AN ANALYSIS OF TWO DIMENSIONALITY REDUCTION TECHNIQUES ON THE PERFORMANCE OF NEURAL NETWORK CLASSIFIERS

Oleh

Ong Siok Lan

Disertasi ini dikemukakan kepada
UNIVERSITI SAINS MALAYSIA

Sebagai memenuhi sebahagian daripada syarat keperluan
untuk ijazah dengan kepujian

SARJANA MUDA KEJURUTERAAN (KEJURUTERAAN ELEKTRONIK)

Pusat Pengajian Kejuruteraan
Elektrik dan Elektronik
Universiti Sains Malaysia **Mac 2005**

ABSTRAK

Projek ini berhubungkait dengan perbandingan antara dua teknik pengurangan dimensi sesuatu set data. Dua teknik yang terlibat ialah *Principal Component Analysis* sebagai teknik yang umum diaplikasikan manakala *Random Projection* merupakan teknik yang baru diperkenalkan. Kajian adalah berdasarkan keputusan daripada dua kawalan neural iaitu *Standard Backpropagation* dan *Fuzzy ARTMAP*. Data piawaian dan data pesakit digunakan dalam kajian ini. Keputusan daripada dua kawalan neural dikira berdasarkan *percentage of correct classification*, *purity*, and *collective entropy*. Pengujian hipotesis iaitu ujian *t* dilaksanakan untuk menguji perbezaan min antara dua min populasi berdasarkan sample yang terhasil daripada keputusan kawalan neural untuk *Principal Component Analysis* dan *Random Projection*. Keputusan yang sah berdasarkan pengujian hipotesis pada ralat, $\alpha = 0.05$ ataupun selang keyakinan 95%, dihasilkan dan ini menyumbang kepada kesimpulan yang kukuh dalam membuat perbandingan antara dua teknik pegurangan dimensi ini. Keputusan daripada data pesakit juga membuktikan *Random Projection* boleh diaplikasikan secara praktikal. Di samping itu, *Random Projection* juga meghasilkan keputusan yang setara berbanding *Principal Component Analysis*. Satu perbincangan disertakan untuk menerangkan keputusan yang diperolehi dan kesimpulan dibuat untuk kajian ini. Cadangan disertakan di akhir disertasi ini untuk perkembangan dan kajian pada masa hadapan untuk teknik pengurangan dimensi.

ABSTRACT

This project involves an analysis of the effectiveness of two dimensionality reduction techniques, i.e., Principal Component Analysis as the standard approach and Random Projection as a recent technique. The study is based on the performance of two supervised neural network classifiers i.e., Standard Backpropagation and Fuzzy ARTMAP. A set of benchmark and real medical databases are used to evaluate the performance of the neural network models. The performance indicators used are percentage of correct classification, purity, and collective entropy. The Student's two-tailed paired *t*-test is used to compare the significance of differences of the results. Based on the estimated 95% confidence intervals, a strong decision which eventually leads to a convincing conclusion on the performance of the dimensionality reduction techniques can be obtained. The perceived experimental results especially from the real medical data sets are encouraging enough to prove that Random Projection exhibits good performance as a dimensionality reduction technique. Surprisingly, Random Projection is effective on low dimensional data, and the outcomes are as good as Principal Component Analysis. A discussion on generalization of the results obtained is included, and a conclusion ensues. Recommendations are also included for further improvements and enhancements in the analysis of dimensionality reduction techniques.

ACKNOWLEDGEMENTS

Firstly, I would like to seize the opportunity to express my gratitude to my supervisor, Dr. Lim Chee Peng for his guidance, assistance, encouragements, and advices which allows me to learn throughout this project. Besides that, I would also like to thank the Linux laboratory assistant, Encik Mohd Roslan Abdul Rahman for giving me the permission to reserve a computer to run experiments. I would also to say thank you to Mr. Tan Yee Chyan, for his guidance in Linux basic programming language. Lastly, I would also like to thank all my friends who helped me throughout this project.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The ability of human recognizing objects such as handwritten characters, identifying faces, differentiating sounds, distinguishing animals, fruits, plants, etc based on important features and patterns involves the art of biological nervous system and a complex process in the human brain. The recognition task, applied by researchers on systems known as neural networks by presenting information of objects defined as patterns based on category to mimic the human nature is sometimes in short termed as pattern recognition. On the other hand, data analysis involves the process of systematically applying statistical and logical techniques to describe, compare, and summarize data based on narratives, charts, graphs, and tables as well as to discover the underlying principles and structures of the data.

Data and information must be gathered prior to applying the pattern recognition or data analysis task. A particular data object contains information represented in numerical or symbolic values known as features or attributes to distinguish from the other data points. Sometimes, a data object is also termed as a data vector in which each feature corresponds to a dimension or a particular direction. For example, a particular salmon could be represented based on its distinctive colour, texture, shape, smell, etc to distinguish it from other fish in a group. Therefore, as there are more features added, more information will be perceived in which corresponds to an increase of dimensions. At times, there could be thousands of dimensions as human attempts to capture more information from the data. However, enormous amount and high dimensionality of data would eventually lay a sheer task on the neural network classifiers and laborious work in data analysis even with the help of statistical software such as Excel and Minitab. Other tasks, for example clustering, are also computationally impractical.

Hence, there is a need to reduce the dimensions of data but remain most of the information of data vectors for mitigating problem in computational time. In short, it is also known as dimensionality reduction. Researchers have came up with many new dimensionality reduction techniques varying from different feature extraction techniques to multidimensional scaling with each having its own underlying mathematical principles, emphasized most on preserving the information of data vectors itself. However, high computational complexity within the dimensionality reduction technique itself again arise the problem in explosion of computational time. For example, feature extraction methods are very much dependent on the nature of data, and therefore generally not applicable, for instance, in all data mining tasks (Kaski, 1998). On the other hand, multidimensional scaling methods such as Principal Component Analysis (PCA) with time complexity of $O\left(nN^2\right)+O\left(N^3\right)$ for estimating the principal components i.e., the new directions of data vectors are computationally costly and almost infeasible if the dimension of the original data vectors is very high. Here, *N* represents the original dimension and *n* is the number of data vectors.

