ISBN: 978-972-8924-40-9 ©2007 IADIS

EVALUATION OF PROJECTION TECHNIQUES USING HUBERT'S Γ STATISTICS

Dorina Marghescu
Turku Centre for Computer Science, Åbo Akademi University, Dept. of IT

ABSTRACT

Projection techniques reduce the data dimensionality by combining the original variables into a smaller number of new dimensions, in a linear or nonlinear manner. The projection methods are particularly useful because they lend themselves to visual representations of data, when the number of new dimensions is one, two or three. In this paper, the aim is to evaluate different visualization techniques based on projection techniques with respect to their effectiveness in preserving the inherent relationships and structure of the dataset. For this purpose, we investigate the use of the Hubert's Γ statistics for evaluating the fit between the distance matrices of original data and projected data. Moreover, we investigate the use of the modified Hubert's Γ statistics for evaluating the effectiveness of projection techniques in preserving the clustering structure inherent in the dataset, if such structure is present.

KEYWORDS

Projection techniques, visualization, evaluation, Hubert's Γ statistic

1. INTRODUCTION

In this paper, we illustrate the use of projection techniques for visualizing high-dimensional data and evaluate their effectiveness in preserving the inherent relationships and structure that exist in the dataset. We study the use of *Hubert's Γ statistics* in order to *objectively* assess the effectiveness of the projection techniques in preserving the relationships and structure in the data. More specifically, we investigate the use of the Hubert's Γ statistics for evaluating the *fit between the distance matrices* of original data and projected data. We also examine the use of the modified Hubert Γ statistics for evaluating the effectiveness of projection techniques in preserving the *clustering structure* inherent in the dataset, if such structure is present.

Section 2 illustrates the projection techniques under analysis. Section 3 presents the definitions of Hubert's Γ statistics and its modified version. Section 4 proposes the procedures for calculating Hubert's Γ statistics and its modified version for evaluating the effectiveness of the projection techniques. Section 5 presents the evaluation results. We conclude with final remarks and future work ideas in Section 6.

2. PROJECTION TECHNIQUES

Projection methods are used to reduce the variable space (i.e., the original variables are combined in a linear or nonlinear manner into a smaller number of new data dimensions). The projection methods are particularly useful because they facilitate visual representations of data. In this section, we look at two classical projection techniques: Principal Components Analysis (PCA) and Sammon's mapping, and three more recently developed techniques: Self Organizing Maps (SOMs), Radviz and Star Coordinates.

For illustrating the capabilities of each projection technique, we use the Iris dataset (Newman et al. 1998), due to its suitability for classification and clustering tasks. The data concerns three species of flowers characterised by four attributes: petal length and width, and sepal length and width. The class variable is the type of flower: Iris-Setosa, Iris-Versicolor, and Iris-Virginica. The dataset contains 150 observations, each class containing 50 flowers. The class Iris-Setosa is linearly separable from the other two classes, but Iris-Versicolor and Iris-Virginica classes are not linearly separable.

2.1 PCA

PCA is a classical statistical technique that maps high-dimensional data items onto a lower-dimensional space (Sharma 1995). The transformation tries to preserve the variance of the original data as well as possible. The PCA technique creates new variables (called principal components), which are linear composites of the original variables. The maximum number of new variables that can be formed is equal to the number of original variables, and the new variables are uncorrelated among themselves. Figure 1 represents the PCA projection of the standardized data using the first two principal components.

2.2 Sammon's Mapping

The Sammon's mapping (Sammon 1969) is a metric multidimensional scaling technique that tries to match the pairwise distances of the lower-dimensional representations of the data items with their original distances. Figure 2 illustrates the Sammon's mapping of the Iris dataset, after applying variance normalization to the original data and using the Euclidean distance in the Sammon's mapping algorithm.

