Data Reduction Techniques in Data Mining

A.Anbarasi & S. Santhosh Baboo

Abstract

Dimension reduction of datasets is very useful in different application including classification, compression, Text documents, microarray gene-expressions, images, etc The challenge in distributed data mining is how to learn as much knowledge from distributed databases as we do from the centralized database without costing too much communication bandwidth. In this paper we propose a data reduction algorithm, which is different from all of these methods, to encode the transactions which reduce the size of transaction that in turn reduces the transfer time, size of the transaction, cost as well as increase the security level of data.

Keywords : Reduction, compression , & encode

Introduction

A dimension refers to a capacity of a certain aspect of an entity. Dimensionality reduction is the learning of methods for reducing the number of dimensions relating the entity[4]. Its general objectives are to remove irrelevant and redundant data to reduce the computational cost and avoid data over-fitting [1],[2] and to improve the quality of data for efficient data- determined processing tasks such as Data transmission and storage system in distributed data mining. Dimensionality reduction is an useful solution to the problem of "curse of dimensionality"[3].

In practice, researchers and practitioners interchangeably use dimension, feature, variable, and attribute[5]. Correspondingly, we will interchangeably use entity, example, vector, and instance. Consider an application in which a system processes data (speech signal, images, or patterns in general) in the form of a collection of vectors[8]. For a particular application, it is more often than not that a subset of features is relevant and in some cases, a large number of features are irrelevant. This problem can be caused by factors such as: (1) many dimensions will have variation smaller than the measurement noise and thus will be irrelevant, and (2) many dimensions will be correlated (through linear combinations or functional dependence) to others and thus will be redundant[7]. Therefore, in many situations, it is recommended to remove the irrelevant and redundant dimensions, producing a more economical

II. 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.

2.1 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.

2.2 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.

2.3. 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.

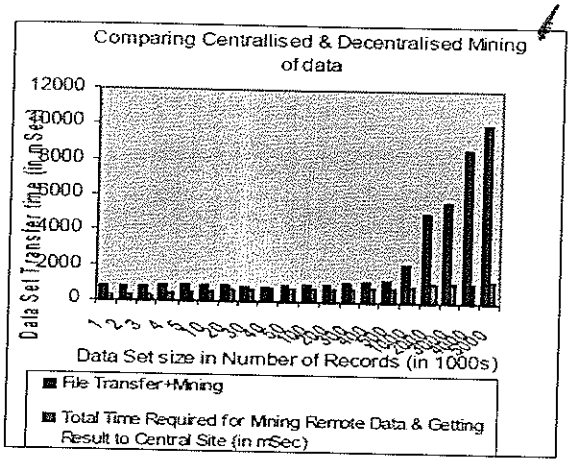

Figure I Comparing & Decentralized mining of data

III. Conclusion

As computers become increasingly powerful, many applications can produce massive data of high dimensionality. Dimensionality reduction is an efficient way of dealing data with high dimensionality. The purpose is to reduce the data so that computational load decreases and better quality can be extracted in data mining algorithms. In this paper, we described the concepts of feature extraction and feature selection, and briefly introduced some representative methods. The need of dimensionality reduction techniques presents new challenges, and novel methods are expected to be developed.

References

1. Ng. A. Y. Preventing overfitting of crossvalidation data. In Proceedings of Fourteenth International Conference on Machine Learning, pages 245–253, 1997.

2. L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 5:1205–1224, Oct 2004.

3. Shisong Yang and Chih-Cheng Hung. Image texture classification using datagrams and characteristic views. In Proceedings of the 2003 ACM symposium on Applied computing, pages 22–26, 2003.

4. E. Xing, M. Jordan, and R. Karp. Feature selection for high-dimensional genomic microarray data. In Proceedings of the Eighteenth International Conference On Machine Learning, 2001.

5. U.P,Kulkarni, K.K. Tangod, S.R.Mangalwede, A.R.Yardi, " Exploring the capabilities of Mobile Agents in Distributed Data Mining, 10th International Database Engineering and Applications Symposium(IDEAS'06), 2006 IEEE.

6. Talia, D."Grid-Based Distributed Data Mining Systems, Algorithms and Services, 9th International Workshop on High Performance and Distributed Mining, Bethesda April 22 2006

7. Wu-Shan Jiang, Ji-Hui Yu., Distributed Data Mining

8. on the Grid, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005

Author's Biography

A.Anbarasi is presently working as Assistant Professor in MCA Dept at Karpagam Institute of Technology. She is doing PhD under Bharathiar University. The research area is Data Mining. She has over 12 years of teaching experience. She has presented above 20 papers in various National and International conferences and published 8 research paper and some articles in leading magazines. She has written three different books, which hold computer science related subjects. Membership in ACM

Mr. S.Santhosh Baboo is presently working as a associate professor, D.G vaisnav college. PG and Research department, Chennai. He has 18 years of teaching Experience. He has published more than 50 research paper in National and International and participated more than 30 Seminars and Conferences.