Compression and Visualization of High-Dimensionality Data Using Auto-Associative Neural Networks

Zalhan Mohd Zin[1], Marzuki Khalid[2], Ehsan Mesbahi[3] and Rubiyah Yusof[2]
[1]Section of Industrial Automation – UniKL-Malaysia France Institute (UniKL-MFI),
[2]Center for Artificial Intelligence and Robotics (CAIRO) – Universiti Teknologi Malaysia (UTM)
[3]Newcastle University, Newcastle Upon Tyne, UK

Abstract

Interpreting the information hidden in multidimensional data can be considered as a challenging and also a complicated task. The compression, dimension reduction and visualization of these multidimensional data provide ways to better understanding and interpretation of the problem. Usually, dimension reduction or compression is considered as the first step to data analysis and exploration. Here, the focus is given on multidimensional data reduction using a supervised artificial neural networks technique namely the Auto-Associative Neural Networks (AANN). The AANN can be considered as very powerful tool in exploratory data analysis. It has the ability to deal with linear and nonlinear correlation among variables. This technique is often referred to as nonlinear principal component analysis (NLPCA) or sometimes is also known as the bottleneck neural network due to its specific structure that consists of a combination of two networks – compression and decompression. By using this structure, AANN can reduce high dimensional data onto lower dimensional data by compressing them on its bottleneck layer that later can be used for data visualization and interpretation. In this paper, the technique of AANN is described, developed and applied on two different multidimensional datasets. The results have shown that the AANN is able to compress multidimensional data into only two nonlinear principal components at its bottleneck layer and these compressed data can provide visualization of different clusters of data.

Keywords—Auto-Associative Neural Networks, Dimension Reduction, Data Clustering, Iris flowers, Olive oils

1. Introduction

The ability of artificial intelligence techniques to extract information hidden in high-dimensional data is rather interesting to be investigated and explored. The compression, dimension reduction and visualization of these multidimensional data could provide ways to better understanding and interpretation of the data itself. There are many techniques of dimension reduction that have been used such as the famous principal component analysis (PCA) (1) (2) (3) (4) (5). However, linear PCA technique is limited to only linear correlated variables. To deal with nonlinear variables problems, nonlinear dimension reduction or projection methods should be used. Some examples of nonlinear dimension reduction techniques are Multidimensional Scaling (MDS), Locally Linear Embedding (LLE), Isomap, Kernel PCA (6), Self-Organizing Maps (SOM) (7) (8) (9) and also Auto-Associative Neural Networks (AANN) (10) (5) (3). AANN, is also known as Bottleneck Neural Networks (BNN), has been usually used for data compression in chemical applications, missing data estimations and image compressions (3) (11) (12). In this research, the focus was given to AANN for multidimensional data compression, clustering and visualization. The technique has been described, developed using high level computer language, applied and analyzed by computing it on high-dimensional data of Iris flowers and Italian olive oils datasets.

2. Methods

The AANN has often been regarded as an alternative to PCA for unsupervised learning, clustering, and outlier detection (13). As described in (5), it is also known as Non Linear Principal Component Analysis (NLPCA) that can improve the performance of data compression. This is because it can deal with nonlinear data problems more efficiently than PCA. AANN is a feed forward network, which is trained to map approximations of input vectors to their corresponding outputs. It can be viewed as circuits of highly interconnected units with adjustable interconnection weights and can be considered as a particular class of artificial neural networks in which the target output patterns is identical to its input patterns (14). The AANN actually works by trying to imitate human brain abilities such as learning and memorizing. Similarly to the biological neural system, the neuron is the smallest functional unit in the AANN. The structure of AANN usually consists of several layers including hidden layers such as input layer, map or compression layer, bottleneck layer, de-map or de-compression layer and finally output layer. This number of nodes at bottleneck layer must have smaller dimension than the number of nodes at input or output layers. During training of AANN, the dataset is compressed to few potential variables, number of which corresponds to the number of nodes in the bottleneck layer. Then the output of the bottleneck is decompressed at the de-mapping layer. Once the training is completed, the

bottleneck outputs of AANN represent a type of nonlinear principal components, which are frequently more relevant than PCA for analyzing nonlinear, real-world datasets. They can be projected as visualization of data clusters.

3. Results

The experiments in this paper used the AANN to cluster high-dimensional data which has been conducted using the structure that consisted of five layers: two layers of input and output and three hidden layers of map, bottleneck and de-map layers. The result of iris flowers dataset's (15) projection by the two nodes at the bottleneck layer is shown in Fig. 1. The high dimensionality iris flowers data has been reduced by the AANN to only two dimensions represented by these two bottleneck neurons. The Setosa and Virginica flowers have been appropriately clustered far from each other, while the class of Versicolor flower has been clustered between these two classes, and closer to Virginica class. This distribution of iris flowers species was consistent with the nature of the iris flowers dataset itself (15). Meanwhile, the projection of Italian olive oils dataset has also resulted similar and consistent behaviour with the results obtained in (3). The data was taken from nine different growing regions in Italy which are North Apulia, Calabria, South Apulia, Sicily, Inland Sardinia, Coastal Sardinia, East Liguria, West Liguria, Umbria.