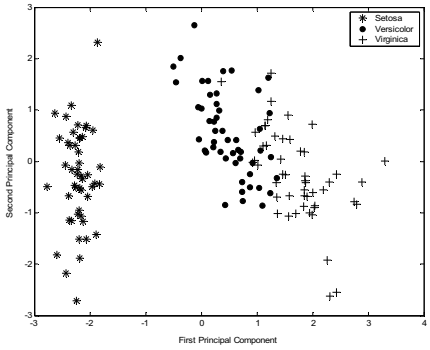

Figure 1. PCA

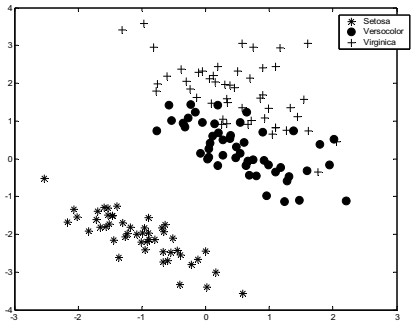

Figure 2. Sammon's mapping

2.3 SOM

Developed by Kohonen in 1982, SOM is a technique widely used for clustering and abstraction based on unsupervised learning neural network (Kohonen 2001). The SOM represents the data items on a two-dimensional grid, where each item is assigned to a node of the grid in an orderly way so that similar data items are mapped to the same node or neighbouring nodes. Figure 3 represents the Iris dataset on a SOM grid of 12x9 nodes. The technique of jittering was used to change with a small value the position of each data item so that the items mapped to the same node will not overlap. The data was first normalized using the variance method. Other parameters of the SOM were initialised as follows: Gaussian neighbourhood, radius [12, 1], batch training, and linear initialization.

2.4 Radviz

The technique was developed by Hoffman et al. (1997). The *n*-dimensional data items are represented as points in a two-dimensional space. The points are located within a circle whose perimeter is divided in *n* equal arcs. The equally spaced points on the perimeter are called *anchorpoints* or *dimensional anchors* (Hoffman et al. 1999), they being associated with each data dimension. The values of each data dimension must be normalized to range within [0…1]. The data item is connected to the *n* anchorpoints through *n* different *springs*. Each data point is then displayed at the position that produces a spring force sum of zero. If all *n* coordinates have the same value (regardless whether they are low or high), the data point lies exactly in the centre of the circle. If the point is a unit vector point, it lies exactly at the fixed point on the edge of the circle, where the spring for that dimension is fixed. Figure 4 shows the Radviz projection of the Iris dataset, after local normalization of the data in the range [0 1].

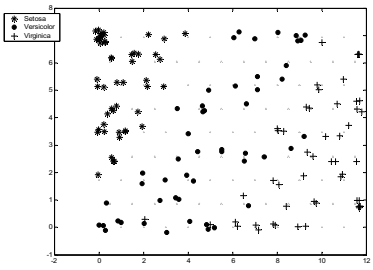

Figure 3. Self-Organizing Map

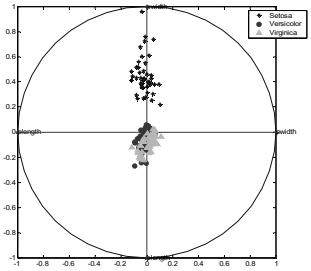

Figure 4. Radviz

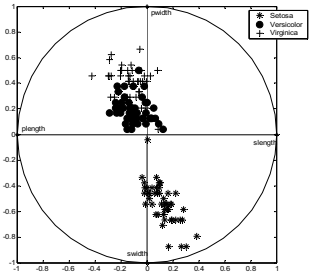

Figure 5. Star Coordinates

2.5 Star Coordinates

Star Coordinates (Kandogan 2001) maps *n*-dimensional data onto a two-dimensional space. The *n* coordinate axes are arranged on a two-dimensional plane, such that all axes share the same origin point, but they are not necessarily orthogonal to each other. The minimum value on each dimension is mapped to the origin, and the maximum value is mapped to the other end of the coordinate axis. Each variable is normalized to range within [0…1]. Each image point corresponding to a data item has a location on the two-dimensional plane determined by the sum vector of all unit vectors on each coordinate, multiplied by the value of the data item for that coordinate. Figure 5 displays the Star Coordinates projection of the Iris data, after normalization.

