2011 International Conference on Information and Electronics Engineering
IPCSIT vol.6 (2011) © (2011) IACSIT Press, Singapore

Dimension Reduction Techniques for Market Basket Analysis

Anbarasi[1], D. Sathya Srinivas[2] and Dr. K. Vivekanandan[3]

[1] MCA Department, Karpagam Institute of Technology, Coimbatore. India.

[2] Department of Computer Applications, Karpagam University, Coimbatore. India

[3] Management Studies, Bharathiar University, Coimbatore. India.

Abstract. Market basket analysis helps us to discover which group of items tends to be purchased together by customers. This is a powerful tool in data mining to understand the purchase behavior. Today the data sizes in the datasets of market basket have increased from gigabytes to terabytes or even larger due to which the complexity of analysis of huge datasets has been a major concern in almost all areas of technology in the past decade. A solution to this crucial problem in distributed data mining is that, the massive dataset can be collected and warehoused in a single site if its dimensionality is reduced. Today many dimension reduction algorithms are there and they are generally classified into feature selection, feature extraction and random projection. In this paper a new dimension reduction 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 communication cost and communication bandwidth.

Keywords: Market basket, Centralized Database, Distributed Data Mining, Dataset Dimension reduction, Power set, Sum of subsets.

1. Introduction

The challenge in distributed data mining is how to learn as much knowledge from the distributed databases as we do from the centralized database without costing too much communication bandwidth

Gathering all the data in a centralized location, as per first framework, is generally neither desirable nor feasible because of communications costs and storage requirements. Techniques that can reduce the communications costs, by reducing the input data size or transmitting the results of analysis (knowledge acquired from the data) to the locations where they are needed, are essential for second framework. Ideally, such techniques should compromise on result quality and comprehensiveness as little as possible. For these reasons, we seek a dimension deduction algorithm that reduces communication in a way that does not require discarding any attribute, rather representation of transaction reduced in size compared to its original form. [4][6].

2. Previous work

Linear Discriminant Analysis (LDA) was the first statistical criterion for low rank linear separation, and it is still the most popular supervised linear feature extractor **[2]**. LDA tries to maximize the dispersion among classes while minimizing the inner dispersion of each class, which is known as Fisher criterion.

PLS is similar to principal components analysis (PCA) **[1, 3].** PLS seeks for a linear combination of attributes whose correlation with the class attribute is maximum. In PLS regression, the task is to build a linear model, $Y = BX + E$, where B is the matrix of regression coefficients and E is the matrix of error coefficients. In PLS, this is done via the factor score matrix $Y = WX$ with an appropriate weight matrix W. Then it considers the linear model, $Y = QY +E$, where Q is the matrix of regression coefficients for Y. Computation of Q will yield $Y = BX+E$, where $B = WQ$.

3. Proposed System

3.1. Market basket analysis

Market basket analysis requires the analysis and mining of large volumes of transaction data for making business decisions. Today it as become a key success factor than ever for vendors to understand their customers and their buying patterns. If they don't they will loose them. In order to gain competitive advantage it is necessary to understand the relationships that prevail across the data items among millions of transactions. The amount of data currently available for studying the buying pattern is extensive and increasing rapidly year by year. Therefore the need to devise reliable and scalable techniques to explore the millions of transactions for the customer buying pattern continues to be important. Above this, the increasing volume of data sets data demands for huge amounts of resources in storage space and computation time. As it is not feasible to have huge storage spaces to store the explosively growing data in a single location they are stored in distributed database and data warehouse located in different geographical location. Inherently data distributed over a network with limited bandwidth and computational resources motivated the development of distributed data mining (DDM).

3.2. Problem Statement

This paper introduces a successful proceed to the difficulty of dimension reduction which makes the reduction method very beneficial and also contribute a large amount data in the compact representation than traditional dimensionality reduction techniques. Here dimensionality reduction is arrange as follows: Given a transaction set of data items S={a, b, c} , the power set of S, is actually written as P(S)={{}, {a}, {b}, {c}, {a,b}, {a,c},{b,c},{a,b,c}),. 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(S) = {{2}, {4}, {8}, {16}, {2, 4}, {2, 8}, {2, 16}, {4, 8}, {4, 16}, {8, 16}, {2, 4, 8, 16}}. An attractive summation found in this power set is that the sum of the subsets are unique i.e. 2, 4, 8, 16, 6, 10, 18, 12, 20, 24, 30.

The encoded transactions are represented by a sequence of numbers. By this way, the new transaction is smaller than the original form and hence the cost of communication is reduced.

Let **D** = {T_1, . . . , T_K} be a database of customer transactions at a store. Each transaction, says T_i, is a collection of items purchased by a customer. A non-empty set of items is called an itemset. An itemset is denoted as **I**={i_1,i_2,…,i_n}, where each i_j is an item from the ordered set either containing all items in the store or interested items in store. Each transaction in the database is an itemset and is a subset of ***I*** (***T*** $\subseteq$ ***I***). Items in itemset **I** is ordered and stored in an **m** × **n** matrix **C**. Each column of **C** corresponds to an item. Item in each column in each row in **C** is assigned a numeric a value from the set {2^1, 2^2, 2^3,.,.,.,., 2^n). For example, the itemset I of 15 items can be stored as shown in Table 1.

Table 1: Representation of item set **I** in **C**

pizza	sauce	sugar	sweet bun	
egg	fruit bread	honey	jam	milk
bread	bun	burger	butter	cheese

T_1 is reduced to 34, 50, 10.
The numeric value assigned to items in row 1 is 2^1, 2^2, 2^3, 2^4, 2^5, 2^6. The same sequence is assigned to items in row 2 and row 3.

The reduced form of transactions in table 2 is given in table 3.
The matrix **E** records the reduced version of transactions in **D**. The number of columns in the row in **E** is equal to the number of rows in **C**. If transaction T_i contains single item from row 1 of **C** then the numeric value that represent that item in **C** is added to the value in column 1 of **E**. If the transaction T_i contains more than one item or all the items in row i of **C** then the sum of the numbers representing the corresponding items in row i in **C** is stored in column i of **E.**

Table 2: Reduced database RD

34	50	10
58	16	02
18	16	18
56	00	00
38	28	10

4. Conclusion

In this paper an algorithm has been introduced for saving network bandwidth. There are two important variables that influence the network bandwidth: the number of transactions and the number of items in a transaction. This dimension reduction algorithm is dependent on the number of transactions. The number of items in a transaction is irrespective. The authors strongly recommend this approach for market basket analysis since it significantly reduces the communication time. That is an approach that we intend to investigate in the near future.

5. References

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

[2] D. Ridder, O. Kouropteva, O. Okun, M. Pietikainen, and R. Duin. Supervised locally linear embedding, in Proc. Artif. Neural Netw. Neural Inf. Process., 2003, pp. 333–341.

[3] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding, Science 290: 2323-2326.

[4] Syed Zahid, Hassan Zaidi, Syed Sibte Raza Abidi and Selvakumar Manickam. Distributed Data Mining From Heterogeneous Healthcare Data Repositories: Towards an Intelligent Agent-Based Framework, Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002)

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

[6] J. J. Verbeek, S. T. Roweis, and N. Vlassis. Nonlinear CCA and PCA by alignment of local models. In Advances in Neural Information Processing Systems 16, 2000

[7] Wu-Shan Jiang, Ji-Hui Yu. Distributed Data Mining on the Grid, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005.