A new rapid dimensionality reduction method is urgently needed for situations where it is impossible to use the original vectors and the existing dimensionality reduction methods are too costly. The Random Projection (RP) method that only involves simple mathematics and low time complexity of $O(Nd)$ in forming the random matrix for obtaining the new directions of vectors is found as a computationally feasible method for reducing the dimensionality of the data so that mutual similarities between the data vectors are approximately preserved. Here, *N* and *d* are the dimensionalities before and after RP. It is clear that the computational complexity for RP is almost negligible as compared to the time complexity of PCA.

It would be impractical and meaningless for merely comparing the theoretical basis and mathematical principals of the dimensionality reduction techniques. Hence, there is a need to gauge the performance of techniques such as performing a classification task, etc prior applying the method for reducing dimensions of real world data. Therefore, in this project, the effectiveness of dimensionality reduction techniques is to be based on the performance of neural network classifiers as high dimensionality is a common major problem arises in the neural nets. Besides that, neural network classifiers are a common computational paradigm found in most practical applications such as medical diagnosis and industrial work for classification and generalization purposes as it imitates the biological nervous system of human nature. Furthermore, the availability of neural network classifiers in databases which could be easily downloaded from the web resources also contributes a reason for applying this technique in the project.

In this project, two neural network models have been experimented, i.e. Standard Backpropagation (Rumelhart et al, 1986) and Fuzzy ARTMAP (Carpenter et al., 1992) networks. The performance indicators of the neural network classifiers for justifying the performance of dimensionality reduction techniques could be basing on the percentage of correct classification, collective entropy, nearest neighbors, data compactness, purity, etc. A standard mathematical test is needed to clarify and evaluate the obtained experimental results as well as to be convincing enough in making decisions and giving strong conclusions on the achieved results besides eliminating controversies subjected to different perspectives of individuals. Hence, all decisions in this project are based upon the decision on the 95% confidence level of the Student's two-tailed paired *t*-test for the differences of mean classification accuracies, differences of mean collective entropies, and differences of mean purities by PCA and RP.

1.1 Project Objectives

Section 1.0 discussed the need for dimensionality reduction especially when data vectors are high and problem having long hours of computational time when a dimensionality reduction technique has high time complexity is applied. Besides that, there is a need to gauge the performance of the techniques based on the neural network classifiers. Therefore, the main objectives of this project are

- to develop a program for implementing RP as a dimensionality reduction technique
- to investigate the effectiveness of PCA and RP in dimensionality reduction
- to evaluate the performance of neural network-based classifiers with PCA and RP as the dimensionality reduction techniques.
- to compare the results from PCA and RP using Student's *t*-test
- to quantify the stability of the classification results using bootstrapping

1.2 Scope

This project mainly includes the following activities:

- understanding of theoretical basis and mathematical concepts applied in PCA that were used to form orthogonal principal components which results in lossless information if the original dimension is kept,
- understanding of the principal and mathematical concepts of RP used in formation of new directions, which results in orthogonal directions when the dimension of data is high based on Fisher's theorem and the findings of Hetch Nielsen (Hetch-Nielsen, 1994),
- understanding the behavior of neural network classifiers,
- conducting performance tests using neural network classifiers to measure the effectiveness of PCA and RP when data of benchmark and real medical data sets are mapped to new directions. The performance indicators are the percentage of correct classification, collective entropy, and purity, and
- performing statistical test to obtain the decision of Student's two-tailed paired *t* test on the performance results of PCA and RP.

1.3 Dissertation Outlines

This section steps through a series of chapters with brief descriptions on the related topics organized in the dissertation.

Chapter 2 begins with a review on various types of dimensionality reduction techniques. This is followed by a thorough survey on the applications of PCA and RP techniques to process real data. Lastly, this chapter reviews on the applications of Backpropagation and Fuzzy ARTMAP neural network classifiers to solve pattern recognition tasks.

Chapter 3 begins with the theoretical insights and the underlying mathematical principles of PCA and followed by RP. Subsequently, the principles of Backpropagation for classification are included. Lastly, the operations of Fuzzy ARTMAP network are explained.

Chapter 4 describes the experimental setup for dimensionality reduction and classification. This chapter begins with the explanation on the data sets used. After that, a brief description on the methods applied to dimensionality reduction by PCA and RP are included. An introduction to the tools used for performing classification is also given. The steps for data preprocessing prior to performing classification are also explained. Lastly, the parameter settings for Backpropagation and Fuzzy ARTMAP networks to perform classification are given.

Chapter 5 gives an introduction on the performance indicators used in the experiment. Detailed analyses and discussion of the results are presented. The principles of the performance indicators, i.e., the percentage of correct classification, collective entropy, and purity are explained. After that, the mathematical principles on bootstrapping for estimating the confidence intervals of the results as well as Student's *t*-test for comparing the experimental results are given. The main focus of this chapter proceeds is the detailed analyses, comparison, and discussion of the experimental results obtained from Backpropagation and Fuzzy ARTMAP networks using PCA and RP.

Chapter 6 gives a conclusion that wraps up the results and discussions. Recommendations for future work are also included.

CHAPTER 2

LITERATURE REVIEW

2.0 Introduction

This chapter presents surveys on dimensionality reduction techniques and classification by neural network classifiers. It is worthwhile to make a review on the dimensionality reduction techniques and the practical applications to solve real world data. This is particularly important to perceive a better understanding and see an overview of techniques used in past researches to reduce dimensionality of data in various aspects, get an idea on general obstacles encountered during dimensionality reduction as well as the need for future improvements.

Section 2.1 begins with an introduction on dimensionality reduction techniques and a list of contributions by PCA and RP. Section 2.2 shows a list of surveys on the applications of Backpropagation and Fuzzy ARTMAP networks. Perceiving knowledge related to the neural network classifiers available from web resources will lead to better understanding in selecting the proper classifiers for practical applications.