3. HUBERT'S Γ STATISTICS AND ITS MODIFIED VERSION

Hubert's Γ statistic (Hubert and Schultz 1976) is an index that measures the correlation between two matrices, *A* and *B*, of dimensions *N x N*, drawn independently of each other (Theodoridis and Koutroumbas 1999). Theodoridis and Koutroumbas discuss the use of this type of statistics as external and internal criteria for assessing clustering validity. For two symmetric matrices, *A* and *B*, this statistic is defined as follows:

$$Hubert's\ \Gamma = (1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} A(i,j)B(i,j) \qquad (1),$$

where $A(i,j)$ and $B(i,j)$ are the (i,j) elements of matrices *A* and *B*, and $M = N(N-1)/2$. High values of Γ indicate close agreement between *A* and *B*. The normalized Γ statistic, denoted $\hat{\Gamma}$, can also be used.

$$\hat{\Gamma} = \frac{(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(A(i,j)-\mu_A)(B(i,j)-\mu_B)}{\sigma_A\sigma_B} \qquad (2),$$

where $\mu_A, \mu_B, \sigma_A, \sigma_B$ represent the means and squared variances of *A* and *B*, respectively. $\hat{\Gamma}$ has values between -1 and 1. Large absolute values of $\hat{\Gamma}$ suggest agreement between the matrices *A* and *B*.

Theodoridis and Koutroumbas (1999) discuss also the **modified Hubert's Γ statistic** as a *relative* measure to compare different clustering solutions obtained by using different clustering algorithms. Let *X* be the original dataset containing *N* data items and *P*, its *proximity matrix*, where the element $P(i,j)$ is the distance between two vectors x_i and x_j of *X*. By applying a clustering algorithm to the dataset *X*, we obtain the partition *C*, in which individual clusters are denoted $C_1, C_2, \ldots, C_m$, *m* being the total number of clusters. Let *Q* be the *N x N* matrix whose (i,j) element, $Q(i,j)$, is equal to the distance $d(w_{c_i}, w_{c_j})$ between the representatives w_{c_i}, w_{c_j} of the clusters where x_i and x_j belong. The same distance measure must be used for both *P* and *Q*. The modified Hubert's Γ statistic and its normalized version are defined using the Eq. (1) and (2), respectively, where the matrix *A* is the proximity matrix *P* of the dataset and the matrix *B* is the matrix *Q*. The normalized modified $\hat{\Gamma}$ has values between -1 and 1. Large absolute values of $\hat{\Gamma}$ suggest agreement between the matrices *P* and *Q*.

4. PROPOSED APPROACH TO EVALUATION OF PROJECTIONS

In this section, we propose two procedures for evaluating the projection techniques based on *Hubert's Γ and modified Hubert's Γ statistics*. First, we use the Hubert's Γ to evaluate the extent to which the obtained projected data preserved the inherent relationships between the data points, measured in terms of *proximity matrix*. The **proximity matrix** of a dataset *X* is a symmetric matrix consisting of the pair-wise distances of elements of *X*. Second, we use the modified index to evaluate the extent to which the obtained projected data preserves the clustering structure inherent in the data, if such structure exists.

Both procedures require that the original data dimensions are standardized (by subtracting the mean and dividing by standard deviation), before calculating the proximity matrix of original data. Moreover, the procedures include the normalization of the proximity matrices using the *global histogram equalization* method. This method works in two steps: first, the data values are replaced by the order index, and then these values are normalized to be in the range [0, 1], by applying a linear transformation.

4.1 Procedure for Calculating Hubert's Γ and $\hat{\Gamma}$ Statistics for Projected Data

1. Calculate the proximity matrix *Dx* of the original *standardized* data.
2. Calculate the proximity matrix *Dp* for projected data.
3. Normalize *Dx* and *Dp* using *global histogram equalization* method so that the distances are comparable.
4. Calculate the Hubert's Γ statistic by applying Eq. (1) to normalized proximity matrices *Dx* and *Dp*. The value of this statistic will indicate the extent to which the two proximity matrices (of original data and projected data) reflect the same inherent relationships of the data points.
5. Similarly, we can calculate the normalized Hubert's Γ statistic given in Eq. (2) using the normalized proximity matrices *Dx* and *Dp*.
6. Repeat steps 1-5 for each obtained projected dataset.
7. The higher the Hubert Γ (or the absolute values of $\hat{\Gamma}$) statistics, the better is the preservation of the data relationships after projection.

