TOWARDS REDUCTION OF DATA FLOW IN A DISTRIBUTED NETWORK USING PRINCIPAL COMPONENT ANALYSIS

Devadatta Sinha[1] and Anal Acharya[2]

[1]Department of Computer Science and Engineering,
Calcutta University, Kolkata, India
devadatta.sinha@gmail.com
[2] Computer Science Department, St. Xavier's College, Kolkata, India.
anal.acharya@sxccal.edu

ABSTRACT

For performing distributed data mining two approaches are possible: First, data from several sources are copied to a data warehouse and mining algorithms are applied in it. Secondly, mining can performed at the local sites and the results can be aggregated. When the number of features is high, a lot of bandwidth is consumed in transferring datasets to a centralized location. For this dimensionality reduction can be done at the local sites. In dimensionality reduction a certain encoding is applied on data so as to obtain its compressed form. The reduced features thus obtained at the local sites are aggregated and data mining algorithms are applied on them. There are several methods of performing dimensionality reduction. Two most important ones are Discrete Wavelet Transforms (DWT) and Principal Component Analysis (PCA). Here a detailed study is done on how PCA could be useful in reducing data flow across a distributed network.

KEYWORDS

Distributed Data Mining (DDM), Principal Component Analysis (PCA), Eigen Vector, Dimensionality Reduction.

1. INTRODUCTION

In the recent years data mining has gained a lot of importance. This is mainly due to what is called as Knowledge Discovery from Data (KDD). This mainly deals with searching of interesting patterns from large data sets. As stated in [2], data mining is an essential step in the process of knowledge discovery. [5] classified three types of data mining algorithms: the first type deals with the application of data mining algorithms on single valued vectors, the second type deals with large data sets and the third type deals with distributed databases.

The focus of this paper is on distributed databases. This is because data today is maintained in a distributed nature due to changing nature of business application. Also these applications are scalable with the growth of business. Distributed Data Mining (DDM) can be carried out following two models [1, 6]: Firstly, all data from each of the locations be transferred to a centralized site and integrated to form a data warehouse. Then the required patterns can be derived using a suitable data mining algorithm. In the second approach, a summary of the data at each location is sent to the centralized locations [9]. The advantage of this approach is that it involves much less data transfer. Basically what is done is that a certain mapping is applied on

Rupak Bhattacharyya et al. (Eds) : ACER 2013,
pp. 287–293, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3227

that data to reduce its dimension. A reverse mapping is then applied to get the original data. Thus data flow is reduced across the distributed network.

To this end we assume a very simple framework that can perform distributed data mining. The framework in brief consists of the following steps:

1. Data Preprocessing: This consists of data cleaning, dimension reduction etc.
2. Data Integration: Data from multiple sources are combined.
3. Data Mining: Various algorithms are applied on data to obtain interesting patterns.
4. Knowledge Discovery: Use the pattern thus obtained to perform analysis.

The paper concentrates on the first step i.e. data preprocessing. Piramuthu[9] concludes that 80 percent of the resources in a majority of data mining applications are spent on preprocessing of data. This is because databases typically consist of data which are noisy, inconsistent and contain missing values due to integration of data from various sources. One of the methods of data preprocessing is dimension reduction. Two methods are studied here: Principal Component Analysis (PCA) and Discrete Wavelet Transforms (DWT). In particular, we study how PCA can be used to reduce data flow in a distributed system. We construct a distributed data mining architecture along these lines. The PCA algorithm is next studied in details. This algorithm is then applied to distributed systems. We next determine the computing complexity of the algorithm. Finally, a simulation is done on sample data to illustrate how dimensionality reduction is done at local sites.

2. LITERATURE SURVEY

The papers surveyed can be divided into two major categories: the models and the corresponding mathematical tools. Kargupta el al[1] has done an exhaustive study on Distributed Data Mining(DDM). In this he proposes a framework for DDM. Various data preprocessing techniques for heterogeneous and homogeneous data sites are discussed in details. In [3] an agent based DDM architecture is developed. The concepts of Parallel Data Mining agents are used. Qi et al[4] in their paper proposes a novel method for performing distributed DDM. Faisal et al.[5] in their work developed an algorithm called Distributed Fast Map for performing dimension reduction. Mohanty et al[8] proposed a classification technique based on chi-square testing and then selecting the best features. In [9] Piramuthu developed a novel method of feature selection using a probabilistic distance based method. Association rules can also be used, particularly in text mining[10]. The use of unsupervised learning has been used for online feature selection in [11]. Finally, an evaluation technique called stability index has been proposed in [12].

