EC601 - PRODUCT DESIGN IN ECE
MINI PROJECT 2

A Survey of Dimensionality Reduction Techniques

INTRODUCTION

Artificial Intelligence is a software that learns and self-improves. It may take a physical form but, in many cases, it is used in services that have no physical form like bots. With the rise of Artificial Intelligence, Machine Learning has become a popular topic. It is a class of software that can self-improve with exposure to useful data and thus forms the basis of AI [4]. The common types of machine learning are listed below:

1. Deep Learning
2. Ensemble Learning
3. Neural Network
4. Supervised Learning
5. Unsupervised Learning
6. Recursive Self-Improvement

From the above listed methods, unsupervised learning is being used extensively for dealing with modern big data problems. In order to solve many real world problems, unsupervised learning methods are being used in deep learning algorithms because of its ability to find the similarities in a data set which helps in easier analysis of new data and thus it has become a core part of the research of machine learning.

UNSUPERVISED LEARNING

Unsupervised Learning is a type of machine of learning method/algorithm which deals with unlabeled data to predict results i.e. the correct solution to the problem is not known beforehand and model works on its own to discover information (similarity among the data). In comparison to supervised learning which deals with data that has expected answers, unsupervised learning has to perform more complex tasks and can be a lot unpredictable. Because of these features, unsupervised learning has great application in the fields of:

- Visual recognition
- Human Behavior Analysis
- Robotics

Unsupervised learning has a lot of real-world applications/methods which help in better analysis of unlabeled data. These are:

1. Clustering
    a. K-Means Clustering
    b. Hierarchical Clustering
    c. K-NN (k nearest neighbors)
    d. Singular Value Decomposition
2. Visualization
3. Dimensionality Reduction

4. Finding Association Rules
5. Anomaly Analysis

In this report, we are going to explore some of the techniques that are used to reduce the dimensionality of datasets.

Many real-world machine learning problems have data (like speech signals, digital photographs, etc.) contain thousands of features for each training instance which affects the training speed of the model being developed and reduces its efficiency. In order to deal with such data adequately, its dimensionality needs to be reduced. ***Dimensionality reduction*** is a technique that is used to reduce the number of random variables in a data set under consideration and getting a set of variables that can completely define the data. Such variables are known as the *Principle Variables*. Ideally the dimensionality of the simplified data should correspond to the intrinsic dimensionality (minimum number of variables required to account for the observed properties) of the data [6].

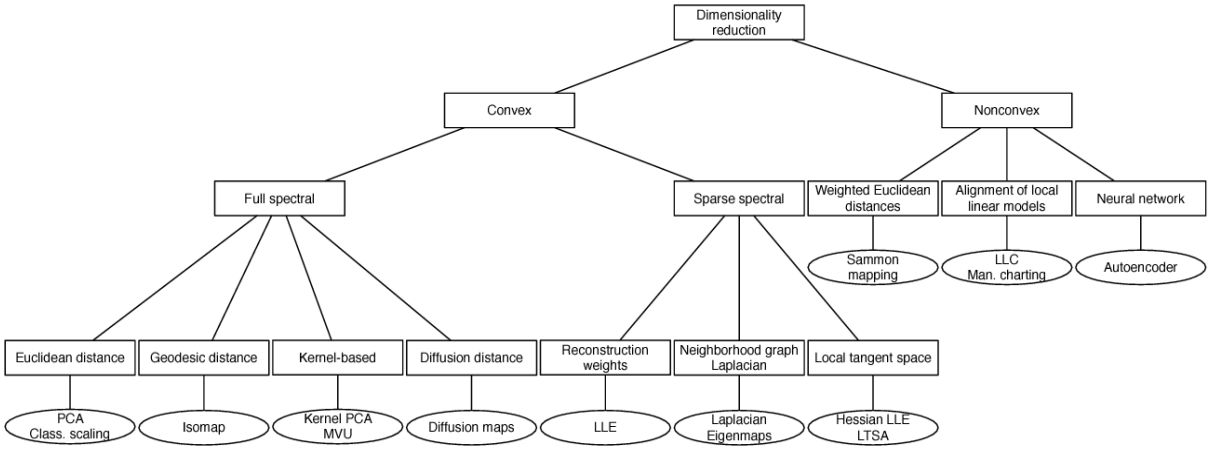

Fig 1. Taxonomy of Dimensionality Reduction Techniques [3]