2.1 A Survey on Dimensionality Reduction Techniques.

Among the main concerns in dimensionality reduction is selecting the appropriate method which preferably subject to less information loss and lower computational complexity. Survey papers by (Fodor, 2002) revised on several methods and categorized into linear and non-linear techniques. Fodor discussed the mathematical principal underlying in each technique such as PCA, RP, Independent Component Analysis (ICA), Multidimensional Scaling, Projection Pursuit, Vector Quantization, Factor Analysis, etc as well as the drawbacks subjected to the technique applied. Generally, even though most of the techniques as described above such as PCA, Principal Curves, Projection Pursuit, etc may fulfill the intended purpose of preserving information in data, these techniques have imposed adverse effect on the computational time needed for reduction when dimensions of data are high. This is subjected to the constraint of high computational complexity within the technique itself. RP on the other hand only involves a simple mathematics.

Section 2.1.1 shows a list of surveys on the application of PCA as a dimensionality reduction technique. Section 2.1.2 describes the utilization of RP as a new dimensionality reduction technique.

2.1.1 Principal Component Analysis (PCA)

PCA or known as the Karhoneun Leunean Theorem, has been widely used by researchers of various field for the purpose of dimensionality reduction. PCA seeks application in areas of research such as medical imaging (Mcleish et al., 2002), computer vision (Menser & Muller, 1999; Oravec & Paradicova, 2004), geophysics (Gothoh et al., 2000), biomedical (Ravazzani et al., 2003; Tarveinan, 2001), process control (Kumar et al., 2002; Yang et al., 2002), and neural computing (Kasgoftaar & Szabo, 1994; Opitz, 1997).

PCA has grown the interest of several researchers in image processing and machine vision. Menser and Muller (1999) performed PCA to enhance face detection in colour images. Result was along with their intended purpose as colour information improved the robustness of colour detection significantly. Besides that, Mcleish et al. (2002) studied the motion and deformation of the heart due to respiration based on patient’s data. PCA was used to produce a statistical model of motion and deformation of the heart. Detailed information regarding how the model could be used to assist motion correction was also described. Recently, Oravec and Pavlovicova (2004) tried to combine PCA and Multilayer Perceptron Neural Network (MLP) for face recognition. The proposed face recognition system yielded good results as compared to the other seven presented methods with more than 80% of correctly recognized faces.

Tarveinan et al. (2001) investigated the goodness of PCA for analyzing the pattern of successive galvanic skin responses (GSR) from twenty health control and thirteen psychotic patients. With application to clustering, a significant discrimination with overall correct ratings of 82% of patients was achieved. A significant fact was that all patients were ranked correctly, giving the proposed method a sensitivity of 100%. There are also applications of PCA in biomedical engineering. For example, Ravazzani et al. (2003) explored the effectiveness of PCA to facilitate fast detection of Transient-Evoked Otoacoustic Emissions (TEOAE i.e., acoustic signals coming out form inner ear after acoustic simulation click). The results seemed to enhance the signal to noise ratio and in turn, allowed correct detection in response.

Besides that, the applications of PCA in neural networks are also explained in this section. Khosgoftaar and Szabo (1994) explored the effectiveness of PCA to neural network modeling as a way of improving the predictive quality of neural networks quality models. Results showed an improvement in prediction by utilizing PCA. Opitz (1997) used PCA to detect the functional redundancy of a neural network. Results revealed that the new algorithm give much more accurate estimation on network complexity than the standard approaches such as weight decay, weight pruning, prediction-risk techniques, etc.

Process control is also another field which needs the applications of PCA. Yang (2002) utilized multiple PCA models based on soft-partition algorithms to monitor continuous processes with more than one stage of operation. Good results were demonstrated on a three-tank plant. Kumar et al. (2002) proposed PCA-based monitoring scheme to detect process changes. This technique proved to be more sensitive to faults which are not detectable in the previous Q-statistics technique.

Other applications of PCA are estimating geomagnetic data .to detect earthquakes in IZU islands (Gothoh et al., 2000) and setting the threshold in a dynamic Idd test process by identifying process corners and computing statistical model (Jiang &Vinnakota, 2002).

2.1.2 Random Projection (RP)

RP is a simple, new yet powerful dimensionality reduction technique that uses random projection matrices to project data onto low-dimensional spaces. This technique has attracted the interest of researchers of various fields. Although it is based on simple mathematical ideas, random mapping has demonstrated good performance in a number of applications including information retrieval (Kleinberg, 1997; Thaper et. al., 2002) machine learning, (Kaski, 1997; Dasgupta, 2000; Brigham & Maninila, 2001; Fradkin & Madigan, 2003; Fern & Brodley, 2003), and optimization (Vempala, 1998), yielding results comparable to the conventional dimensionality reduction techniques, such as PCA, while reducing the computational requirements.

RP has grown the interest of several researchers in database applications, especially in the nearest-neighbor similarity search problem. Indyk and Motwanti (1998) applied RP in the form of locality-sensitive hashing, as part of a randomized algorithm for solving the nearest-neighbor problem in high dimensional Euclidean distance. They used random projection to reduce the original problem to a series of tractable, low-dimensional problems and produced an algorithm that scales better with dimension that determines methods. It is implemented successfully in practice for sparse indexing of databases. Kleinberg (1997) combined randomly chosen one-dimensional projections of the underlying data to develop an algorithm for finding approximate nearest neighbors.

Papadimitriou et. al. (1998) combined RP with Latent Semantic Indexing (LSI). Even though LSI is an elegant and accurate technique for document categorization and classification, its high computational cost make it impossible for large databases as it slows down the process in LSI tremendously. Prior to applying LSI, RP was used to reduce the dimensionality of the data. Kurimo (1999) applied RP for the similar purpose as Papadimitriou to the indexing of audio documents, prior to using LSI and Self-Organizing Maps (SOMs).

Dasgupta (2000) combined RP with Expectation-Maximization (EM) in learning high-dimensional Gaussian mixture models. His results illustrate that data separation can be retained even though from a mixture of *k* Gaussians and projected down to

O(*logk*) dimensions. The results were encouraging when the algorithm was applied on a hand-written digit data set. Motivated by the results of Dasgupta, Fern and Brodley (2003) investigated the application of RP for clustering high-dimensional data. They proposed using ensembles of RP as earlier work demonstrated that clustering results were very unstable when using single runs of RP. Better performance was achieved when three different data sets were used as compared to using individual runs of RP. Positive results were achieved in comparison to clustering using the combination of EM algorithm and PCA for dimensionality reduction.