Li[6] in his work made a survey on the use of wavelets in data mining. It contains a detailed study on the use of wavelets in Data Management, Preprocessing and various data mining algorithms. The use of wavelets in clustering and classification is particularly interesting. Finally, Smith [7] has given very lucid account of the mathematics involved in the working of PCA.

3. DIMENSION REDUCTION

While processing large databases the problem of high dimensionality of feature space is encountered. For example a row in a particular table may contain several columns. In a distributed database, it is virtually impossible to import data of such dimension to a centralized location. Thus to improve computational efficiency, data is projected in a smaller space[6]. Smaller space also helps in visualizing and analyzing data. The goal here is to create a low dimension space which yields maximum information. We propose the use two mathematical tools for this purpose. PCA is discussed in details here. We give a basic outline of our use of Wavelets here. We leave the details for future work.

3.1 Wavelets

We suppose that we want to reduce the data to p columns where the original data set consists of n columns, p<<n. We apply the simplest type of wavelet called Haar Wavelets on the feature vector. This gives us a set of approximations and detailed coefficients. Their number decreases as we increase the number of resolutions. Thus we can keep the first p coefficients and approximate the others with zeros. A reverse transformation is applied on these to yield the original data.

3.2 Principal Component Analysis(PCA)

Now we propose the use of PCA algorithm for dimension reduction. It is to be noted that the resulting vector set is orthogonal in nature. Thus we are able to reduce the vector set into much smaller space. Here we enumerate the steps for performing PCA:

Step 1: For simplicity we assume a bi-variate data set {X,Y}.

Step 2: Compute the mean. We assume they are m1 and m2. We apply the transformation {X-m_1,Y-m_2} on this data set. For simplicity we call them {P,Q}. This matrix is renamed M.

Step 3: We compute the covariance matrix,

$$\begin{bmatrix} \mathrm{Cov(P,P)} & \mathrm{Cov(P,Q)} \\ \mathrm{Cov(Q,P)} & \mathrm{Cov(Q,Q)} \end{bmatrix}$$

We call this matrix C.

Step 4: We compute the Eigen values E and Eigen vectors of the covariance matrix.
E= (e1,e2). The Eigen vector is given as

$$\begin{bmatrix} V_{11} & v_{12} \\ V_{21} & v_{22} \end{bmatrix}$$

We call this vector V. We note that this vector is a unit vector.

Step 5: We find the feature vector. This is the Eigen vector with the highest Eigen value. This could be any one of $[v_{11}\ v_{21}]^T$ or $[v_{12}\ v_{22}]^T$.

Step 6: Derive the transformed data matrix D:

$$D= V^T*M^T$$

Step 7: Get back the original data using the transformation:

$$O= V^T*D$$

Step 8: Add the mean that was originally subtracted.

The above algorithm can be extended to any number of variables. The use of this algorithm in distributed database is discussed in the next section.

4. THE FRAMEWORK

The data mining architecture that we use is shown in the figure below:

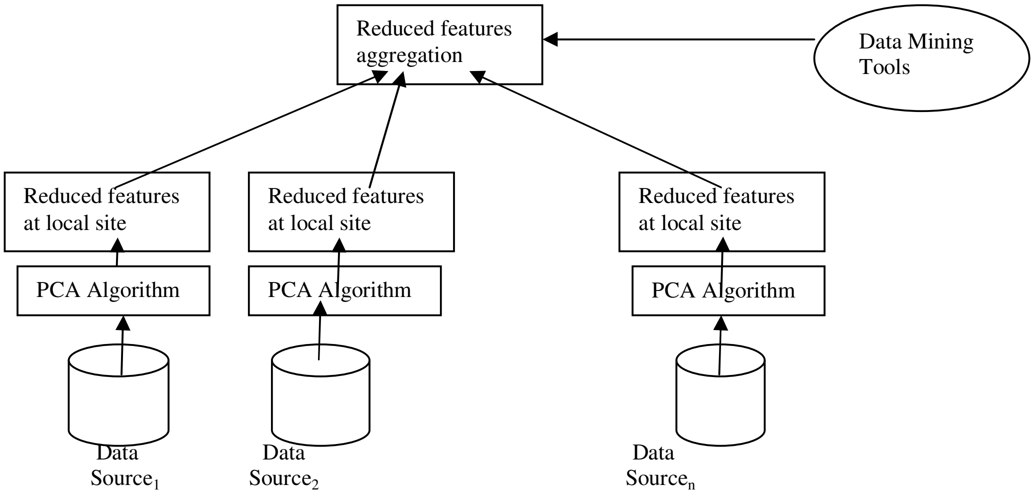
Reduced features aggregation
Data Mining Tools
Reduced features at local site
Reduced features at local site
Reduced features at local site
PCA Algorithm
PCA Algorithm
PCA Algorithm
Data Source$_1$
Data Source$_2$
Data Source$_n$

