Thesis Proposal

Khan Hafizur Rahman

August 29, 2017

1 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 all necessary documents. Then we need to apply a lot of mathematical techniques for example vector space transformations. After transforming applying method to reduce dimension, then we are ready for the design of the classifier system and finally now its time for evaluating the system [Yan+06]. Among all, dimensionality reduction have a great importance because the quality and efficiency of the system depends on it. The classification will give you a poor result caused by the high dimension of the feature space.

Dimension Reduction is not only important for classification of data but also critically important for storing data efficiently[Ye+05]. Dimensionality reduction also help to speed up queering procedure. If Data is dimensionally reduced then applying indexing structure help to speed up the queering procedure.

Though Dimensionality reduction is a very old problem but still this problem exists. We are somehow able to reduce dimension on static data but in case of streaming data still we are looking for a better solution. Streaming data are referred to those data which arrive continuously with a high-speed and huge volume[YLO07].The important characteristic of streaming data is we can only have the data for once to analysis which ensures real time processing. It means in streaming data time and memory is the important thing to consider.

Stream Data and challenges related to it has been studies extensively in recent years because of the availability of data. There are many sources from where data evolves continuously for example, clicking web pages, data in stock market and data evolves from many sensors. Due to the characteristics of streaming data we are force to use one pass algorithm over the data.This is the main difference in approach between static data and dynamic data. In static data we can iterate over the data as many times as we want i.e. batch mode but in

streaming data we can't. To the main obstruction of streaming data is the high dimensionality which adds more challenge for the data analysis in real time.

2 Background

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. Linear Discriminant Analysis (LDA), Maximum Margin Classification (MMC) & Principle Component Analysis (PCA) are Feature extraction algorithm[Yan+06].

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.

To adjust traditional algorithm with the Stream data researcher developed new algorithms by modifying these but still there are some more challenges exist.Both PCA and LDA seek directions of the component but the main difference among them is PCA seeks direction for representation where LDA seek directions for efficient discrimination.The both considers that training data set are available in advance but in reality a complete training set might not given beforehand.The main part of the LDA algorithm is a generalised eigenvalue problem in within-class and between-class scatter matrices. LDA computes a linear transformation by maximising the ratio of between-class distance to the within-class distance to achieve the maximal discrimination. Linear transformation is computed by Singular Value Decomposition.These two matrices are computed in batch mode that means whole data set should be available to compute that. It is very much difficult to design an incremental solution for the eigenvalue problem on the product of scatter matrices[Ye+05].

In case of PCA,it represents 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[Li+03]. 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”.

One of the solution of this approach is to collect data whenever new data are presented and apply batch learning approach for the collected data so far. But the drawback of this approach is that it requires a large memory to store the data and high computational expenses are required. Moreover the system also forget about the knowledge that it acquires in past batch mode.

The batch method is now 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 where d=5632 which means we need a co-variance matrix of size d*(d+1)/2 elements which amounts to 15,862,828 entries[WZH03].

To solve this problem another way is to have an incremental method to compute principal components for sequential observation. The incremental method must be one-pass incremental learning.In this learning scheme, a system acquire knowledge wit a single presentation of the training data and retaining the knowledge acquired in the past without keeping a large number of training samples. To achieve this, several methods [HMM98; Cha+97; DR90] have been developed based on updating eigenspace models by perform incremental learning However all these learning considered one new sample to an eigenspace model at time [POK05].

In PCA, the principal component are updated based on each observation vector without not using covariance matrix.For example there are several Incremental Principle Component Analysis (IPCA) have been proposed to compute principal components without the co-variance matrices [LO96; OK85; San89]. However all of them run into convergence problems when facing with high dimensional factors [WZH03].To solve this problem a new algorithm called candid covariance-free incremental principal component (CCIPCA) is presented. This algorithm also compute the principal component of samples incrementally without estimating the covariance matrix by keeping the scale of observations and computes the mean of observations incrementally [WZH03].This algorithm is developed based on a well-known statistical concept called efficient estimate like some well-known distribution for example Gaussian distribution.This method works for real-time application that means it does not allow iterations.They used amnesic average technique instead of fixed learning rate to retain the old and new data.

