2012 International Conference on System Engineering and Modeling (ICSEM 2012)
IPCSIT vol. 34 (2012) © (2012) IACSIT Press, Singapore

Efficient and Scalable Data Transmission in Dimension Reduction Techniques of Distributed Data Mining

A. Anbarasi[1] [+]and S.Santhosh Baboo[2]

[1] Research Scholar/Dept of Computer Science, Bharathiar University
Coimbatore, 641 046, Tamil Nadu, India.
[2] Postgraduate and Research department of Computer Science
D G Vaishnav College, Chennai, Tamil Nadu, India.

Abstract. Dimension reduction for large volume of data is attracting much concentration nowadays due to the fast growth of the World Wide Web. We can classify those popular dimension reduction algorithms into two groups: feature extraction and feature selection algorithms. In the former, new features are combined from their original features through algebraic transformation. In this research paper, a new Encode and decode algorithm is proposed, which is different from all the existing methods, to encode the transactions which reduces the size of transaction that in turn reduces the time and communication cost.
Existing data mining methods for distributed data are of communication rigorous. Many algorithms for data mining have been proposed for a data at a single location and some at multiple locations with improvement in terms of efficiency of algorithms as a part of quality but effectiveness of these algorithms in real time distributed environment are not addressed, as data on the network are distributed by very of its nature. As an upshot, both new architectures and new algorithms are needed. In this paper we introduce the encode and decode technology that supports building of distributed data mining architecture and explore the capabilities of data transmission, it is suited for distributed data Mining compared to traditional methods like client server computing.

Keywords: Encode, Decode, Dimension reduction, & Communication

1. Introduction

The development of network based computing environments has introduced a new and important dimension i.e distributed foundation of data and computing. The coming of laptops, palmtops, mobile phones, embedded systems, and wearable computers are also making everywhere access to a large quantity of distributed data a reality. Advanced analysis of distributed data for extracting useful knowledge is the next natural step in the increasingly connected world of ubiquitous and distributed computing. Most of the popular data mining algorithms are designed to work for centralized data and they often do not pay attention to resource constraints of distributed and mobile environments. Recent research in this area has demonstrated that handling these resource constraints in an optimal fashion requires a new breed of data mining algorithms and systems that are very different from their centralized counterparts

Many dimensionality reduction algorithms have been proposed, most of which are linear, such as the commonly used Principal Component Analysis (PCA) and classical multidimensional scaling. Linear methods can discover the true structure of data lying on a linear space but may be inadequate to those with nonlinear structures.

All these manifold learning algorithms share a common characteristic in that globally optimal dimensionality reduction is achieved by assembling together the local geometrical information learned from the data. But the local geometrical information is evaluated in different ways in these algorithms.

2. Existing Systems

[+] Corresponding author.
E-mail address: anbarasi2@gmail.com.

Distributed storage is the technique of storing a single data set across multiple hosts. This paper focuses on massive localized systems for distributed storage which are directed at overcoming the capacity and performance limitations of single-host, or huge, storage systems. One of the problems with high-dimensional datasets of astronomy, biology, remote sensing, economics, and consumer transactions, is that, in many cases, not all the measured variables are important for understanding the underlying phenomena of interest. While certain computationally expensive novel methods [4] can construct predictive models with high accuracy from high-dimensional data, it is still of interest in many applications to reduce the dimension of the original data prior to any modeling of the data.

3. Problem Methodology

Data Storage System transforms a transaction into a single dimension transaction with all attributes that appears in its original form. The encoded transactions are represented by a sequence of numbers. The sum of subset approach techniques should be followed. By this way, the new transaction is smaller than the original form and hence the cost of storage is reduced.

Certain transaction set of data items Z={x, y, z}, the power set of Z, is in fact written as possible P(Z)={{}, {x}, {y}, {z}, {x,y}, {x,z},{y,z},{x,y,z}),. If set S is assumed as set of powers of 2, i.e. for example S = {2, 4, 8, 16}, then the power set P1(Z) = {{2}, {4}, {8}, {16}, {2, 4}, {2, 8}, {2, 16}, {4, 8}, {4, 16}, {8, 16}, {2, 4, 8, 16}}. In this fashion, the sum of the subsets are matchless i.e. 2, 4, 8, 16, 6, 10, 18, 12, 20, 24, 30.

3.1 Algorithm

For Calculating Dimension

Input: Number of data's needed from set
Output: Extract the data's as per rows and columns

```
k=input ('Enter the Datasets to be extracted :');
c=k;
disp(c);
row=round(c/2)
col=round(k-row)
for i=1:m:n
   for j=1:m:n
 p=[i:m,j:n];
 X=A(i:m,j:n);
 end
end
K=X
```

The above algorithm gives us the extract data's from the set based on the calculation of rows and columns.

