Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 48 (2015) 548 – 554

Procedia Computer Science

International Conference on Intelligent Computing, Communication & Convergence

(ICCC-2015)

Conference Organized by Interscience Institute of Management and Technology,

Bhubaneswar, Odisha, India

Lowering Data Dimensionality in Big Data For The Benefit of Precision Agriculture

Sabarina K [1], Priya N[2]

[1]*PG Student, Jerusalem College of Engineering, Chennai*

[2]*Assistant Professor, Jerusalem College of Engineering, Chennai*

Abstract

Predictive analytics can be used to make smarter decisions in farming by collecting real-time data on weather, soil and air quality, crop maturity and even equipment and labor costs and availability. This is known as precision agriculture. Big data is expected to play an important role in precision agriculture for managing real-time data analysis with massive streaming data. The data analysis efficiency and throughput would be a challenge with the massive increase in size of big data. The unstructured streaming data received from different agricultural sources would contain multiple dimensions and not the entire content is needed for performing analysis. The core data which is small but that alone enough to represent the entire content should be extracted. This paper explains how to systematically reduce the size of big data by applying a tensor based feature reduction model. The data decomposition and core value extraction is done with the help of IHOSVD algorithm. This way it reduces the overall file size by eliminating unwanted data dimensions. The time involved in data analysis and CPU usage will be significantly reduced when dimensionality reduced data is used in place of raw (unprocessed) data.

Keywords: Big data; Precision agriculture; Dimensionality reduction; HOSVD.

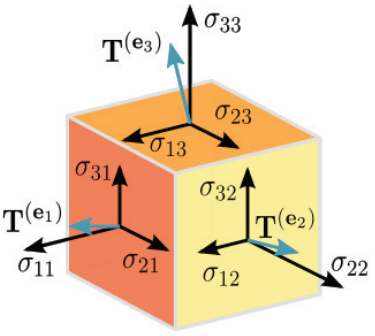
1. Introduction

The current ratio of agricultural land to world population is sufficient to meet the food requirements of the world and hence we are not much concerned about agricultural productivity. However the global population is expected to grow exponentially in the coming years wherein the total area of agricultural land is getting reduced day by day.

Peer-review under responsibility of scientific committee of International Conference on Computer, Communication and Convergence (ICCC 2015)
doi:10.1016/j.procs.2015.04.134

Hence it is very essential to switch from conventional agricultural methods to modern precision agriculture for increasing the agricultural productivity and maintaining safer food supply. The precision agriculture functions together with information technology to make a difference in the agriculture methodologies. It is expected that the precision agriculture would help us to reach sustainable agriculture where the food demand and supply will always be maintained at safer level. The Data Mining and Big Data technologies are advancing in agriculture, changing how various agricultural activities are sequenced. Making predictions about crops based on the real-time data gathered from the farms will change the current scenario in agriculture [10]. Improvising the speed of data analysis by reducing the data dimensions will help the farmers getting the tips on time. Agricultural predictions done using data mining techniques may not be effective when the data dimensions are high. The query accuracy and efficiency degrade rapidly as the dimension increases. Dimensionality reduction is an effective approach for reducing the dimensions and size of data.

The most popular approaches that are involved in data dimensionality reduction include Principal Component Analysis (PCA) [2], Incremental Singular Value Decomposition (SVD) [7], and Dynamic Tensor Analysis (DTA) [4]. These methods works fine for low dimensional data but suffer from serious performance issues when being applied on high-dimension data and fail to extract the core data sets from streaming big data. This paper explains an incremental dimensionality reduction method for reducing data dimensions and extracting the core set of data from massive big data.



Recursive Incremental HOSVD Method: The high dimensional streaming data is converted into tensors and then the tensors are unfolded into matrices. Later the IHOSVD algorithm is recursively applied to extract the core tesnsor by decomposing the matrices [1]. The IHOSVD algorithm is explained in a later section of this paper.

