

A simple method for visualizing labelled and unlabelled data in high-dimensional spaces

J. R. Greene

Department of Electrical Engineering, University of Cape Town
Rondebosch, 7001, South Africa.

jrgreene@eng.uct.ac.za

Abstract

The low-dimensional visualisation of high-dimensional data is a valuable way of detecting structure (such as clusters, and the presence of outliers) in the data, and avoiding some of the pitfalls of blind data manipulation. Projection based on principal component analysis is widely employed and often useful, but it is a variance-preserving projection which takes no account of class labels, and may, for this reason, hide significant structure.

Here we present a very simple method which appears to yield useful visualizations for many datasets. It is based on a random search for a linear transformation, and projection into a two-dimensional visual space, which maximises an objective measure of class separability in the visual space. The method, which can be thought of as a variant of projection pursuit with a novel interest measure, is demonstrated on datasets from the UCI Repository. Tentative interim results are also given for a proposed extension based on spectral clustering, for extending the method to unlabelled data.

1. Introduction

The low-dimensional visualization of high dimensional data is often used to gain insight into the presence of structure (such as clustering) or to identify outliers in the data, to aid in the selection of classifiers and generally to try to avoid some of the pitfalls inherent in blind data manipulation. A popular and frequently effective approach is to project the data onto its first few principle components. However this is a variance-preserving linear transformation which takes no account of class labelling, and a principal component projection may be far from optimal for revealing structure. It should be noted that a recent proposal (1) for modified versions of principal component projection addresses this problem and in

many (but not all) cases provides markedly superior visualizations either through a normalisation (half-whitening) approach or by biasing the projection taking class labels into account.

Here we present a very simple, but often highly effective, visualization strategy based on random linear transformation and projection into a two-dimensional visual space. Multiple random visualisations are performed, and that one selected which maximises a simple objective measure of separation in the visual space. In detail, it is based on the following principles and observations.

2. Outline and motivation

1. Complex nonlinear mappings are powerful in revealing distributional structure, but they are also capable of constructing false structure – artefacts of the transformational process (in a process somewhat analogous to overfitting in supervised learning). We therefore opt for the conservative strategy of using only linear transformations. These are partially structure-preserving, in that they cannot break up existing clusters (i.e. make a unimodal distribution multi-modal). Since we are interested in transformations which involve projection into a space of lower dimensionality they can, of course, create false clusters by the adventitious superposition of points in the projected space.

Clearly the set of linear transformations is not sufficiently powerful for visually revealing structure in cases where the data resides on a nonlinear manifold of the representation space. However powerful methods (auto-associative neural networks,

multidimensional scaling, Self-organising maps, Sammon mapping, Principal Curves, Laplacian eigenmaps, ISOMAP, kernel PCA and other manifold learning methods) exist to deal with this case. The remarkably limited range of benchmark data typically used to justify and demonstrate these methods (e.g. toy data such as the 'swiss roll', images of facial expressions, 2-D projections of rotated 3-D objects, and hand written numerals such as in the NIST data set) suggest that the kind of nonlinear inter-variable correlation which gives rise to such manifolds is uncommon in real-world classification data sets. It has been our experience that, at least in the data sets typically used for classifier benchmarking, well-chosen linear transformation and projection onto R^2 often suffices to reveal clusters and outliers, and nonlinear manifold learning approaches do not contribute to significantly improved performance.

2. Multiplication of the data by a matrix of random numbers spans the space of all possible linear transformations. Normal- or uniformly-distributed numbers seem an obvious choice; we found the latter slightly more effective in rapidly finding good projections for class-separation. To project $[p \times d]$ data onto R^2 we postmultiplied by a uniformly random $[d \times 2]$ matrix. Adventitious correlation between the columns of the random matrix (especially in the case of lower-dimensional data) often resulted in a strongly elliptical projection, so we used singular value decomposition to transform the matrix into one with uniformly-random but mutually orthogonal columns of unit length (i.e. an orthonormal random matrix), ensuring that the projected data fully spans the projection space.

