**Supervised Dimensionality Reduction on Streaming Data:**

Introduction Idea: (My opinion: supervised vs unsupervised: feature extraction vs feature selection)

Dimensionality reduction is a very old problem but still this problem remains. We can reduce the dimensions on static data by applying Batch mode but in Streaming data due to time and memory constraint.by definition, streaming data are those which arrive continuously having High speed and huge volume. The main characteristic of streaming data is we can only see once for the purpose of real time processing. We can mining streaming data in different approaches for example through sliding window approach or the clustering approach.

Limitation of Traditional LDA:

Traditional LDA is a generalized eigenvalue problem with within-class and between-class scatter matrices. These two matrices are computed in batch mode, which implies that we have collected all data in advance.

In real world application, a complete training set might not be given beforehand. Traditional LDA algorithm may be useless because it requires all high dimensional data and all covariance matrices be kept in the memory and ipdated altogether.

Limitation of ILDA:

It is one-pass incremental LDA algorithm which can perform incremental dimensionality reduction but the limitation of this algorithm is it needs to solve a high dimensional generalized eigenvalue problem. As a consequence of that if dimensions are very high then memory becomes an issue.

Advantage & limitation of IDR/QR:

Here, the whole data matrix needs not to be saved in main memory. But the problem is here they consider appropriate matrices instead of the covariance matrix which leads the possibility of approximation when new data arrives.

**Incremental Linear Discriminant Analysis for Classification of Data Streams**

Difference between LDA & PCA:

LDA which is also known as FDA seek directions for efficient discrimination where PCA seeks direction for representation.

Common problem of LDA & PCA:

The both considers that training data set are available in advance and learning is carried out in one batch.

Approach for the real-world application:

We can collect data whenever new data are presented and do the batch learning over the collected data so far. The drawback of this approach is

We require a large memory to store the data and high computation expenses. The System also forget about the knowledge that it acquires in past because

It starts from the beginning when new data arrives.

Problems of IPCA [Hall et.all]:

Consider adding exactly one new sample to an Eigen space model at a time.

Challenge of ILDA:

The difference between IPCA and ILDA is the corresponding class data may not be present with the arrival data. The new class data may be presented after several learning stage because of randomness.

**Effective and Efficient Dimensionality Reduction for Large-scale and Streaming Data Preprocessing**

Introduction:

The necessity for dimensional reduction is mainly for many large-scale information processing problems. Due to World Wide Web growth, many

Traditional classification techniques require a big amount of memory and CPU usage. For an example, if we want to classify the documents we have to collect the documents, vector space transformations, dimensionality reduction, design the classifier system and finally evaluation. Among all these tasks dimensionality reduction have great importance because of the quality and efficiency of a classifier. The classification will give you a poor result caused by the high dimension of the feature space.

The traditional and state-of-the-art dimensionality reduction method can be classified into Feature Extraction and Feature Selection. Feature extraction algorithm aim to extract features by projecting high-dimensional space into lower dimensional space using algebraic expression. LDA, MMC

& PCA are Feature extraction algorithm.

Feature Selection is a greedy approach by aiming at finding out the subset of most representative features based on some criteria. Hence FS approaches are greedy so it’s a challenge to find the optimal solution. Orthogonal centroid algorithm is a FS algorithm.

Feature extraction algorithm always apply linear algebraic transformation to find the optimal solution of a problem. On the other side FE does not always compute the optimal solution rather it finds the optimal solution in discrete space. The tradeoff is FE is computationally faster than the

FE algorithm because Fs algorithm has no need to perform algebraic transformation. Due to it, Fs algorithm have much more implication than the

FE algorithm.

Difference among all IPCA Algorithm:

The main difference among IPCA is the incremental representation of the covariance matrix but IPCA ignores the valuable label information and not optimal for general class task.

ILDA:

Singularity problem of LDA and ILDA algorithm is instable.

Incremental Maximum Margin Criterion:

Parameter estimation for IMMC is to select the parameter estimation.

Incremental Orthogonal Centroid Algorithm:

Introduce IOC which is an online FE algorithm. It is a supervised FE algorithm. It utilizes orthogonal transformation on centroid.

Unsupervised approach:

PCA projects the original data to a new lower-dimensional data space according to some criteria without utilizing the label information.

Comparison between LDA and MMC:

LDA and MMC both are supervised FE algorithm. However computing MMC is easier than LDA because in LDA, we have matrix inverse operation.

Limitation of QR decomposition:

The time and space cost of QR decomposition are too expensive for large-scale data such as web documents.

**IDQ/QR: An Incremental Dimension Reduction Algorithm via QR Decomposition**

Introduction:

Dimension reduction is critically important for many database and data mining applications such as efficient storage and retrieval of High-dimensional data. There are many efforts have been made for speeding up Query processing. There have been many efforts but all their effectiveness decays because of the dimensions which is regarded as “curse of Dimensionality”. One standard approach is transforms the High Dimensional Data to Low Dimensions and then apply the indexing techniques.

LDA:

Linear Discriminant Analysis computes a linear transformation by maximizing the ratio of the between-class distance to the within-class distance so that it can achieve maximal discrimination. The common aspect of previously proposed LDA based algorithms is the use of Singular Value Decomposition. It is very much difficult to design an incremental solution for the eigenvalue problem on the product of scatter matrices.

Advantage of IDR/QR:

It applies QR decomposition than the SVD calculation. The advantage of using QR decomposition over SVD is in QR does not require the whole matrix in main memory. In addition, It also applies constrain on computational cost by applying efficient QR-updating techniques. The performance is same like the other LDA algorithm but in term of computational cost it is much better than others.