2. Preliminaries

2.1 Tensors

Tensors are mathematical elements that describe physical properties similar to vectors and scalars. They help to establish mathematical relationship between vectors, scalars, and other tensors. Typical examples of such relations include the cross product, the dot product, and linear maps. Vectors and scalars are also individually considered as tensors. Tensors are normally represented using multi-dimensional array of numerical values. The order of a tensor is determined based on the number of dimensions needed to represent it in an array. Scalars are viewed as 0th-order tensors as they are single numbers. Vectors are viewed as 1st-order tensors as they can be represented using one dimensional arrays. A linear map is considered as a 2nd-order tensor as it can be represented by a 2-dimensional array. The array dimensionality should not be confused with the dimension of the underlying vector space while determining the tensor order.

2.2 Singular Value Decomposition (SVD)

The singular value decomposition (SVD) is a mathematical approach used for splitting a complex matrix into simple matrices by applying factorization. This approach is extensively used in data analytics and signal processing applications. The singular value decomposition of a complex matrix R is represented in factorization form as follows [1] [2] [8],

$R = U_1 \Sigma U_2^*$

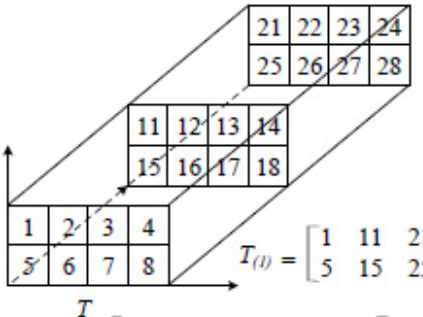
Where U_1 is an m × m real or complex unitary matrix, Σ is an m × n rectangular diagonal matrix with non-negative real numbers on the diagonal, and U_2^* (the conjugate transpose of U_2, or simply the transpose of U_2 if U_2 is real) is an n × n real or complex unitary matrix. The diagonal entries Σi,i of Σ are known as the singular values of R. The m columns of U_1 and the n columns of U_2 are called the left-singular vectors and right-singular vectors of R, respectively.

2.3 Tensor unfolding

Given a P-order tensor $T \in RI1 \times I2 \times \ldots \times IP$, the tensor unfolding [1]

$T(p) \in R^{Ip \times (Ip+1 Ip+2 \ldots Ip\ I1 I2 \ldots Ip-1)}$ contains the element $t_{i1 i2 \ldots ip ip+1 \ldots ip}$ at the position with row number ip and column number that is equal to $(i_{p+1} - 1) I_{p+2} \ldots IP\ I1 \ldots I_{p-1} + (i_{p+2} - 1)\ I_{p+3} \ldots IP\ I1 \ldots I_{p-1} + \ldots + (i_2 - 1)\ I3 I4 \ldots I_{p-1} + \ldots + i_{p-1}$.

Consider a three-order tensor $T \in R^{2x4x3}$, Fig. 2 shows the three unfolded matrices $T_{(1)}$, $T_{(2)}$ and $T_{(3)}$



$$T_{(1)} = \begin{bmatrix} 1 & 11 & 21 & 2 & 12 & 22 & 3 & 13 & 23 & 4 & 14 & 24 \\ 5 & 15 & 25 & 6 & 16 & 26 & 7 & 17 & 27 & 8 & 18 & 28 \end{bmatrix}$$

$$T_{(2)} = \begin{bmatrix} 1 & 5 & 11 & 15 & 21 & 25 \\ 2 & 6 & 12 & 16 & 22 & 26 \\ 3 & 7 & 13 & 17 & 23 & 27 \\ 4 & 8 & 14 & 18 & 24 & 28 \end{bmatrix} \quad T_{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 \\ 21 & 22 & 23 & 24 & 25 & 26 & 27 & 28 \end{bmatrix}$$