Some of the traditional and most common linear algorithms used for dimensionality reduction are listed as below:

1. Principal Component Analysis (PCA)
2. Locally-Linear Embedding
3. Linear Discriminant Analysis

The above-mentioned techniques for dimensionality reduction make the data less complex, much faster and also reduces the memory requirement for the data. But one of the major drawbacks of these dimensionality reduction algorithms is the loss of information which eventually worsens the performance of the system for complex non-linear data. Therefore, some non-linear techniques such as Kernel PCA, Maximum Variance Unfolding [3], Random Projection (RP) have been developed to deal with such types of data sets. In this report, we will be limiting our discussion to the RP techniques only.

Random Projection technique maps the high-dimensional data on a low-dimensional subspace and the time requirement for this is very low. The essential idea of RP is based on the Johnson-Lindenstrauss lemma, which states that it is possible to project n points in a space of arbitrarily high dimensions onto an O(log n)- dimensional space such that the pairwise distances between points are approximately preserved [1].

In random projection, the original d-dimensional data is projected to a k-dimensional (k<<d) subspace through the origin, using a random k × d matrix R whose columns have unit lengths. Using matrix notations, if X_{dXN} is the original set of N d-dimensional observations, then:

$$X^{RP}_{kXN} = R_{kXd} X_{dXN}$$

is the projection of the data onto a lower k-dimensional subspace [2]. Apart from this, RP is computationally very simple. The complexity for this is of the order O(dkN). If data matrix X is sparse with c non-zero values per column, then the complexity is of the order O(ckN). In addition to this, the RP technique is not sensitive to impulse noise and hence it proves to be a great alternative for the existing noise reduction methods.

Some of the implementations of Random Projection are:

- RandPro: An R package for random projection
- sklearn.random_projection: Python module for random projection
- Weka implementation [5]

RECOMMENDATIONS

The people who want to work with the unsupervised learning algorithms to develop machine learning models can explore the Radom Projection technique for dimensionality reduction (one of the main methods required to train a much efficient and accurate model) when working with complex, non-linear data sets (image data, text data, etc.) in order to achieve low computation complexity, less time consumption and immunity to noise.
However, they must be careful while implementing this technique because the matrix generated with reduced dimensionality doesn't incorporate the intrinsic structure of the original data and this might lead to high distortion in the results.

CONCLUSION

This report on the non-linear techniques for Dimensionality Reduction in Unsupervised Learning has shown that Random Projection (RP) is a very efficient and powerful technique and can counteract the *Curse of Dimensionality* (high computation requirements for high-dimension real-world data with large number of variables). When working with such data sets, a dimensionality reduction technique is required which produces low distortion and less computation complexity.

Reference [1] discusses various methods to reduce the high distortion in results produced by RP with the use of various techniques such as *Feature Extraction Approaches*, *Dimensionality Increasing Approaches*, and *Ensemble Approaches*. In addition, this paper discusses about the future applications of RP in unsupervised learning like video recognition, speech and voice recognition.

Reference [2] discusses about the various real-world problems in which RP can be implemented to improve performance. They work with a number of data sets which have varying nature: noisy and noiseless images of natural scenes, and text documents from a newsgroup corpus.

Therefore, these both papers can together help towards developing a better Unsupervised Learning model for complex data sets which has low distortion and high computation efficiency.

SUMMARY OF REFERENCES

[1] Haozhe Xie, Jie Li, Hanqing Xue, "*A Survey of Dimensionality Reduction Techniques Based on Random Projection*"

[2] Ella Bingham and Heikki Mannila, "*Random projection in dimensionality reduction: Applications to image and text data*"

[3] Laurens van der Maaten, Eric Postma, Jaap van den Herik, "*Dimensionality Reduction: A Comparative Review*"

[4] https://simplicable.com/new/machine-learning

[5] https://en.wikipedia.org/wiki/Random_projection

[6] https://pythonistaplanet.com/applications-of-unsuperviseds-learning/