Kaski (1997) presented the experimental results using RP in the context of using SOMs i.e., WEBSOM, a system for organizing textual documents. His results show that RP needed moderate number of dimensions for producing a good mapping. In this case, the results were as good as those obtained using PCA, and almost as good as those obtained using the original vectors. Brigham and Manilla (2001) compared several dimensionality reduction techniques such as PCA, RP, and DCT on image and text data. Their results again indicate that RP preserves distances and has performance comparable to that of PCA while being much faster.

More recently, Fradkin and Madigan (2003) evaluated RP in the context of supervised learning. In particular, RP was compared with PCA on a number of different problems using different machine learning algorithms which came to a conclusion that RP was slightly worst than PCA. However, its computational advantages might make it attractive in certain applications.

Other applications of RP include solving VLSI layout with minimum area consumption (Vempala, 1998), performing approximate kernel computations (Achlioptas et. al., 2001), similarity computations for histogram models (Thaper et. al., 2002), protein similarity search (Rigoutsos & Califano, 1993), and DNA motif discovery (Buhler & Tompa, 2002).

2.2 A Survey on Neural Network Classifiers

The large amount of neural network classifiers available on the web resources has always raised the curiosity of human in selecting the appropriate algorithm for better performance in classification. The performance is indeed dependant on the built in structure of the algorithm itself. Section 2.2.1 lists the applications of Backpropagation algorithm while section 2.2.2 describes the utilizations of FAM to perform pattern recognition tasks.

2.2.1 Backpropagation

This section will make a survey on the applications of Backpropagation, first of all because it is powerful, useful, and relatively easy to understand but also many other training methods can be seen as a modification from it. The training method is simple even for complex models having thousands of parameters (Duda et al., 2000). The applications of Backpropagation for classification in the following research areas are explained below.

- Brunet et al. (1994) applied Backpropagation for recognizing phonemes. The ability of the network was tested with samples from a few men and women. Although there was limitation for Backpropagation in recognizing similar phonemes, the network was able to function in an independent manner.
- Chen and Hwang (1994) developed a Multi-Level Backpropagation Network (MLBPN) for pattern recognition which was practically needed for massive computations to extract features. The MLBPN could improve the pattern recognition systems and keep good characteristics of Backpropagation.
- Funderbuck et al. (2000) investigated the feasibility of Backpropagation to predict the performance of chronic users of cocaine after one month of abstinent. Although results were not optimal, Backpropagation proved that it worked well and handled the noise that was invariably present in the data. This provided a first approximation of a clinical useful tool.
- Rodic et al. (2002) proposed a new concept for integrating Backpropagation as part of the active control system to ensure robustness and better adaptability upon the system uncertainties and inaccuracies. The system had shown to be

valid and effective. The fast convergence of learning process was achieved by Backpropagation

- Jin et al. (2002) applied Backpropagation to identify fingerprints and successfully proved that the network was able to recognize the core part of the fingerprint images.

Other applications of Backpropagation include prediction of protein structural class (Metfessel & Saurugger, 1993) and recognition of facial affect (Avent et al, 1994).

2.2.2 Fuzzy ARTMAP (FAM)

ART (Adaptive Resonance Theory) is found to be able to overcome the stability-plasticity dilemma (Carpenter & Grossberg, 1987a, 1988) suffered by most of the neural network and perform incremental learning. Fuzzy ARTMAP, a derivative of ART is able to perform classification on binary and analogue inputs (Carpenter et al., 1992), has been popularly used by researchers. The goodness of Fuzzy ARTMAP could be seen from a few application in the examples as below.

- Srinivasa and Ziggert (1994) applied Fuzzy ARTMAP to approximate the thermal error maps in machine tools. Even though FAM was not able to learn thermal errors in real-time, it could make correct predictions on test data.
- Murshed et al (1995) proposed the use of Fuzzy ARTMAP for offline signature verification and trained the system with genuine signatures. The system could be trained even though with only genuine signatures. The authors believed that FAM could produce a solution of unresolved and very difficult problem in area of signature verification. They proposed to evaluate the signatures on large databases in the following research.
- Jervis et al. (1996) evaluated the effectiveness of FAM to detect Contigent Negative Variation (CNV) as neural nets trained on CNV data offer an additional tool for diagnosis of Huttington Disease, Parkinson Disease, and Schizophrenia as well as detecting and monitoring the pre-onset Huttington Disease. FAM had showed most promising in this domain.
- Ham and Han (1996) performed classification of cardiac arrhythmias using FAM. Classification with 99% specificity and 97% sensitivity were achieved.

- Dagher et al. (2002) attempted fingerprint classification using FAM. Results showed that the system was able to achieve high acceptable identication accuracies and similar to the performance of the current implemented matchers.

Other applications of FAM are speech recognition (Woo et al., 2000), discovery of gene function and classes of cancer (Azuaje, F., 2001), and fault detection and diagnosis in power generation plant with symbolic rule extraction (Tan &Lim, 2004).

2.3 Summary

From the literature review, an overall view of the entire project to be performed is given. It is clearly shown that PCA has been a popular technique which seeks application in many aspects of researches. Although RP has been recently introduced, this dimensionality reduction technique is popularly used by researchers in machine learning, information retrieval, and optimization fields. Fuzzy ARTMAP and Backpropagation are used in a variety of classification tasks, e.g., speech, text, or image data despite there may be some limitations in the classifier itself.

CHAPTER 3

SYSTEM'S GENERAL ARCHITECTURE

3.0 Introduction

Prior to proceeding to the detailed experimentation, it is worthwhile to understand the theoretical insights of PCA and RP as well as Backpropagation and Fuzzy ARTMAP used in this project. The system proposed for comparing the performance of PCA and RP is composed of two main stages i.e., the first stage for dimensionality reduction as described in section 3.1 while the second stage involves classification by Backprogation and Fuzzy ARTMAP networks as in section 3.2. The general architecture of the system is shown in Figure 3.1.