Solving LDA’s Singularity Problem:

There are many approaches taken place to solve the Singularity problem for example regularized LDA, PCA+LDA and LDA/GSVD. But all of these methods does not give a solution other than storing large datasets into the memory and requiring a high computation. Moreover, these approaches are not capable of handling chunk data.

Stage of the algorithm:

There are two stages for the algorithm. The first stage is to maximize the separability between different classes which is fulfilled by QR decomposition. The second stage incorporates in both between-class and within-class information by applying LDA on the “reduced” scatter matrices. The main advantage is it does not require the whole matrix in main memory.

Neural Network based LDA:

Iterative based LDA also proposed for neural network based LDA which require **O** (d2) time for each update where d denotes the dimension of the data.

\*\* Need to study more about the neural network stuff

**A Framework for Projected Clustering of High Dimensional Data Streams**

Introduction:

Data Stream problem has been studies extensively in recent years because of the great ease in collection of stream data. The characteristic of streaming data force to use algorithms one pass over the data. Now a days, many single-scan streaming analysis methods have been proposed but hence stream data is high dimensional in nature it is always difficult to generalize data streams.

The problem of data streams has gained importance in recent years because of hardware technology advancement. These advances have made it easy to store and record numerous transactions and activities in everyday life in an automated way.

**An Integrated algorithm of incremental and robust PCA**

Principle Component Analysis efficiently represent high dimensional vector with a small number of orthogonal basis vectors. The conventional methods of PCA always perform in batch mode which is computationally expensive when dealing with large scale problems. To address this there are many new algorithms developed which are generally similar in accuracy and speed while the difference are mainly on how to approximate covariance matrix. Traditional PCA is always susceptible to outlying measurements that is vulnerable to “outliers”.

**Candid Covariance-Free Incremental Principal Component Analysis**

Uniqueness of the paper:

The candid covariance-free incremental principal component (CCIPCA) used to compute the principal component of samples incrementally without estimating the covariance matrix. To achieve this, it keeps the scale of observations and computes the mean of observations incrementally which is an efficient estimate for some well-known distribution for example Gaussian distribution. This algorithm is developed based on a well-known statistical concept called efficient estimate. The method is for real-time applications and thus it does not allow iterations. To retain the old and new data they used amnesic average technique instead of fixed learning rate.

Problem of Batch Method:

The batch method no longer satisfies the incrementally derive from any kind of streaming data source. Online Development of streaming data requires the system performs when there is new data. Furthermore, if the dimension is high the computation and storage complexity grow dynamically and dramatically. For example a moderate gray image has 64 rows and 88 columns which results in a d-dimensional vector with d=5632 which means we need a covariance matrix of size d\*(d+1)/2 elements which amounts to 15,862,828 entries.

Solution of the problem:

Thus it is now necessary to have an incremental method to compute principal components for observations arriving sequentially. Here the principal component are updated based on the each observation vector. It is not necessary to have a covariance matrix.

Several IPCA techniques have been proposed to compute principal components without the covariance matrix [see the references] but they run into convergence problems when facing high dimensional vectors.

**Reference by the paper:**

**1**. E. Oja, Subspace Methods of Pattern Recognition. Letchworth, U.K.: Research Studies Press, 1983.

**2.** E. Oja and J. Karhunen, “On Stochastic Approximation of the Eigenvectors

and Eigenvalues of the Expectation of a Random Matrix,” J. Math. Analysis

and Application, vol. 106, pp. 69-84, 1985.

**3.** T.D. Sanger, “Optimal Unsupervised Learning in a Single-Layer Linear

Feedforward Neural Network,” IEEE Trans. Neural Networks, vol. 2, pp. 459-

473, 1989.

**Efficient Visualization of Large-scale Data Tables through Reordering and Entropy Minimization**

Introduction:

Entropy Minimization is a theoretic approach which order the table such that the similar rows and similar column are grouped together. For ordering of rows, EM-ordering repeats until convergence the steps of

1. Rescaling columns
2. Solving a travelling Salesman Problem (TSP) where rows are considered as cities

The time complexity of TSP is **O** (n\*log\*(n))

What this paper is about:

Data Visualization has a long history in scientific research and useful to understand the data more deeply. But due to the huge number of the data along with the high dimension, the state on art visualization tools failed to gain the deeper insight of the data for example histograms, scatter plots, pie and bar charts. In some cases calculation become too cumbersome and computationally costly for example parallel coordinates.

The more advanced visualization approach is low dimensional data projection approach i.e. data are visualized into lower dimensional subspace and then visualized using scatter plot. The common algorithms are PCA, LDA and many more. The tradeoff for this method is the possibility of significant amount of information loss.

There are many possible approached for ordering of data tables. As in PCA, the main task is to find the principle component and then order the examples by traversing the line or the curve. But ordering is not the main task of the PCA rather it is a byproduct of manifold search. In hierarchical clustering the same ordering can be done by traversing throughout the leaves of the binary tree. Another way of ordering can be done through spectral clustering but it gives suboptimal results when the data consists of several clusters that are not well separated.

\*Comparison with the heatmap.

**Reference by the paper:**

Parallel coordinates:

1. A. Inselberg, “The plane with parallel coordinates,” The Visual Computer, vol. 1, no. 2, pp. 69–91, 1985.

2. A. Inselberg and B. Dimsdale, “Parallel Coordinates,” in Human-

Machine Interactive Systems. Springer, 1991, pp. 199–233

**Summary of the Notes:**

1. Not included the visualization introduction
2. Include PCA,MMC, LDA and their limitation
3. Not include the time series Bitmap