Similar to PCA, one pass incremental algorithm is also proposed for the LDA algorithm as well. For example in paper [POK05] they give an incremental Linear Discriminant Analysis (ILDA). In their algorithm they give a solution of the challenge of not having corresponding class data with the arrival data however

it needs to have a high dimensional generalised eigenvalue problem.When the number of dimensions are too high then memory becomes an issue [YLO07].

The problem for Incremental Maximum Margin Criterion (IMMC) is to estimation selection of the parameter [Yan+06].

In replace of Singular Value Decompostion researcher also propose algorithm based on the QR decomposition [Ye+05].There are two stages for the algorithm where first stage is to maximise the separability between different classes. The corresponding stage belong to both between-class and within-class information by applying LDA on the "reduced" scatter matrices. The main advantage of using QR decomposition over SVD is in QR decomposition does not require the whole matrix to be saved because they consider appropriate matrices instead of the covariance matrix. The performance for this algorithm is same like other LDA algorithm but in term of computational cost is better than others. However, in some large-scale data the time and space cost are too expensive for example in web documents [Yan+06].

3 Our approach:

In this thesis, we will try to reduce the dimensions in streaming data by applying Entropy Minimization (EM). Before going to the the total approach, let us describe the Em algorithm in short.

3.1 Entropy Minimisation Algorithm:

EM is a theoretic approach which order the table such that the similar rows and similar column are grouped together [DV13].For ordering of rows, EM-ordering repeats until convergence of the following steps:

- Re-scaling columns
- Solving a travelling Salesman Problem (TSP) where rows are considered as cities and traverse around the examples and produce a traversing cost minimum path.

The main advantage of this method is it exchange rows and columns but does not lead to loss of informations.By doing so we reveal unknown regularities and patterns of the data. In many sector data ordering plays an important role for example, archaeology, anthropology, bioinformatics and geographical data.

EM algorithm basically helps us to visualise data efficiently when there is a high number of dimensions available.As mentioned in background section, there are many limitation of using PCA, LDA; along with that in these approaches there is a high chance of information loss.

There are many possible approaches for ordering data tables. In PCA, the main task is to find the principal component and then order the examples by traversing the line or the curve. Here ordering is the byproduct of PCA manifold search.In hierarchical clustering [Eis+98] approach the same ordering can be done by traversing the leaves of the tree. In spectral clustering [DH04] it gives not the optimal results if the cluster is not well formed.

In paper [DV13] they propose a novel ordering method which contribution are:

- Ordering can be done by doing permutation of rows or columns that ensures maximally compressible data set. The maximally can be determined by the entropy of the residuals of predictive coding.
- The problem is determined by an Expectation-Maximization algorithm which alternatively solves a TSP and assign reweights features based on the quality of the resulting tour.
- The proposed TSP solver known as TSP-means find the path comparable to those by LK algorithm. By applying K-means (k = 2) recursively the algorithm construct a Binary tree. The runtime of TSP solver is O (nlog(n))

3.2 Our Proposed Method:

We will try to use EM algorithm on Streaming Data. To achieve that we will propose an incremental approach similar the ILDA approach and then by using the same concept described in [Yan+06] we will propose Streaming Entropy Minimization procedure.

We will consider a sliding window approach. Sliding Window approach means in a particular time series data will be only seen once. The window will be keep on moving along the data streams and data that have been passed out of the window will be deleted. In my Thesis, the dimensionality reduction will be done in two steps:

- New data instances will be held on the proposed Incremental Entropy Minimisation algorithm
- Handle the deletion of the old data based on the proposed solution.

During the time of deletion of data there may be two situations arises. One situation may be the instance is the last one of the frame and the other is instance may be present on that cluster as well. We will try to handle both the situations.

References

[Cha+97] S. Chandrasekaran et al. "An Eigenspace Update Algorithm for Image Analysis". In: *Graphical Models and Image Processing* 59.5 (1997), pp. 321–332. ISSN: 1077-3169. DOI: http://dx.doi.org/10.1006/gmip.1997.0425. URL: http://www.sciencedirect.com/science/article/pii/S1077316997904251.