3. Manually observing a sequence of random projections is sometimes enlightening but tedious so we sought to automate the selection of the most favourable projection. Initially we tried Thornton's Separability Index s as a suitable criterion. This is a simple, rapidly computable measure of class separation, defined as the fraction of points which share a class label with their nearest neighbour. In a previous paper (2) we showed its effectiveness in feature subset selection; here we apply it to the planar transformed representation and use it as a

surrogate of visual separability. Empirical tests showed that it performs well in this role, with one potential limitation, relating to its limited 'dynamic range'.

For realistic data s is never less than 0.5 ($s = 0$ would require artificially constructed data such as an interlaced pair of grids of points with opposite class membership). At the other extreme, s 'saturates' at a value of unity. Complete separation, such that each point has a nearest neighbour of the same class, results in $s = 1$, which does not change with further distancing of the class centroids; thus s is insensitive to class separation as soon as the nearest-neighbour criterion has been realised. The limited range over which s is effective sometimes results in a projection which is not optimally effective in visual terms. To counter this we considered using a different criterion of separability – that of the 'hypothesis margin' h . This is the summed difference between the inter-class and intra-class nearest-neighbour pairwise distances. The hypothesis margin is an unbounded measure of separability so it does not exhibit the saturation effect exhibited by s . On the other hand it is less effective when the data classes overlap with a horizontal asymptote of zero. We conjectured that due to this complementary aspect of their behaviour, a hybrid separation index hs consisting of the sum of h and s might correlate better with visual separability over a wide range of datasets. This conjecture seemed to be supported by informal tests over a wide range of datasets from the repository. However the difference is rather marginal, so for speed and simplicity, Thornton's Separability Index s is used as a criterion of separation in the examples shown below.

4. The search for the optimal projection is non-convex, with many local maxima. Stochastic and evolutionary search methods are effective in this situation, and our first effort was to use Population-based Incremental Learning (a simple but powerful abstraction of the genetic algorithm). It was noticed however that frequently good solutions were obtained in the first generation, with little or no subsequent improvement. It appears

that this is one of the situations (feature selection can be another) where extremely sparse random searches in vast search spaces can nevertheless be effective, due to the multiplicity of local maxima which are competitive with the global optimum. If a substantial fraction of possible solutions are acceptable, it is likely that one of them will rapidly be encountered in a sparse random search. Pure random search, with its minimal computational overhead, may actually be more efficient in this situation than a more directed guided stochastic search (or a more principled search based on mathematical programming or gradient descent, which is almost certain to be entrapped at a sub-optimal point.

The algorithm is summarised in the following pseudocode:

Given data $X [p \times d]$, $t [p \times 1]$, ± 1
and a maximum number of iterations *maxiter* (a user-chosen parameter, typically 30-100)

iter = 0
max_sepindex = 0

```
while iter < maxiter & max_sepindex < 1
  create a uniformly random [p x 2] matrix R
  and transform it into an orthonormal matrix
  R = Rorth
  Xp = X * Rorth
  s = sepindex(Xp, t) (separability index)
  if s > max_sepindex
    max_sepindex = s
    Rbest = R
  end
  iter = iter + 1
endwhile
```

Plot column 1 vs Column 2 of $X_p = X * R_{best}$

The plots in the appendix show the results of applying the method to six datasets from the UCI repository. It can be seen that an effective visualisation results, with the dichotomous clustering of the data being vividly evident in the case of the Breast Cancer, Wine, Heart and Thyroid datasets. In the case of the Thyroid data it is further evident that the classes, though readily separable, require a nonlinear discriminant.

3. Extension to unlabelled data

Using class separability in visual space as a figure of merit requires labelled data. We tried extending the method to the visualisation of unlabelled data by using a spectral method (3) for partial dichotomous labelling of the data in the original representation space. This is achieved by a thresholding of the 2nd ('Fiedler') eigenvector of an affinity matrix formed by normalising a kernel matrix constructed using the data. It is based on the observation that points clearly belonging to different clusters are likely to have markedly different values in the Fiedler eigenvector. The subset of points so labelled is used to find the optimal transformation for visual mapping and this is applied transductively to the whole dataset.

