next level.

6. COMPUTATIONAL COST

Although the rapid development of high-performance computing has to some extent altered our perception of computational complexity, this issue cannot be ignored in a visual data mining framework where user interaction is important.

The computational cost for PCA scales as linear, $\mathcal{O}(N)$, in

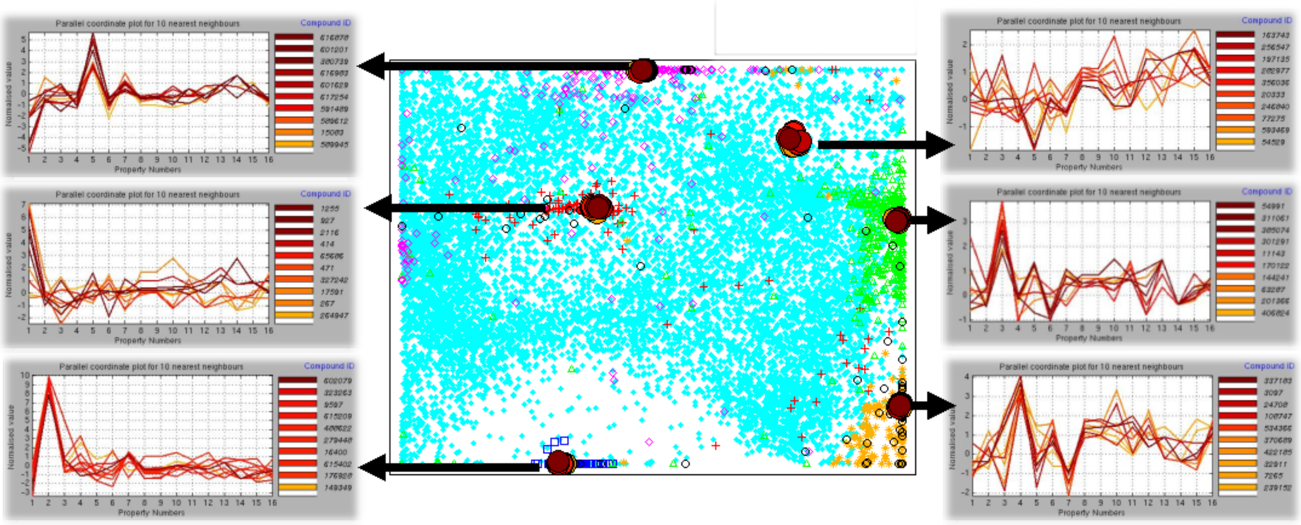


Figure 5: Local parallel coordinates demonstrating variations in the patterns in different regions of the GTM projection (plot 4(d)). Legend information is given in Table 1.