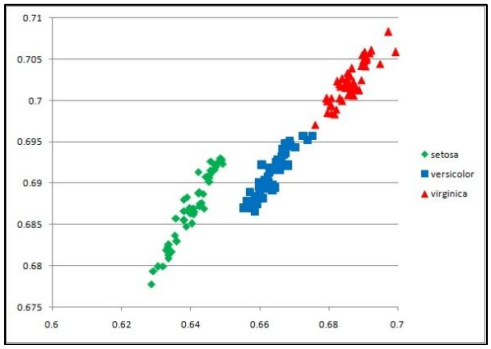

Fig. 1: The projection of iris flowers dataset defined by the two activation nodes at bottleneck layer in AANNs.

4. Conclusions

In this research, the compression ability of AANN has been used to compress the multidimensional data into few principal components which later were projected for visualization of the clusters of data. The research mainly involved in the studying, developing and applying AANN algorithm using high level computer language. From the experimental results, the AANN has been able to cluster iris flowers and Italian olive oils datasets according to their three and nine different classes respectively. This could give us the way to explore the potential of AANN to create a model of multidimensional data clustering. Additional research could also be undertaken to apply and analyze the AANN with other clustering technique such as the Self Organizing Maps (SOM) (7) that might be useful for clustering other complicated type of data such as Gene expression profiles.

References

1. **Smith, Lindsay I.** *A Tutorial on Principle Component Analysis.* [Online] 2002. [Cited: 05 February 2012.] http://neurobot.bio.auth.gr/2005/a-tutorial-on-principal-components-analysis/.

2. *A Comparison of Self Organizing Map Algorithm and Some Conventional Statistical Methods for Ecological Community Ordination.* **J. L. Giraudel, S. Lek.** s.l. : Elsevier, 2001, Journal of Ecological Modelling, Vol. 146, pp. 329-339.

3. *A Journey Into Low-dimensional Spaces With Autoassociative Neural Networks.* **M. Daszykowski, B. Walczak, D. L. Massart.** s.l. : Elsevier Inc., 2003, International Journal of Pure and Applied Analytical Chemistry Talanta, Vol. 59, pp. 1095-1105.

4. *Missing Data Estimation Using Principle Component Analysis and Autoassociative Neural Networks.* **Jaisheel Mistry, Fulufhelo V. Nelwamondo, Tshilidzi Marwala.** 3, s.l. : iiisci.org, 2009, Journal of Systemics, Cybernatics and Informatics, Vol. 7, pp. 72-79.

5. *Nonlinear Principle Component Analysis Using Autoassociative Neural Networks.* **Kramer, Mark A.** 1991, AIChE Journal, Vol. 37, pp. 233-243.

6. *A Framework for High Dimensional Data Reduction in the Microarray Domain.* **Ali Anaissi, Paul J. Kenned, Madhu Goyal.** s.l. : IEEE Explore, 2010. IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA). pp. 903-907.

7. *The Self-Organizing Maps.* **Kohonen, Teuvo.** s.l. : IEEE Xplore, 1990. Proceedings of the IEEE. Vol. 78, pp. 1464-1480.

8. *On the Use of Self Organizing Maps for Clustering and Visualization.* **Flexer, Arthur.** Prague, Czech Republic : Springerlink, 1999. International Conference on Principle on Data Mining and Knowledge Discovery. Vol. 5, pp. 80-88.

9. *Clustering of the Self Organizing Maps.* **Juha Vesanto, Esa Alhoniemi.** 3, 2000, IEEE Transaction on Neural Networks, Vol. 11, pp. 586-600.

10. *The Autoassociative Neural Network-A Network Worth Considering.* **Stone, Victor M.** Hawaii : s.n., 2008. World Automation Congress (WAC).

11. *Non-linear PCA: A Missing Data Approach.* **Matthias Scholz, m Fatma Kaplan, Charles L. Guy, Joachim Kopka.** 20, 2005, Journal of Bioinformatics, Vol. 21, pp. 3887-3895.

12. *Image Compression Using Hybrid Neural Networks Combining The Auto-Associative Multilayer Perceptron and The Self Organizing Feature Map.* **M. A. Abidi, S. Yasuki,P. B. Crilly.** 4, 1994, IEEE Transaction on Consumer Electronics, Vol. 40, pp. 796-811.

13. *Augmented Efficient BackProp for Backpropagation Learning in Deep Autoassociative Neural Networks.* **Mark J. Embrechts, Blake J. Hargis, Jonathan D. Linton.** Barcelona : s.n., 2010. International Joint Conference on Neural Networks (IJCNN). pp. 1-6.

14. *Feature Extraction Using Auto-Associative Neural Networks.* **Gaetan Kerschen, Jean-Claude Golinval.** 1, 2004, Smart Materials and Structures, Vol. 13.

15. **Fisher, R. A.** UCI Machine Learning Repository. *Center for Machine Learning and Intelligent System University of California Irvine.* [Online] 2006. [Cited: 31 January 2011.] Iris Database. http://archive.ics.uci.edu/ml/datasets/Iris.