2.4 Tensor Order and Tensor Dimension

The tensor order and tensor dimension are two key elements used in dimensionality reduction methodologies. We should understand the difference between them before using them in dimensionality reduction. Tensor $T \in R^{I1 \times I2 \times \ldots \times IP}$ has P orders, and order i $(1 \leq i \leq P)$ has I_i dimensions. A P-order tensor can be unfolded to P matrices [1] [6] [8]. For the mode-i unfolded matrix $T(i)$, the number of rows is equal to I_i, while the number of columns is equal to $\prod_{1 \leq j \leq P;\, j \neq i} I_j$. It will be very difficult to establish relationship among the data records when the data contains redundant, repetitive and unwanted data. Hence it is essential to identify the core data from the input dataset after removing the redundancy and unwanted data. Normally the number of tensor orders remains the same while the dimensionality is significantly reduced during the dimensionality reduction process.

2.5 Unified Tensors

Big data are heterogeneous in nature and it is composed of unstructured data du, semi-structured data dsemi and structured data ds. Due to the requirement of processing all types of heterogeneous data, a unified data tensorization operation is performed using the following equation [1]

$$f : (d_u \cup d_{semi} \cup d_s) \rightarrow \underbrace{T_u \cup T_{semi} \cup T_s}_{T}$$

3. Incremental High Order Singular Value Decomposition

This paper propose an IHOSVD method for incremental dimensionality reduction on streaming data. The real-time data obtained from sensors placed in an agricultural field can be considered for this data reduction exercise [9]. The IHOSVD method consists of three algorithms that are used for recursive matrix singular value decomposition and incremental tensor decomposition [1]. The three algorithms are discussed below,

Algorithm 1 is a recursive algorithm with recursive function given in the below equation,

$$f(M_i, C_i) = \begin{cases} svd(M_1), & i = 1 \\ mix(f(M_{i-1}, C_{i-1}), C_i), & i > 1 \end{cases}$$

During the running process, function f will call itself over and over again to decompose matrices Mi and Ci. Each successive call reduces the size of matrix and moves closer to a solution until matrix M1 is reached finally, the recursion stops and the function can exit.

Algorithm 1: Recursive singular value decomposition on matrices, (U, ∑, V) = R −MSvd(Mi, Ci).

Input:

Initial matrix Mi.

Incremental matrix Ci.

Output:

Decomposition results U, S, V of matrix [Mi Ci].

Algorithm 2: Merging the incremental matrix with decomposition results of previous matrix, (U, ∑, V) = mix (Mi−1, Ci−1, Uj , ∑j , Vj).

Input:

Initial matrix Mi−1 and incremental matrix Ci−1.

Decomposition results Uj , ∑j , Vj of matrix M.

Output:

New decomposition results U, ∑, V.

Algorithm 3: Incremental singular value decomposition on tensors,

(S, [U, ∑, V] new) = I – T Svd (χ, T, [U, ∑, V] initial).

Input:

New tensor $\chi \in R^{I_1 \times I_2 \times \ldots \times I_P}$

Previous tensor $T \in R^{I_1 \times I_2 \times \ldots \times I_P}$

Previous unfolded matrices SVD results [U, ∑, V] initial.

Output:

New truncated SVD results [U, ∑, V] new.

New core tensor S.

The consolidated steps comprising all the three algorithms are listed below,

1. Extend tensor χ and tensor T to identical dimensionality.
2. Unfold new tensor χ to matrices χ(1), ..., χ(P).
3. Decompose the matrix χ(1) into U(1), S(1), V(1) by applying SVD
4. Execute step 10 to 14 recursively after incrementally passing matrices χ(2), ..., χ(P) together with decomposition results of previous matrix
5. Receive orthogonal bases U, S, V for the tensor χ
6. Truncate the new orthogonal bases.
7. Combine new tensor χ with initial tensor T.
8. Obtain new core tensor S with n-mode product
9. Return S, and [U, S V]new.
10. Project χ(2) on the orthogonal space spanned by U(1)
11. Compute matrix H which is orthogonal to U(1).
12. Obtain the unitary orthogonal basis J from matrix H
13. Determine the new decomposition results U(2), S(2), V(2) by applying SVD on [U(1) J]
14. Return the new decomposed results [U(2) S(2) V(2)] (Note: This decomposed results are created for consolidated matrices χ(1) & χ(2))