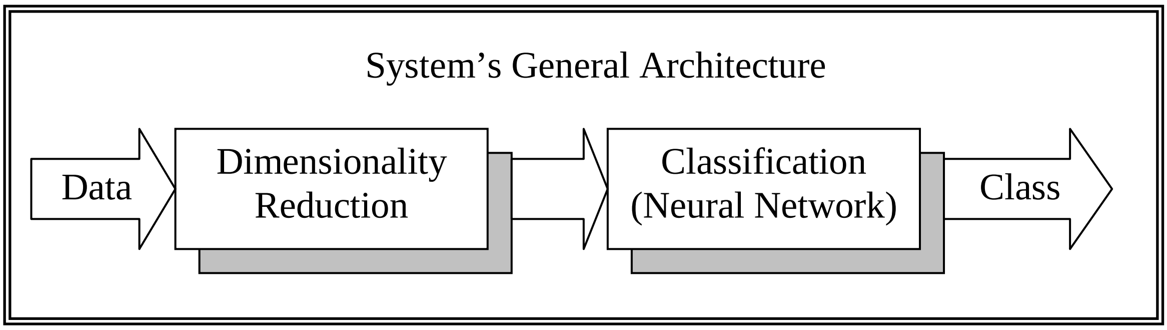
System's General Architecture
Data
Dimensionality Reduction
Classification (Neural Network)
Class

Figure 3.1: General Architecture for Measuring Performance of PCA and RP.

3.1 Dimensionality Reduction Techniques

Dimensionality reduction involves the process of reducing dimensions from data vectors by mapping to a lower dimensional subspace after going through certain functions that may capture important information from the original data vectors. Note that the dimensionality reduction techniques described in this dissertation use the same data representation model i.e., the vector space model in which each sample is represented as a vector. Each dimension of the vector corresponds to one feature, and the value of each component is the relative frequency of occurrence or measurement for the corresponding feature in the sample.

3.1.1 Principal Component Analysis (PCA)

Principal Component Analysis is an unsupervised approach to finding the right "features" from the data (Duda et al, 2000). PCA transforms correlated or uncorrelated features to independent principal components which has the maximum variability on the first principal components, followed by the second principal component and so on. This dimensionality reduction technique is widely used as described in section 2.1.1. The basic theory of PCA is as follows.

3.1.1.1 Theoretical Basis of PCA

Suppose that we want to represent all the vectors in a set of n *d*-dimensional samples, **X** $\vec{x}_1,.......\vec{x}_n$ by a single vector $\vec{x}_0$. However, there are many vectors that could be represented by $\vec{x}_0$. Therefore, the vector $\vec{x}_0$ is chosen as such that the sum of squared distances between $\vec{x}_0$ and the various $\vec{x}_k$ is as small as possible. The sum of squared-error criterion function, $J_0\left(\vec{x}_0\right)$ by

$$J_0\left(\vec{x}_0\right) = \sum_{k=1}^{n} \left\| \vec{x}_0 - \vec{x}_k \right\|^2 \tag{3.1}$$

and find the value of $\vec{x}_0$ that minimizes J_0. It is simple to know that the solution to this problem is given by $\vec{x}_0 = \vec{m}$, where $\vec{m}$ is the sample mean,

$$\vec{m} = \frac{1}{n}\sum_{k=1}^{n} \vec{x}_k \tag{3.2}$$

Therefore, the solution to this problem could be easily shown by writing

$$J_0\left(\vec{x}_0\right) = \sum_{k=1}^{n} \left\| \left(\vec{x}_0 - \vec{m}\right) - \left(\vec{x}_k - \vec{m}\right) \right\|^2$$

$$= \sum_{k=1}^{n} \left\| \vec{x}_0 - \vec{m} \right\|^2 - 2\sum_{k=1}^{n} \left(\vec{x}_0 - \vec{m}\right)^t \left(\vec{x}_k - \vec{m}\right) + \sum_{k=1}^{n} \left\| \vec{x}_k - \vec{m} \right\|^2$$

$$= \sum_{k=1}^{n} \left\| \vec{x}_0 - \vec{m} \right\|^2 - 2\left(\vec{x}_0 - \vec{m} \right)^t \sum_{k=1}^{n} \left(\vec{x}_k - \vec{m} \right) + \sum_{k=1}^{n} \left\| \vec{x}_k - \vec{m} \right\|^2$$

$$= \sum_{k=1}^{n} \left\| \vec{x}_0 - \vec{m} \right\|^2 + \sum_{k=1}^{n} \left\| \vec{x}_k - \vec{m} \right\|^2 \qquad (3.3)$$

The second sum is independent of $\vec{x}_0$ and the minimum value of J_0 could be achieved when $\vec{x}_0 = \vec{m}$. The sample mean is a zero dimensional representation of data set. Although this method of representation is simple, it does not reveal any variability of the data. Therefore, the one-dimensional data representation could be obtained by projecting the data onto a line running through the sample mean. Assuming $\vec{e}$ be a unit vector in the direction of the line, then the equation for $\vec{x}_0$ could be written as,

$$\vec{x} = \vec{m} + a\,\vec{e} \qquad (3.4)$$

where a denotes the distance of any point $\vec{x}$ from the mean $\vec{m}$.