Fig 1: Our framework for DDM using PCA

We assume that there are n data sources represented by db_1, db_2,....... ,db_n. We assume that the database is homogeneous in nature. Each of the rows has n vectors. We apply the PCA algorithm stated earlier at each site and reduce it to k vectors, $k<n$. This gives us a uniform reduced set of vectors at each local site. These vectors are then transferred to a centralized location to form aggregate data set. Data mining tools can then be applied on this data.

5. COMPUTATIONAL COMPLEXITY

For simplicity, we perform this computation on a single site db_1 and the rest follows from it. As stated earlier, using PCA we reduce the dimension to p from N where $p<<N$. If the number of rows in the table is m, without the use of PCA the amount of data to be transmitted is of the order O(mN) whereas if we use PCA it reduces to O(mp). Clearly, considering that there are n data sites, the complexity reduces to O(nmp) from O(nmN). Thus a large amount of reduction of computational complexity is achieved by using PCA in a distributed network.

6. RESULTS

The above algorithm was implemented on a core i3 processor of 3.3 GHZ. For simplicity, we perform PCA on the following bi-variate data set:

X	0.5	2.2	3.1	2.3	1.0	1.5	1.1
Y	0.7	2.9	3.0	2.7	1.1	1.6	0.9

Table 1: Original Data

This yields the following graph:

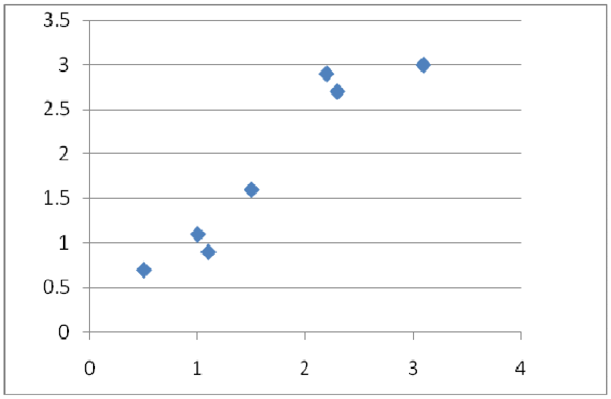

Fig 2: Original data Plotted

We follow the steps enumerated earlier and obtain the covariance matrix:

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix}$$

Using a 'C' program we obtain the following Eigen values [0.0491 1.2840]

These yield the following set of Eigen Vectors

$$\begin{bmatrix} -0.7352 & -0.6779 \\ 0.6778 & -0.7352 \end{bmatrix}$$

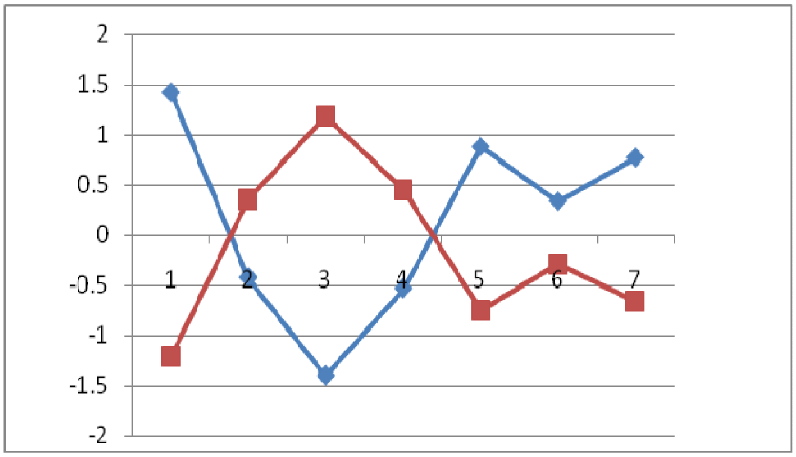

Fig 3: Normalized data on application of the Eigen Vectors obtained from the covariance matrix.