4. Conclusion

Big data and predictive analytics is expected to make a difference in the agricultural industry by providing on-time tips to the farmers based on real-time data gathered from different agricultural sources. Considering the volume and heterogeneous nature of future data, a well-defined mechanism is needed to reduce the memory and time involved in data analysis. Reducing the file size by extracting core value from massive big data is one of the approach to improve the data analysis performance.

References:

[1] Liwei Kuang,Fei Hao,Laurence T.Yang,Man Lin,Changqing Luo,and Geyong Min "A Tensor-Based Approach for Big Data Representation and Dimensionality Reduction"IEEE Transaction on Emerging Topics in Computing-2014

[2] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A Multilinear Singular Value Decomposition," SIAM Journal on Matrix Analysis and Applications, vol. 21, no. 4, pp. 1253–1278, 2000.

[3] I. Horrocks, P. F. Patel-Schneider, and F. Van Harmelen, "From SHIQ and RDF to OWL: The Making of a Web Ontology Language," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 1, no. 1, pp. 7–26, 2003.

[4] J. Sun, D. Tao, and C. Faloutsos, "Beyond Streams and Graphs: Dynamic Tensor Analysis," in Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006, pp. 374–383.

[5] Georg Rub, Rudolf Kruse, Martin Schneider, and Peter Wagner,"Data Mining with Neural Networks for Wheat Yield Prediction,"Springer-Verlag Berlin Heidelberg 2008

[6] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos, "Tag Recommendations Based on Tensor Dimensionality Reduction," in Proc. of the 2008 ACM Conference on Recommender Systems. ACM, 2008, pp. 43–50.

[7] Q. Li, X. Shi, and D. Schonfeld, "A General Framework for Robust HOSVD-Based Indexing and Retrieval with High-Order Tensor Data," in Proc. of the 36th IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2011, pp. 873–876

[8] M. Kim and K. S. Candan, "Approximate Tensor Decomposition within a Tensor-Relational Algebraic Framework," in Proc. of the 20th ACM International Conference on Information and Knowledge Management. ACM, 2011, pp. 1737–1742

[9] Hideya Ochiai, Member, IEEE, Hiroki Ishizuka, Yuya Kawakami, and Hiroshi Esaki, Member, IEEE"A DTN-Based Sensor Data Gathering for Agricultural Applications," IEEE sensors journal, vol. 11, no. 11, November 2011

[10] GopalaKrishna Moorthy .K, Dr.C.Yaashuwanth, Venkatesh.K, "A Wireless Remote Monitoring Of Agriculture Using Zigbee", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 8, February 2013

[11] Xindong Wu,Fellow,IEEE,Xingquan Zhu,Gong-Qing Wu and Wei Ding,Senior Member,IEEE "Data Mining with Big Data"-IEEE Transaction on Knowledge And Data Engineering -2014

[12] Lin Gu,Student Member,IEEE,Deze Zeng,Member,IEEE, and Song Guo,Senior Member,IEEE"Cost Minimization for Big Data Processing in Geo-Distributed Data Centers"IEEE Transactions on Emerging Topics in Computing-2014

[13] Chunxiao Jiang,Member,IEEE,Yan Chen,Member,IEEE,and K.J Ray Liu,Fellow,IEEE "Graphical Evolutionary Game for Information Diffusion Over Social Networks"-IEEE journal of Selected Topics in Signal Processing-2014

[14] Sangeeta Bansal, Dr. Ajay Rana, "Transitioning from Relational Databases to Big Data", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014