For making the given data's as set and perform for checking.

```
N=input('Enter the number of values:');
names=cell(1,N);
for t=1:N
   e2=input ('Enter the items:','s');
```

```
names {t} = e2
end
```

The above algorithm creates a cell arrays for the user defined data's that is need to be checked with the data set.

Comparing the values in dataset and Encoding

```
if(strcmp(names(t),K(i1,j1)))
    fprintf('\n Item Found\n')
          K1=2^j1;
 k2=cell(1,N);
names1{t}=K1
 else
 fprintf('\n Item Not Found\n')
end
```

The above algorithm performs the comparison operation between the user data and the data set. If the item is found in the set it encodes the data based on the sum of subset approach specified and creates a cell arrays which gives the encoding data's in the form of numerical representation values.

3.2 Process of Algorithms

Mention the dimension of the matrix (cell arrays) so that the specified items will be displayed. From the dataset, which is available to sort and select the needed items. The needed items are chosen based on the number of values needed. After that the position of the selected data are taken and based on the formulae which is depicted the values are calculated and applied to cell array. If the item is not found in the dataset which is available go for the selection of next item. Or else if the data set seems to be empty or the searched item is not found, insert the needed items in the data set and again continue the same process. If the selected item is there continue the calculation of values and update the newly created dataset or cell arrays.

3.3. Comparison of Centralized and Decentralized Mining of Data

The comparative performances of these two cases are shown in figure1. From figure-1, it is clear that, this kind of data transformation better in terms of time and network bandwidth usage compared with traditional system built over client and server technology. The percentage improvement of performance increases with decrease in the number of patterns find in each site because, more the number of patterns, implies, more data to be carried by migration to central site.

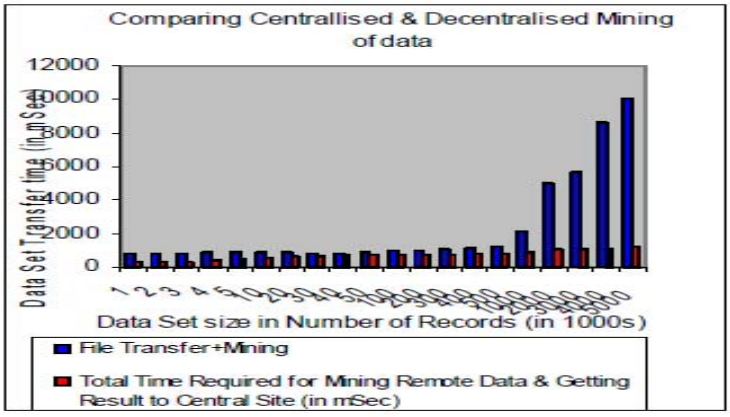

Figure I Comparing & Decentralized mining of data

4. Conclusion

While the technique of building storage clusters at this level is relatively new and the essential technology in a state of quick change, sound design methodologies can be developed. Existing systems provide examples of design considerations and a proof of concept for future systems.

Compared to the existing applications of similar kind, this encoding and decoding techniques application proves to be one of the best and healthy methods to handle distributed data and consequently distributed data mining, in faster way. This also addresses the issue of handling of dynamically generated data, because the data is maintained at remote sites, which can be updated continuously, or as and when it is required. Since this approach does not require the huge amount of data transfer from remote to central site, the network resources are used optimally. The development of methodology for consolidating the collected Knowledge from different sites is in progress.

5. References

[1] H. Abdi. Partial least squares (PLS) regression. 2003.

[2] Y. Akbas, C. Takma. Canonical correlation analysis for studying the relationship between egg production traits and body weight, egg weight and age at sexual maturity in layers Czech Journal of Animal Science, 50, pp.163–168, 2005 (4)

[3] A. Boulesteix. PLS dimension reduction for classification with microarray data. Statistical Applications in Genetics and Molecular Biology, 2004.

[4] L. Breiman. Random forests, Technical report, Department of Statistics, University of California, 2001.

[5] Deon Garrett, David A. Peterson, Charles W. Anderson, and Michael H. Thaut. Comparison of Linear, Nonlinear, and Feature Selection Methods for EEG Signal Classification IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 11 Issue. 2, pp.141 – 144, 2003

[6] P.H. Garthwaite. An interpretation of partial least squares. Journal American Statistical. Association. 89, pp.122–127, 1988.

[7] A. J. Guarino, A Comparison of First and Second Generation Multivariate Analyses: Canonical Correlation Analysis and Structural Equation Modeling 1, Florida Journal of Educational Research, 2004, Vol. 42, pp. 22 – 40 22

[8] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with applications to learning methods, Neural Comput., vol. 16, pp. 2639– 2664, 2004.

[9] A. HIoskuldsson. PLS regression methods, Journal of Chemometrics. 2, 211–228. 1988.

[10] J.E. Jackson. A User's Guide to Principal Components, New York: John Wiley and Sons, 1991.

[11] John Aldo Lee, Amaury Lendasse, Michel Verleysen. Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis Neuro computing, 2004 (49 – 76)