[DH04] Chris Ding and Xiaofeng He. "Linearized Cluster Assignment via Spectral Ordering". In: *Proceedings of the Twenty-first International Conference on Machine Learning*. ICML '04. Banff, Alberta, Canada: ACM, 2004, pp. 30–. ISBN: 1-58113-838-5. DOI: 10.1145/1015330.1015407. URL: http://doi.acm.org/10.1145/1015330.1015407.

[DR90] R. D. DeGroat and R. A. Roberts. "Efficient, numerically stabilized rank-one eigenstructure updating [signal processing]". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 38.2 (Feb. 1990), pp. 301–316. ISSN: 0096-3518. DOI: 10.1109/29.103066.

[DV13] N. Djuric and S. Vucetic. "Efficient Visualization of Large-Scale Data Tables through Reordering and Entropy Minimization". In: *2013 IEEE 13th International Conference on Data Mining*. Dec. 2013, pp. 121–130. DOI: 10.1109/ICDM.2013.63.

[Eis+98] Michael B. Eisen et al. "Cluster analysis and display of genome-wide expression patterns". In: *Proceedings of the National Academy of Sciences* 95.25 (1998), pp. 14863–14868. eprint: http://www.pnas.org/content/95/25/14863.full.pdf. URL: http://www.pnas.org/content/95/25/14863.abstract.

[HMM98] Peter M. Hall, David Marshall, and Ralph R. Martin. "Incremental Eigenanalysis for Classification". In: *in British Machine Vision Conference*. 1998, pp. 286–295.

[Li+03] Y. Li et al. "An integrated algorithm of incremental and robust PCA". In: *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*. Vol. 1. Sept. 2003. DOI: 10.1109/ICIP.2003.1246944.

[LO96] Jorma Laaksonen and Erkki Oja. "Subspace dimension selection and averaged learning subspace method in handwritten digit classification". In: *Artificial Neural Networks — ICANN 96: 1996 International Conference Bochum, Germany, July 16–19, 1996 Proceedings*. Ed. by Christoph von der Malsburg et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 227–232. ISBN: 978-3-540-68684-2. DOI: 10.1007/3-540-61510-5_41. URL: https://doi.org/10.1007/3-540-61510-5_41.

[OK85] Erkki Oja and Juha Karhunen. "On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix". In: *Journal of Mathematical Analysis and Applications* 106.1 (1985), pp. 69–84. ISSN: 0022-247X. DOI: http://dx.doi.org/10.1016/0022-247X(85)90131-3. URL: http://www.sciencedirect.com/science/article/pii/0022247X85901313.

[POK05] Shaoning Pang, S. Ozawa, and N. Kasabov. "Incremental linear discriminant analysis for classification of data streams". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 35.5 (Oct. 2005), pp. 905–914. ISSN: 1083-4419. DOI: 10.1109/TSMCB.2005.847744.

[San89] Terence D. Sanger. "Optimal unsupervised learning in a single-layer linear feedforward neural network". In: *Neural Networks* 2.6 (1989), pp. 459–473. ISSN: 0893-6080. DOI: http://dx.doi.org/10.1016/0893-6080(89)90044-0. URL: http://www.sciencedirect.com/science/article/pii/0893608089900440.

[WZH03] Juyang Weng, Yilu Zhang, and Wey-Shiuan Hwang. "Candid covariance-free incremental principal component analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.8 (Aug. 2003), pp. 1034–1040. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2003.1217609.

[Yan+06] Jun Yan et al. "Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing". In: *IEEE Trans. on Knowl. and Data Eng.* 18.3 (Mar. 2006), pp. 320–333. ISSN: 1041-4347. DOI: 10.1109/TKDE.2006.45. URL: http://dx.doi.org/10.1109/TKDE.2006.45.

[Ye+05] Jieping Ye et al. "IDR/QR: an incremental dimension reduction algorithm via QR decomposition". In: *IEEE Transactions on Knowledge and Data Engineering* 17.9 (Sept. 2005), pp. 1208–1222. ISSN: 1041-4347. DOI: 10.1109/TKDE.2005.148.

[YLO07] M. Ye, X. Li, and M. E. Orlowska. "Supervised Dimensionality Reduction on Streaming Data". In: *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*. Vol. 1. Aug. 2007, pp. 674–678. DOI: 10.1109/FSKD.2007.548.