The method proposed is as follows(3):

1. Construct a gaussian kernel (with the spread factor set to a user-determined factor of the modal nearest-neighbour distance as determined by histogram).
2. Use additive normalisation to convert the kernel to a Laplacian stochastic affinity matrix L.
3. Perform an eigenstructure decomposition on L.
4. Sort the second eigenvector (the 'Fiedler eigenvector') L₂ by value and reorder the dataset instances accordingly.
5. Label a small fraction of the topmost and bottom-most instances with opposite class labels, leaving the instance in-between unlabelled. This results in a partial labelling of the dataset, with the labelled instances being assigned to distinct classes with a high degree of confidence.
6. Perform the random projection method outlined above using only the instances labelled in the step above.

7. Plot all the points are plotted using the random transformation matrix which was found to optimise the visual separation of the points on the subset of data in the previous step.

This process requires a small modification to the calculation of visual separability, since with the limited labelled dataset constructed a separability index of 1 is rapidly reached, resulting in a loss of information about the merits of further transformations. Instead of the more complex hybrid separability index proposed above, Thornton's separability index was augmented, once it reached the value of 1, by adding the Euclidean distance between class means. Thus an unbounded measure of separability results, which continues to guide the search in the direction of greater separation even when complete separability has been achieved for the spectrally-labelled data.

A problem remains with the above proposal, which is still under investigation. Spectral clustering methods, though very powerful, depend fairly critically on the parameter σ of the gaussian kernel used to determine the affinity matrix, and it is unclear, in the present setting, how an optimal value for σ should be obtained. To circumvent this problem and test the proposed method in broad principle, an approximation was used in which the leading eigenvalue of the input correlation matrix $X^T X$ was used as a surrogate for the Fiedler vector.

Results were very promising with the Wisconsin breast-cancer dataset, and excellent visualizations were obtained (virtually indistinguishable from the results illustrated above) without the use of class labels. However further work is required to see how well the method generalises to other datasets. It is conjectured that a robust solution will require the use of the full spectral method with a more principled approach to selecting σ . A possibility under investigation is to set the value of σ equal to a fixed fraction of the modal pairwise nearest neighbour distance as determined by a histogram. Another possibility is to adjust sigma using a maximisation of the 'eigengap strategy'. These are all under investigation.