The red line gives us the feature vector

$$\begin{bmatrix} -0.6779 & -0.7353 \\ -0.7352 & 0.6772 \end{bmatrix}$$

The transformed data is obtained as

P	1.7775	-0.9922	-1.6758	-0.9129	1.1445	0.4380	1.2238
Q	0.1428	0.3843	-0.2094	0.1753	0.0464	0.0177	-0.1626

Table 2: Data obtained after applying PCA

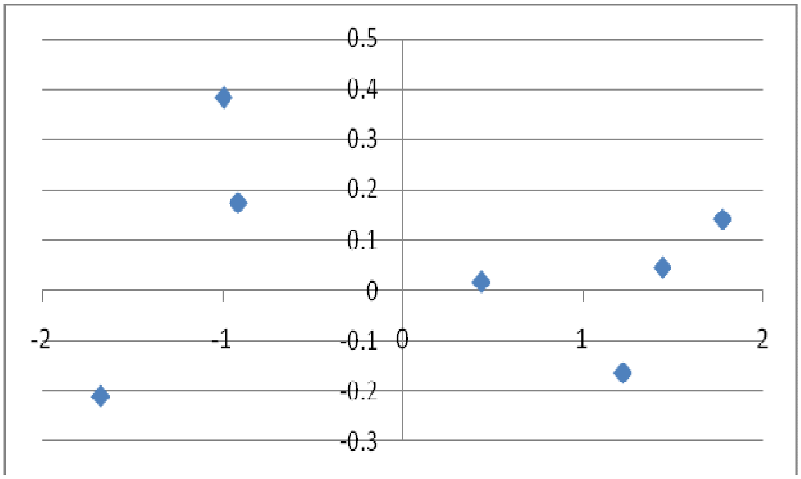

Fig 4: Plot of the data obtained after applying PCA

Original data can be obtained by applying the reverse transformation as discussed in the earlier section.

7. Conclusion And Future Work

In this paper we made a study of a method used for dimension reduction in context of DDM. There are other methods as well like applying Discrete Wavelets Transforms (DWT) on feature vector to reduce dimensions. The architecture we have developed in rudimentary in nature. We could expand this framework by adding agents to perform suitable tasks. Also some modelling techniques could be used to model these. We leave these all for our future work.

References

[1] Kargupta,Park, (2002), “Algorithms, Systems and Applications”, ABC *Transactions on ECE*, Vol. 10, No. 5, pp120-122.

[2] Haan and Kamber (2011) Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufman Publishers.

[3] Kargupta (2003), “Scalable, Distributed data mining using agent based architecture”.

[4] Qi,Wang,Birdwell, (2001), “Global Principal Component Analysis for performing dimension reduction in distributed data mining”.

[5] Faisal,Samatova,Ostrouchov, (2007), “Distributed Dimension Reduction Algorithms for Widely Dispersed Data”.

[6] Li,Zhu,Oghihara, (2009), “A survey on the use of wavelets in data mining”, SIGKDD Explorations, Volume 4, Issue 2, page 49-69

[7] Smith (2009),”A Tutorial on Principal Component Analysis”

[8] Bhuyan,Mohanty,Das (2012), “Privacy Preservation for feature selection in data mining using Centralized network” IJCSI, Vol. 3, No. 2.

[9] Piramuthu, (2004), “Evaluating feature Selection methods for learning in data mining applications”, European Journal on Operations Research”, Vol. 9.

[10] Do,Hui,Fong, (2007), “Associative feature Selection methods for text mining”.

[11] Hoi,Wang,Zhao, (2005), “Online Feature Selection for Mining big data”, Bignine, China

[12] Kurcheva, (2012), “A stability index for feature selection” 25th IASTAD International Conference on Artificial Intelligence, Austria

[13] Kargupta, Huang, SivaKumar,Johnson, (2001), “Distributed clustering using Collective Principal Component analysis”, Knowledge and information Systems , Springer-Verlag,, London.

AUTHORS

Prof Devadatta Sinha is a Professor in University of Calcutta. He has over 30 years of teaching and research experience. His interests include Distributed Processing, Software Engineering.

Prof Anal Acharya is currently Head of the Department of Computer Science, St Xavier’s College, Kolkata. His research interests include Distributed Data Mining Algorithms.