4.2 Procedure for Calculating Modified Hubert's Γ And $\hat{\Gamma}$ Statistics for Projected Data after Clustering

1. Apply a clustering algorithm (e.g., K-means algorithm (MacQueen 1967)) on projected data.
2. Calculate the proximity matrix *Dx* of the original *standardized* data.
3. Calculate a *N x N* matrix *Q*, whose (i, j) element $Q(i, j)$ is equal to the distance $d(w_{c_i}, w_{c_j})$ between the representatives of the clusters where x_i and x_j belong. The same distance measure must be used for calculating both *Dx* and *Q*.
4. Normalize *Dx* and *Q* using *global histogram equalization* method so that the distances are comparable.
5. Calculate the modified Hubert's Γ statistic by applying Eq. (1) to the normalized proximity matrices *Dx* and *Q*. This statistics will indicate the extent to which the clustering of the projected data reflects the inherent structure of the original data (given by the normalized proximity matrix *Dx*).
6. Similarly, we can calculate the normalized modified Hubert's Γ statistic by applying Eq. (2) to the normalized proximity matrices *Dx* and *Q*.
7. Repeat steps 1-6 for each obtained projected dataset.
8. The higher the modified Hubert Γ (or the absolute value of $\hat{\Gamma}$) statistics, the better is the preservation of the data structure after projection.

To investigate the use of these indices for the evaluation of projection techniques, we performed a number of experiments on known datasets. The computations were realized in Matlab (The MathWorks 2000). For obtaining the SOM and Sammon's mapping, we used the SOM Toolbox (Vesanto et al. 1999).

5. RESULTS

Three datasets with clustering structure available at UCI Machine Learning Repository (Newman et al. 1998) were used for analysis in this study: *Iris dataset*, *Voting records database* and *Wine recognition data*. We also used two artificially created datasets, one which contains three well separated clusters (*Artificial 1*) and another that does not have a clustering structure (*Artificial 2*). **Artificial 1** dataset contains 150 data points of dimensionality 4. The data was randomly generated from three normal distributions with known means and variances. Therefore, the dataset consists of three distinct groups, each group having 50 data points. The means and variances characterising each group in this dataset are given by the matrices M_1 and V_1. M_1 represents the means of the 4 variables (columns) in each group (rows), and V_1 represents the respective variances:

$$M_1 = \begin{bmatrix} 1 & 2 & 3 & 3 \\ 4 & 7 & 8 & 8 \\ 10 & 10 & 15 & 15 \end{bmatrix} \text{ and } V_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Artificial 2 dataset contains 200 data points of dimensionality 5. The data was randomly generated from one normal distribution with means $M_2 = [1\ 2\ 3\ 3\ 5]$ and variances $V_2 = [1\ 1\ 1\ 1\ 1]$.

5.1 Results of Normalized Hubert's Γ Statistics for All Datasets

Table 1 shows the values of the normalized Hubert's Γ statistics calculated for the proximity matrices of the projected datasets. We used for calculating these indices the procedure 4.1. It is observed that for these datasets, the PCA, Sammon's mapping and SOM are the most effective techniques in preserving the inherent relationships between the data points calculated in terms of Euclidean distance.

Table 1. Normalized Hubert's Γ statistics of projected data (the higher the index, the better the projection)

	Iris	*Vote*	*Wine*	*Artif. 1*	*Artif. 2*
PCA	**0.98**	**0.81**	**0.82**	**0.99**	0.64
Sammon's mapping	**0.97**	**0.84**	**0.89**	**0.99**	**0.76**
SOM	**0.96**	**0.77**	**0.78**	**0.97**	0.55
Radviz	0.74	0.49	0.68	0.32	0.63
Star Coordinates	0.83	0.59	0.61	0.20	0.64

Table 2. Normalized modified Hubert's Γ statistics after clustering (the higher the index, the better the obtained clustering)

	Iris	*Vote*	*Wine*	*Artif 1*	*Artif 2*
Original	**0.88**	**0.65**	0.33	**0.93**	0.26
PCA	**0.87**	**0.65**	**0.62**	**0.93**	0.26
Sammon's mapping	**0.86**	**0.65**	**0.63**	**0.93**	0.26
SOM	**0.87**	**0.64**	**0.62**	**0.93**	0.24
Radviz	0.75	0.42	**0.61**	0.21	0.28
Star Coordinates	**0.80**	0.56	0.51	0.11	0.25

The results also reveal that the poorest projection for Iris dataset is given by Radviz technique (see also Figure 4). By examining in parallel Figures 1-5 with the results showed in Table 1, we observe that the values of the indices are good indicators of the effectiveness of projection techniques in preserving the relationships between data points. It is interesting to observe that in the case of Artificial 2 data (which does not have a clustering structure) the best projection technique that preserves the original distances between the data points is Sammon's mapping. This result confirms the capability of Sammon's mapping of preserving the pair-wise distances of the data points. In this case, SOM technique has the lowest performance.