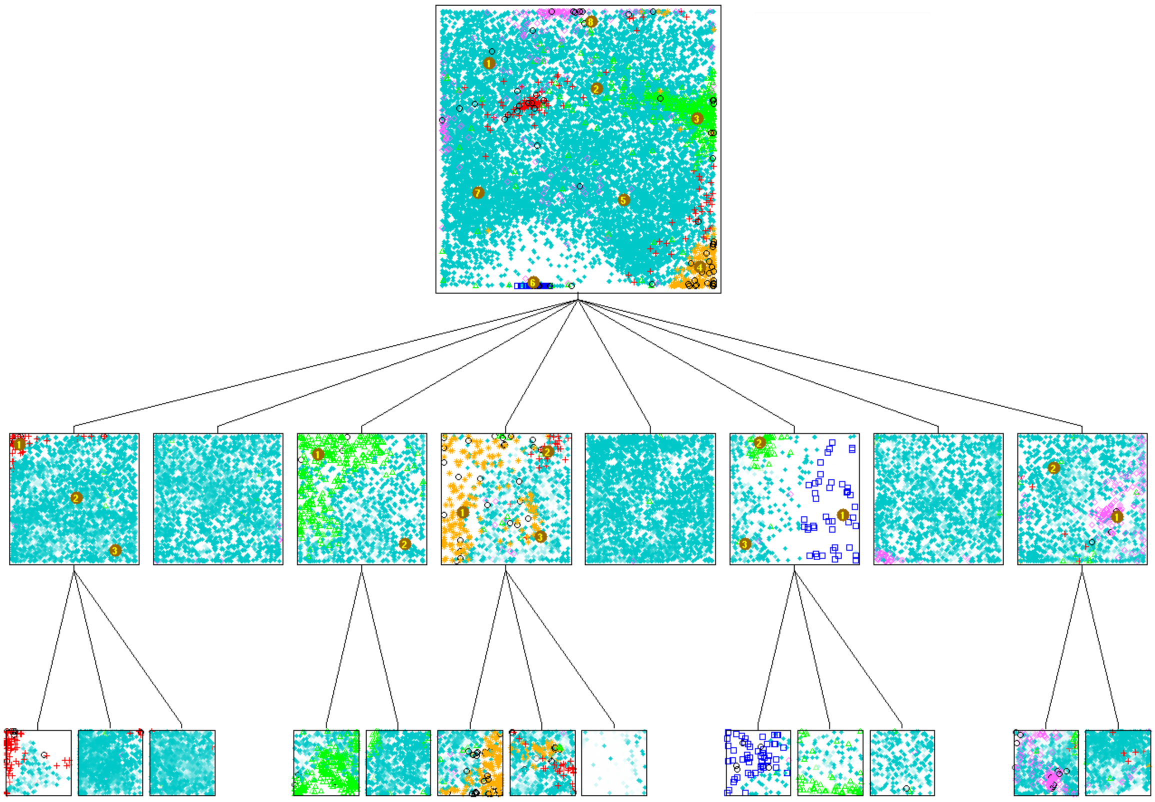


Figure 6: HGTM projection of the chemoinformatics dataset. Legend information is given in Table 1.

the number of data points (N). Neuroscale suffers from the fact that the computational demands grow with the square of the number of data points, $\mathcal{O}(N^2)$. This is because each evaluation of the STRESS error requires the computation of $N(N-1)/2$ inter-point distances. In practice, for large data sets, it is common to apply an initial clustering phase to the data set (using for example the *K-means* algorithm), to generate a set of K prototype vectors (where $K \ll N$). Neuroscale can then be applied to these prototype vectors at a much reduced computational cost. Here we used Neuroscale with the fast shadow targets training algorithm [32].

The distance calculation between data points and mixture components of reference vectors, respectively, is identical in SOM and GTM training algorithms. Updating the parameters in SOM training depends on the neighborhood function. In the experiments presented here it was continuous on the latent space so the parameter updating scales as $\mathcal{O}(L^2ND + L^2)$, where L is the number of grid points in the SOM map and D is the dimension of the data space. When updating parameters, the GTM requires a matrix inversion of an $M \times M$ matrix, where M is the number of basis functions, followed by a set of matrix multiplications. The matrix inversion scales as $\mathcal{O}(M^3)$, while the matrix multiplications scales as $\mathcal{O}(LND)^3$, where L is the number of grid points in the GTM latent space.

Table 4 shows the time taken to train different projection models on the training set using an Intel Pentium 4 - 2.4GHz machine with 2GB of RAM. The implementation of the algorithms in C/C++ instead of MATLAB can further improve the speed.

Table 4: Training time for different projection models ($N = 11800$, 20 iterations).

The model	Time (seconds)	Architecture
PCA	1	-
Neuroscale	546	-
SOM	36	$L = 256$
GTM	42	$L = 256, M = 64$

Once the models are trained, the computational cost to project data for the subsequent test set scales as linear, $\mathcal{O}(N)$, in the number of data points (N) in the test set.

7. CONCLUSIONS

To understand a large high-dimensional dataset, close integration of principled projection methods and information visualization techniques are useful to develop an effective visual data mining framework. The interface developed using this framework and following Shneiderman's mantra for design provided us with a useful tool for visual data mining of large high-dimensional datasets.