If $\vec{x}_k$ is represented by $\vec{x} = \vec{m} + a_k\,\vec{e}$, a set of “optimal” coefficients, a_k are found by minimizing the squared-error criterion function.

$$J_1\left(a_1, \ldots\ldots, a_n, \vec{e} \right) = \sum_{k=1}^{n} \left\| \left(\vec{m} + a_k\,\vec{e} \right) - x_k \right\|^2$$

$$= \sum_{k=1}^{n} a_k^2 \left\| \vec{e} \right\|^2 - 2\sum_{k=1}^{n} a_k\,\vec{e}\left(\vec{x_k - m} \right) + \sum_{k=1}^{n} \left\| \vec{x}_k - \vec{m} \right\|^2 \qquad (3.5)$$

Since $\left\| \vec{e} \right\|$ is equivalent to 1, partially differentiating with respect to a_k,

$$\frac{\partial J_1\left(a_1, \ldots\ldots a_n, \vec{e} \right)}{\partial a_k} = 2\sum_{k=1}^{n} a_k - 2\sum_{k=1}^{n} \vec{e}\left(\vec{x}_k - \vec{m} \right)$$

$$= 2\sum_{k=1}^{n} \left(a_k - \vec{e}\left(\vec{x}_k - \vec{m} \right) \right)$$

Setting the derivative to zero,

$$2\sum_{k=1}^{n}\left(a_k - \vec{e}\left(\vec{x}_k - \vec{m}\right)\right) = 0$$

$$a_k = \vec{e}\left(\vec{x}_k - m\right) \tag{3.6}$$

is obtained. This results shows that a least-squares solution is obtained by projecting the vector $\vec{x}_k$ onto a line in the direction of $\vec{e}$ that passes through the sample mean.

(a) Scatter Matrix

This part demonstrates the finding of best direction $\vec{e}$ for the line involving the computation of scatter matrix as defined by

$$s = \sum_{k=1}^{n}\left(\vec{x}_k - \vec{m}\right)\left(\vec{x}_k - \vec{m}\right)^t \tag{3.7}$$

The scatter matrix is only *n* – 1 times the sample covariance matrix. It arises here when we substitute a_k found in equation (3.6) into equation (3.5) to obtain

$$\begin{aligned} J_1\left(\vec{e}\right) &= \sum_{k=1}^{n} a_k^2 - 2\sum_{k=1}^{n} a_k^2 + \sum_{k=1}^{n}\left\|\vec{x}_k - \vec{m}\right\|^2 \\ &= -\sum_{k=1}^{n}\left[\vec{e}\left(\vec{x}_k - \vec{m}\right)\right]^2 + \sum_{k=1}^{n}\left\|\vec{x}_k - \vec{m}\right\|^2 \\ &= -\sum_{k=1}^{n}\vec{e}\left(\vec{x}_k - \vec{m}\right)\left(\vec{x}_k - \vec{m}\right)^t\vec{e} + \sum_{k=1}^{n}\left\|\vec{x}_k - \vec{m}\right\|^2 \\ &= -\vec{e}^t S\,\vec{e} + \sum_{k=1}^{n}\left\|\vec{x}_k - \vec{m}\right\|^2 \end{aligned} \tag{3.8}$$

This clearly shows that $\vec{e}$ minimizes J_1 and also maximizes $\vec{e}^t S\,\vec{e}$. Using Langrage multipliers to maximizes $\vec{e}^t S\,\vec{e}$ subject to the constraint that $\left\|\vec{e}\right\| = 1$ and assuming λ as the undetermined multiplier and differentiate

$$u = \vec{e}^t S\,\vec{e} - \lambda\,\vec{e}^t\,\vec{e} \tag{3.9}$$

with respect to $\vec{e}$ to obtain

$$\frac{\partial u}{\partial \vec{e}} = 2S\vec{e} - 2\lambda\vec{e} \tag{3.10}$$

Setting the gradient vector to zero, it is shown that $\vec{e}$ must be an eigenvector of the scatter matrix:

$$S\vec{e} = \lambda\vec{e} \tag{3.11}$$

In particular because $\vec{e}^{t} S \vec{e} = \lambda \vec{e}^{t} \vec{e} = \lambda$, it follows that to maximize $\vec{e}^{t} S \vec{e}$, the eigenvector with the corresponding largest eigenvalue of the scatter matrix have to be selected with the best in least sum-of-squared error sense. The data are projected onto the line or eigenvector that passes through the sample mean. This results can be readily extended from a one-dimensional dimensional projection to a *d'* dimensional projection. In replace of equation (3.4),

$$\vec{x} = \vec{m} + \sum_{i=1}^{d} a_i \vec{e} \tag{3.12}$$

where $d^{'} \leq d$ '

This leads to the following criterion function,

$$J_{d^{'}} = \sum_{k=1}^{n} \left\| \left(\vec{m} + \sum_{i=1}^{d^{'}} a_{ki} \vec{e} \right) - \vec{x}_k \right\|^2 \tag{3.13}$$

is minimized when the vectors $\vec{e}_1, \ldots\ldots. \vec{e}_{d^{'}}$ are the *d'* eigenvectors of the scatter matrix having the largets eigenvalues. The eigenvectors, $\vec{e}$ are orthogonal and nonzero vectors while λ is real positive as the scatter matrix is real and symmetric. They form the principal components representing any feature vector $\vec{x}$. The coefficients a_i in (3.12) are the components of $\vec{x}$ in the new directions and known as the principal components.

Therefore, it is mathematically proven that PCA reduces the dimensionality of feature space by restricting attentions to those directions along which the scatter of data points are the greatest.

3.1.1.2 Algorithm for PCA

In summary, PCA could be performed as follows.

1 Collect $\vec{x}_k$ of a d dimensional data set, **X**, k = 1, 2, 3, …n.

2. Calculate the mean, $\vec{m}$ as shown in equation (3.2) for data adjust to mean zero

$$\mathbf{X}_{\text{adjust}} = \sum_{k=1}^{n}\left(\vec{x}_k - \vec{m}\right) \qquad (3.14)$$

3. Calculate the variance-covariance matrix, ***C***. However, sometimes the correlation matrix, ***R*** is used instead when the units between 2 features. The computation method is shown in equation (3.7) with only $\frac{1}{d-1}$ times the scatter matrix.

4. Determine the eigenvalues, λ and eigenvectors, $\vec{e}$ of the matrix, **C.** To find a nonzero $\vec{e}$, the characteristic equation $|C - \lambda I| = 0$ must be solved. The eigenvalues and eigenvectors could be solved using

- Singular Value Decomposition, (SVD) because of its numerical stability,

- Transform the matrix to a tridiagonal form (using Householder transformation) and decompose it to ***C = QR***, where ***Q*** is orthogonal and ***R*** is upper triangular, and
- Hotelling's power method, an iterative technique to find the largest $d_i \le d$ eigenvectors and eigenvalues, result that step 5 is unnessary.