5.2 Results of Modified Hubert's Γ Statistics for All Datasets

Table 2 shows the values of the normalized modified Hubert's Γ statistic obtained after clustering the data using the K-means algorithm. We used for calculating these indices the procedure 4.2, in which Euclidean distance was chosen for calculating the proximity matrices. The column *Original* contains the values of the

indices after applying the clustering on the original dataset. It is observed that for these datasets, the PCA, Sammon's mapping and SOM are the most effective techniques in preserving the inherent clustering structure that exists in the datasets. Because Artificial 2 dataset does not have a clustering structure, the indices have low values. In the case of Wine data, the clustering of the original data does not reveal the true clustering structure of the dataset, and the use of any projection technique may be used to visualize and detect the clusters inherent in the dataset.

The results obtained for all these indices (Tables 1 and 2) are dependent on the selection of a number of parameters. These parameters may regard the projection techniques, the distance metric used for calculating proximity matrices, and the clustering algorithm.

6. CONCLUSION

In this paper, we investigated the use of Hubert's Γ statistic and its modified version for evaluating the effectiveness of five projection techniques (PCA, Sammon's Mapping, SOM, Radviz, and Star Coordinates) in data relationships and structure preservation. We illustrated the use of these projection techniques on the well known Iris dataset, in order to facilitate visual comparison of the capabilities of each projection technique with respect to their effectiveness for solving a data-mining task such as clustering. Because this visual inspection is highly subjective, we investigated the use of objective measures such as Hubert's Γ statistics for the evaluation.

We proposed two procedures for calculating the Hubert's Γ statistics and its modified version in order to assess the extent to which the projected data preserves, first, the inherent relationships and, second, the structure that exists in the dataset. The results showed that our approach can be used to evaluate objectively the data visualizations based on projection techniques. The performance of the projection techniques depends on the dataset under analysis, but generally the PCA, Sammon's mapping and SOM were found the most effective projections for our datasets. The evaluation approach proposed in this paper can be used to assess whether a given projection is good enough in preserving the data relationships and structure, before using that projection for further processing the data. For future work, we intend to use this approach in investigating the capabilities of other projection techniques to represent different high-dimensional datasets.

REFERENCES

Hoffman, P. E., Grinstein, G. G., Marx, K., Grosse, I., and Stanley, E., 1997. DNA Visual and Analytic Data Mining, *Proceedings of IEEE Visualization '97*, Phoenix, Az, 437–441.

Hoffman, P., Grinstein, G., and Pinkney, D., 1999. Dimensional Anchors: a Graphic Primitive for Multidimensional Multivariate Information Visualizations, *Proc. 1999 Workshop on New Paradigms in Information Visualization and Manipulation, in Conjunction with the 8h ACM Int'l. Conf. Information and Knowledge Management (CIKM '99),* pp. 9-16.

Hubert, L. J., and Schultz, J., 1976. Quadratic Assignment as a General Data-Analysis Strategy, *British Journal of Mathematical and Statistical Psychology*, Vol. 29, pp. 190-241.

Kandogan, E., 2001. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates, *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 107-116.

Kohonen, T., 2001. *Self-Organizing Maps, 3rd* ed. (Berlin; New York: Springer).

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations, *Proc. Fifth Berkeley Symposium on Math. Stat. and Prob., vol. 1*, pp. 281-297.

The MathWorks, 2000. *MATLAB-The Language of Technical Computing* (Natick, MA: The MathWorks, Inc).

Newman, D. J., Hettich, S., Blake, C. L. and Merz, C. J., 1998. UCI Repository of machine learning databases, http://www.ics.uci.edu/~mlearn/MLRepository.html.

Sammon, J. W., 1969. A nonlinear mapping for data structure analysis, *IEEE Trans. Comput., vol. C-18*, pp. 401-409.

Sharma, S., 1995. *Applied Multivariate Techniques* (New York, NY: John Wiley & Sons, Inc.).

Theodoridis, S. and Koutroumbas, K., 1999. *Pattern recognition* (Academic Press, San Diego, CA).

Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J., 1999. Self-organizing map in Matlab: the SOM toolbox, *Proceedings of the Matlab DSP Conference,* pp. 35-40.