Traditional projection algorithms used in data mining domain, such as PCA, Neuroscale, and SOM, are not powerful enough for a good projection for many real life scientific problem where one has to deal with a complex high-dimensional dataset. The GTM algorithm, a principled projection algorithm, has provided a better projection (in terms of cluster separation) than the projection obtained from the traditional methods, such as PCA, Neuroscale, and SOM. For the chemoinformatics dataset analysed here, the KL divergence matrix sum results presented in Table 3 clearly show the effectiveness of the strong class grouping we obtain. With the use of interactive visual aids provided during the projection output, we not only can characterize hit populations from different target classes (i.e. peptidergic GPCRs vs. aminergic GPCRs vs. kinases) but can also understand areas of overlap.

Magnification factor and directional curvature plots of GTM helped to provide a better understanding of the projection manifold and its fitting on data in the data space. The local parallel coordinates technique proved to be a useful tool to understand data points in interesting regions of the projection manifold more in detail. Since the structure of compounds is very important in drug discovery, billboarding could be a useful feature for the domain experts (chemists and screening scientists) at Pfizer to visualize chemical structures in the projection manifold.

HGTM models are useful to explore clusters and interesting local regions in details in a large dataset. Note that this results are only for 11,800 compounds. The number of compounds one has to consider during the drug discovery process is enormous; in such situations, a single GTM projection can look cluttered but a well trained HGTM model could be very useful to provide a better grouping.

The computational cost to train a GTM algorithm is acceptable for inclusion in the visual data mining framework. The GTM and HGTM algorithms are scalable so having a large number of data points during training, however, causes no difficulty beyond increased computational cost.

The interface we developed supports powerful machine learning algorithms and information visualization techniques. The interface is also flexible enough to let a user export data points from selected region on a projection manifold and analyze them further using industry standard data mining tool such as PIPELINE PILOT [39] and SpotFire [40].

In the future we plan to develop data visualization using GTM with simultaneous feature selection and model selection for unsupervised learning problems.

8. ACKNOWLEDGEMENT

D. M. Maniyar is grateful to Pfizer Central Research and the Overseas Research Scholarship award for funding his PhD. We thank Bruce Williams and Andreas Sewing of Pfizer for useful discussions and feedback on the results.

9. REFERENCES

[1] M. Ankerst. Visual data mining with pixel-oriented visualization techniques. *Proc. Workshop on Visual Data Mining*, 2001.

[2] F. Aurenhammer. Vornoi diagrams - survey of a fundamental geometric data structure. *ACM Computing Surveys*, 3:345–405, 1991.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[4] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, New York, 1st edition, 1995.

[3] To be exact, the matrix multiplications scales as $\mathcal{O}(LMD + LND)$, but normally the number of data points, N, exceeds the number of basis functions, M.

[5] C. M. Bishop, M. Svensen, and C. K. I. Williams. GTM: a principled alternative to the self-organizing map. In *Proc. of ICANN '96*, pages 165–70. Springer-Verlag, Berlin, Germany, 1996.

[6] C. M. Bishop, M. Svensén, and C. K. I. Williams. Magnification factors for the GTM algorithm. *Proc. IEE Fifth International Conference on Artificial Neural Networks*, pages 64–69, 1997.

[7] C. M. Bishop, M. Svensén, and C. K. I. Williams. Magnification factors for the som and gtm algorithms. *Work. proc. on Self-Organizing Maps*, pages 333–338, 1997.

[8] C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10:215–234, 1998.

[9] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, New York, 1st edition, 1991.

[10] M. C. F. de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: A survey. *IEEE Trans. on Vis. and Comp. Graphics*, 9(3):378–394, 2003.

[11] P. Englebienne. High throughput screening: Will the past meet the future? *Frontiers in Drug Design & Discovery*, 1:69–86, 2005.

[12] C. Faloutsos and K. Lin. Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, pages 163–174, 1995.

[13] H. H. Harman. *Modern Factor Analysis*. Univ. of Chicago Press, 1967.

[14] A. Hinneburg, D. A. Keim, and M. Wawryniuk. HD-Eye: Visual mining of high-dimensional data. *IEEE Comp. Graph. and App.*, 19(5):22–31, 1999.

[15] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley. DNA visual and analytic data mining. *Proc. of the 8th IEEE Visualization Conf.*, 1997.