5. Sort the eigenvalues λ,(and corresponding eigenvectors) so that $\lambda_1 \ge \lambda_2 \ge ....\lambda_d$.

6. Select the first $d_i \leq d$ eigenvectors and generate the data set in the new (usually compressed) representation. The first principal component consists of $\vec{e}_1^{\,t} \cdot \mathbf{X}_{\text{adjust}}$, the second principal component consists of $\vec{e}_2^{\,t} \cdot \mathbf{X}_{\text{adjust}}$, and so on.

3.1.2 Random Projection (RP)

RP has recently emerged as a powerful dimensionality reduction tool in the demand for high dimensional data. This section will detail the theoretical basis of RP with the underlying mathematical principles.

3.1.2.1 Theoretical basis of RP

(a) Random Projection Technique (RP)

In the linear RP method, the original data vector for a sample, denoted by $\vec{n} \in R^N$, is multiplied by a random matrix **R.** The mapping

$$\vec{x} = \mathbf{R}\,\vec{n} \tag{3.15}$$

results in a reduced dimensional vector $x \in R^d$ (Kaski, 1998). The random matrix consists of random values and the Euclidean length of each column is normalized to unity. If the *i*th column of **R** is denoted by $\vec{r}_i$ and the *i*th component of $\vec{n}$ is denoted by n_i, the random mapping operation can be expressed as

$$\vec{x} = \sum_i n_i \, \vec{r}_i \tag{3.16}$$

In the original vector, the components n_i are weights of orthogonal unit vectors, whereas in expression (3.16), each dimension *i*, of the original data space has been replaced by a random, non-orthogonal direction $\vec{r}_i$ in the reduced dimensional space.

(b) Properties of the RP

"*There exist a much larger of almost orthogonal than orthogonal directions in a high dimensional space. Therefore, in a high dimensional space, even vectors having random directions might be sufficiently close to orthogonal to provide an approximation of a basis*" (Hetch-Nielsen, 1994).

(c) Transformation of the Similarities.

The closer the vectors are to being orthogonal, the better the similarities of the vectors obtained by random mapping correspond to the original similarities. The similarities of the vectors are measured using the cosine function, $A.B = |A||B|\cos\theta$. Each column of the random matrix is normalized to unit length so that the cosine can be computed as the inner product of the vectors.

Assuming two vectors obtained by random mapping, $\vec{y}_1 = \mathbf{R}\vec{n}$ and $\vec{y}_2 = \mathbf{R}\vec{m}$. The inner product of two vectors, $\vec{y}_1$ and $\vec{y}_2$ that have been obtained by the random mapping of the vectors n and m, respectively, can be expressed as follows:

$$\vec{y}{}_1^t \vec{y}_2 = \vec{n}{}^t \mathbf{R^t}\,\mathbf{R}\vec{m} \tag{3.17}$$

$$\mathbf{R^t}\,\mathbf{R} = 1 + \varepsilon \tag{3.18}$$

$$\varepsilon_{i\,j} = \vec{r}{}_i^t \vec{r}_j \text{ for } i \neq j, \text{ and } \varepsilon_{i\,i} = 0 \text{ for all } i. \tag{3.19}$$

The diagonal entries in $\mathbf{R^t}\,\mathbf{R}$ are collected into the identity matrix, *I*. These entries are always equal to unity since the column vectors, $\vec{r}_i$ has been normalized. The units off the diagonal has been collected into ε which consists of the cosine angles between the vectors r_i and r_j. Therefore if all of the entries in ε are zero, the vectors $\vec{r}_i$ and $\vec{r}_j$ are orthogonal as cosine 90° = 0. Hence, $\mathbf{R^t}\ \mathbf{R} = \mathbf{I}$ and the similarity of the vectors are preserved in RP. However in practice, the entries in ε are just approximately zero.

(d) Statistical Properties of ε

It is possible to analyze the properties of ε if the distribution of the entries into the random matrix, **R** is fixed. In the present experiment, the components $\vec{r_i}$ and $\vec{r_j}$ are initially generated by a normal random number generator to be independent, identical, and normally distributed with mean zero, and thereafter the length of $\vec{r_i}$ and $\vec{r_j}$ is normalized to unity. Hence, the direction of $\vec{r_i}$ and $\vec{r_j}$ will be distributed uniformly. It is clear that $E\lfloor \varepsilon_{ij} \rfloor = 0$ for all i and j, where *E* denotes the average over all random choices for the entries of ***R***. Besides that, the distribution of $\varepsilon_{\mathrm{ij}}$ could be derived. $\varepsilon_{\mathrm{ij}}$ is in fact an estimate of correlation coefficient between two independent, identical and normally distributed random variables. This can be proven that for the units off the diagonal entries that

$$\varepsilon_{ij} = \frac{\sum_d (r_i)\cdot(r_j)}{\sqrt{\sum_d r_i^2}\sqrt{\sum_d r_j^2}} \tag{3.20}$$

$$\approx \frac{\sum_d \left(r_i - \overline{r_i}\right)\left(r_j - \overline{r_j}\right)}{\sqrt{\sum_d \left(r_i - \overline{r_i}\right)^2}\sqrt{\sum_d \left(r_j - \overline{r_j}\right)^2}} \tag{3.21}$$

where $\overline{r_i}$ and $\overline{r_j}$ are approximately zero.

Equation (3.20) is the exact formula for correlation coefficient.

(e) Fisher's Theorem

As proven by Fisher,

$$\frac{1}{2}\ln\left(\frac{1+\varepsilon_{i\,j}}{1-\varepsilon_{i\,j}}\right) \sim N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{d-3}\right) \tag{3.22}$$

where ρ denotes the population correlation coefficient of two random variables

d denotes the number of samples in the estimate. In this context, *d* is the number of dimensions of the original vector.