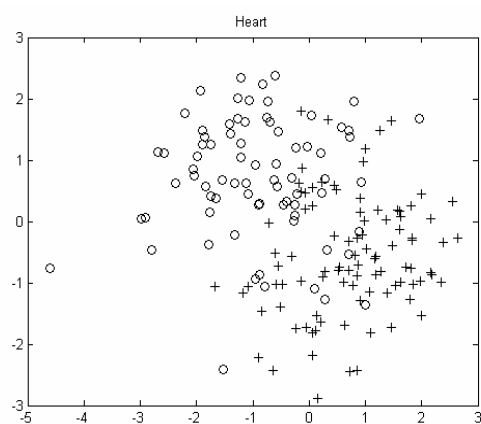
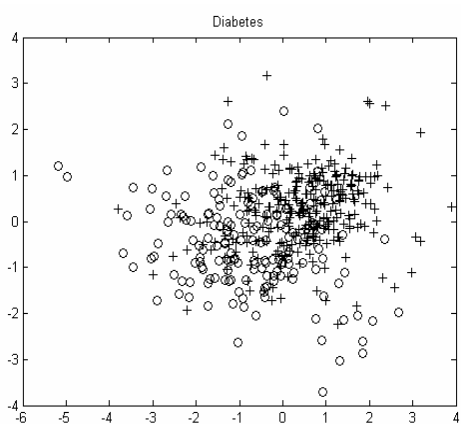
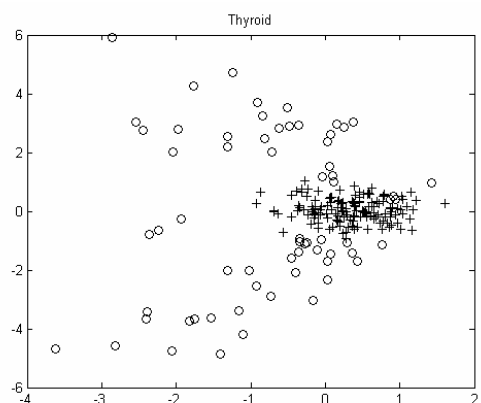
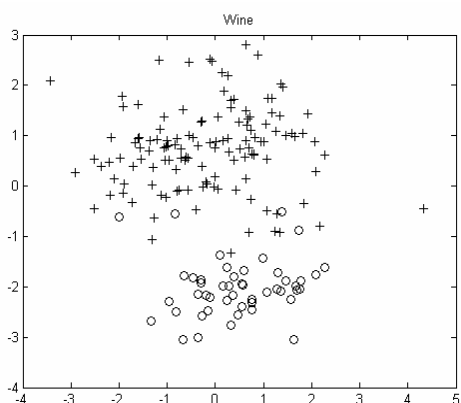
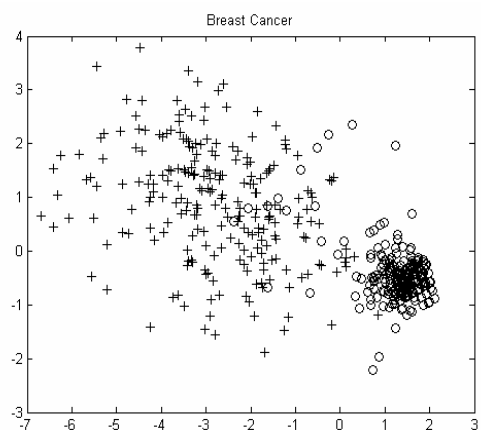
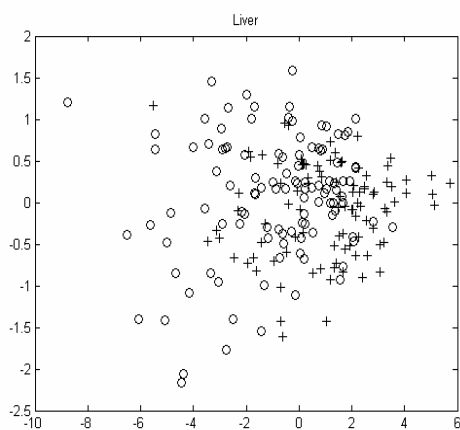
4. Conclusion

A simple but effective method has been presented for visualising structure (dichotomous clustering, outliers) in high-dimensional data making use of class labels, and demonstrated on some datasets. It is related to well-known projection pursuit methods, but makes efficient use of low-overhead random search, and employs Thornton's separability index as a novel interest measure. It has also been proposed that the method can be extended to unlabelled datasets by inferring a partial labelling using a spectral method, and applying the mapping so found transductively to the dataset as a whole. The method was implemented in broad principle and shown to be effective on a single dataset, but much work remains to be done solve some outstanding problems and to show that the extension is both robust and generally applicable.

5. References

1. Y Koren and L Carmel, "Visualization of Labelled Data Using Linear Transformations". *Proceedings of IEEE Information Visualization Conference InfoVis03*, (2003), IEEE pp121–128, 2003.
2. J Greene, "Feature subset selection using Thornton's separability index and its applicability to a number of sparse proximity-based classifiers". *Proceedings of the Pattern Recognition Association of South Africa*, PRASA, Franschhoek, 2001.
3. D Verma and M Meila, "A Comparison of Spectral Clustering". *Technical Report 2003.UW CSE Technical Report 03-05-01*
4. C.J.C. Burges. "Geometric methods for feature extraction and dimensional reduction: A guided tour". Microsoft research Technical report, 2004.

Appendix 1. Result of the visualisation method for labelled data



In each case the visualization provides useful information on the distributional structure. In the case of the Breast Cancer, Wine, Thyroid and Heart datasets, dichotomous clustering was very clearly visible, with visual separability indices in the range 0.5-1. It is clear that the Breast Cancer, Wine and Heart datasets are linearly separable, while the Thyroid data is readily separable, but requires a non-linear discriminant. In all cases 50 iterations sufficed to find an optimal projection.