[16] A. Inselberg and B. Dimsdale. Parallel coordinates : A tool for visualizing multi-dimensional geometry. *Proc. IEEE VISUALIZATION '90*, pages 361–375, 1990.

[17] D. A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(3):255–269, 2002.

[18] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin, 1995.

[19] E. L. Koua and M.-J. Kraak. Geovisualization to support the exploration of large health and demographic survey data. *International Journal of Health Geographics*, 3(12):1–13, 2004.

[20] M. Kreuseler and H. Schumann. A flexible approach for visual data mining. *IEEE Transactions on Visualization and Comp. Graphics*, 8(1):39–51, 2002.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11):2278–2324, 1998.

[22] S. Levy. Interactive 3-d visualization of particle systems with Partiview. *Proc of the Int Astronomical Union Symposium*, 208:85–91, 2001.

[23] D. Lowe and M. E. Tipping. Neuroscale: Novel topographic feature extraction with radial basis function networks. *Advances in Neural Information Processing Systems*, 9:543–549, 1997.

[24] D. M. Maniyar and I. T. Nabney. DVMS 1.5 : A user manual. Technical Report NCRG/2005/003, Neural Computing Research Group, UK, 2005.

[25] D. M. Maniyar and I. T. Nabney. Guiding local regression using visualisation. *In Deterministic and Statistical Methods in Machine Learning, LNAI, Springer-Verlag*, 3635:98–109, 2005.

[26] D. M. Maniyar, I. T. Nabney, B. S. Williams, and A. Sewing. Data visualization during the early stages of drug discovery. *Journal of Chemical Information and Modelling*, to appear in vol. 46, 2006.

[27] I. T. Nabney. *Netlab: Algorithms for Pattern Recognition*. Springer, London, 1st edition, 2001.

[28] I. T. Nabney, Y. Sun, P. Tiňo, and A. Kabán. Semisupervised learning of hierarchical latent trait models for data visualization. *IEEE Trans. on Know. and Data Eng.*, 17(3):384–400, 2005.

[29] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Tran. on Comp.*, C-18:401–409, 1969.

[30] B. Shneiderman. The eyes have it: A task by data type taxonomy for information. visualizations. *Proc. IEEE Symp. on Visual Lang.*, 3(6):336–343, Sep 1996.

[31] D. Surendran and S. Levy. Visualizing high dimensional datasets using partiview. *IEEE Symp. on Info. Visu., INFOVIS '04*, 2004.

[32] M. E. Tipping and D. Lowe. Shadow targets: A novel algorithm for topographic projections by radial basis functions. *Neurocomputing*, 19:211–222, 1998.

[33] P. Tiňo and I. T. Nabney. Constructing localized non-linear projection manifolds in a principled way: hierarchical generative topographic mapping. *IEEE Trans. on Patt. Ana. and Mach. Inte.*, 24:639–656, 2002.

[34] P. Tiňo, I. T. Nabney, and Y. Sun. Using directional curvatures to visualize folding patterns of the GTM projection manifolds. *Int. Conf. on Artificial Neural Networks*, pages 421–428, 2001.

[35] P. Tiňo, I. T. Nabney, Y. Sun, and B. S. Williams. A principled approach to interactive hierarchical non-linear visualization of high-dimensional data. *Computing Science and Statistics*, 33, 2001.

[36] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, second ed. edition, 2001.

[37] J. Vesanto. SOM-based data visualization method. *Intelligent Data Analysis*, 3(2):111–126, 1999.

[38] S. Wang and H. Wang. Knowledge discovery through self-organizing maps: data visualization and query processing. *Know. and Info. Sys.*, 4(1):31–45, 2002.

[39] Website. © SciTegic. *http://www.scitegic.com/*.

[40] Website. © Spotfire. *http://www.spotfire.com/*.

[41] P. C. Won. Visual data mining. *IEEE Computer Graphics and Applications*, 19(5):20–21, 1999.

[42] F. Young. *Multidimensional Scaling: History, Theory, And Applications*. Lawrence Erlbaum Assoc., Hillsdale, N.J., 1987.