Using series expansion, it can be shown that

$$\frac{1}{2}\ln\left(\frac{1+\varepsilon_{ij}}{1-\varepsilon_{ij}}\right)=\frac{1}{2}\times 2\left(\varepsilon_{ij}+\frac{\varepsilon_{ij}^{3}}{3!}+\frac{\varepsilon_{ij}^{5}}{5!}+\frac{\varepsilon_{ij}^{7}}{7!}+........\right) \tag{3.23}$$

However, if ε_{ij} is small enough that the higher order ε_{ij} could be ignored,

$$\frac{1}{2}\ln\left(\frac{1+\varepsilon_{ij}}{1-\varepsilon_{ij}}\right)\approx\varepsilon_{ij}\,. \tag{3.24}$$

The expected mean for ε_{ij} is denoted by $E\lfloor\varepsilon_{ij}\rfloor$ and since the random variables $\overset{\rightarrow}{r_i}$ and $\overset{\rightarrow}{r_j}$ are uncorrelated, $\rho=0$, therefore, $E\left[\varepsilon_{ij}\right]=\frac{1}{2}\ln\left(\frac{1+0}{1-0}\right)$ = 0.

If *d* is large enough,

$$\sigma_{\varepsilon}^{2}\approx\frac{1}{d}\,. \tag{3.25}$$

Therefore, since this equation is linearized around zero and if the claim follows for large *d*, the matrix $\mathbf{R}^{\mathbf{T}}\,\mathbf{R}$ will approximate the identity matrix the better because for higher dimensional vectors, most of the entries in ε approximates zero.

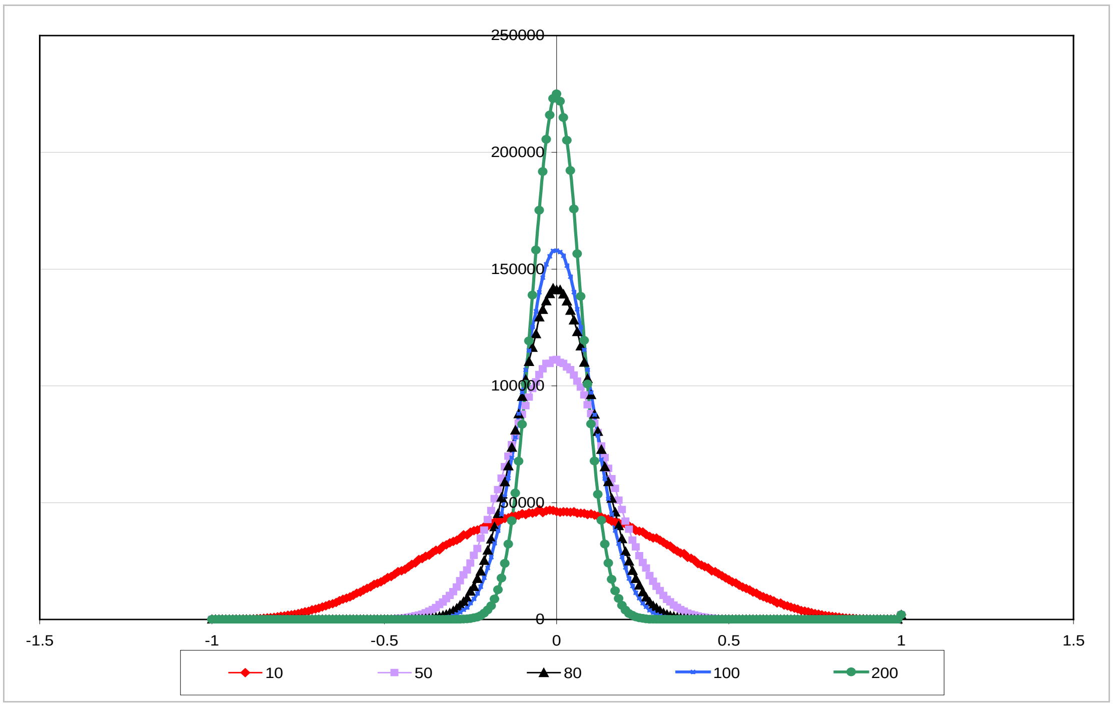

Figure 3.2: Distributions of ε_{ij} for different dimensionalities.

The normal distributions of the inner products between pairs of random vectors $\vec{r_i}$ for different dimensionalities with variance equal to $\frac{1}{d-3}$ are arbitrarily generated by a normal random number generator and plotted in Figure 3.2. It is noticed from the empirical curves that when d increases, the inner products become smaller and the vectors become more orthogonal. Generally, small inner products contribute only small distortions in the similarity computations.

(f) Statistical Properties of Mutual Similarities

It is possible to investigate more closely how the similarities of the original vectors are transformed in random mapping. Given a pair of original vectors, $\vec{n}$ and $\vec{m}$ it is possible to derive the distribution of the similarity of the vectors $\vec{x}$ and $\vec{y}$ obtained by random mapping.

Using equations (3.17), (3.18) and (3.19), the inner product between the mapped vectors $\vec{x}$ and $\vec{y}$ can be expressed as

$$\begin{aligned}\vec{x}^t \vec{y} &= (\mathbf{R}\vec{n})^t.(\mathbf{R}\vec{m}) \\ &= \vec{n}^t \mathbf{R}^\mathbf{T}\mathbf{R}\vec{m} \\ &= \vec{n}^t (I+\varepsilon)\vec{m} \\ &= \vec{n}^t \vec{m} + \vec{n}^t \varepsilon \vec{m} \\ &= \vec{n}^t \vec{m} + \sum_{k\neq l}\varepsilon_{kl} n_k m_l \end{aligned} \tag{3.26}$$

where

$$\delta = \sum_{k\neq l}\varepsilon_{kl} n_k m_l\,. \tag{3.27}$$

The mean of δ is zero as the mean of each term in the sum is zero and the random variables are uncorrelated.

$$\begin{aligned}\sigma_\delta^2 &= E\left[\left(\sum_{k\neq l}\varepsilon_{kl} n_k m_l\right)\left(\sum_{p\neq q}\varepsilon_{pq} n_p m_q\right)\right] \\ &= \sum_{k\neq l}\sum_{p\neq q} E\left[n_k m_l n_p m_q \varepsilon_{kl}\varepsilon_{pq}\right]\end{aligned}$$