Computers and Electronics in Agriculture 120 (2016) 72–78

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Analyzing the behavior dynamics of grain price indexes using Tucker tensor decomposition and spatio-temporal trajectories

F.E. Correa [a,*], M.D.B. Oliveira [b], J. Gama [b], P.L.P. Corrêa [a], J. Rady [a]

[a] Department of Computer Engineering – Polytechnic School – University of São Paulo, PO Box 61548, São Paulo, SP 05424-970, Brazil
[b] Laboratory of Artificial Intelligence and Decision Support – INESC TEC, University of Porto, R. Doutor Roberto Frias 378, 4200-465 Porto, Portugal

ARTICLE INFO

Article history:
Received 13 October 2014
Received in revised form 16 November 2015
Accepted 22 November 2015
Available online 7 December 2015

Keywords:
Data mining
Tucker decomposition
Spatio-temporal trajectories
Grain market

ABSTRACT

Agribusiness is an activity that generates huge amounts of temporal data. There are research centers that collect, store and create indexes of agricultural activities, providing multidimensional time series composed by years of data. In this paper, we are interested in studying the behavior of these time series, especially in what regards the evolution of agricultural price indexes over the years. We explore data mining techniques tailored to analyze temporal data, aiming to generate spatio-temporal trajectories of grains price indexes for six years of data. We propose the use of Tucker decomposition to both analyze the temporal patterns of these price indexes and map trajectories that represent their behavior over time in a concise and representative low-dimensional subspace. The case study presents an application of this methodology to real databases of price indexes of corn and soybeans in Brazil and the United States.

1. Introduction

The agricultural commodities are very important to economies of several countries, especially Brazil, where these assets account for 7.3% of the Gross National Product – GNP. Moreover, agricultural activities are the backbone of most economic systems, in the sense they represent an important source of raw materials to other industries (e.g., cotton, sugar) and provide many employment opportunities for the labor force (IBGE).

Two of the most important agricultural products in the Brazilian economy are corn and soybean. These products belong to the family of grains. For example, Brazil exportation of soybeans grains was 32 M. tons, representing an estimated figure of 17.5 billion dollars in 2012 (ALICEWEB).

Despite the great amount of money involved, we do not have, in agribusiness activities, accurate information for the whole process. Therefore, research centers in Brazil, such as the Center for Advanced Studies on Applied Economics – CEPEA, collect and provide price indexes of these commodities (Correa, 2009).

Studies to understand the temporal trajectory of a variable, such as prices for products like soybean and corn, provide the market players with strategic information regarding the international market transaction behavior over the last years. Considering that the models were applied on real data, it is possible to update these models with new collected data and use them to infer or predict if some events will continue to happen. Moreover, we can mention some other benefits that arise from the international market analysis. For example, it is possible to observe that the trajectory of the Chicago stock market prices is the base for price indexes over the Brazil internal prices. As a result, further research could try to measure what is the impact of some public policies, e.g. American incentives for corn producers, by adding such information on new models and simulations (Aruga, 2014; Rosa et al., 2014).

In order to analyze agro economic data, it is necessary to join several databases of distinct types and subjects (Plant, 2012). Databases with this kind of information are usually multidimensional, i.e., they have more than two dimensions (variables that affect their behavior).

Examples of multidimensional data are common in agriculture. Usually, the products are negotiated in different types of markets, e.g., domestic market and stock market. Further, there are a variety of products, like corn, soybeans grain and meal. Moreover, these multidimensional data are temporally ordered, i.e., they are time series collected and stored over several years (King, 2010).

There are data mining techniques able to deal with multidimensional and temporal data. In this research, our aim is to explore two of those techniques in order to provide a methodology for the analysis of agro economic data. The main techniques and framework used were Tucker decomposition (Tucker, 1966; Oliveira and Gama, 2012) and spatio-temporal trajectories (Oliveira and

* Corresponding author.
E-mail addresses: fecorrea@usp.br (F.E. Correa), mdbo@inescporto.pt (M.D.B. Oliveira), jgama@fep.up.pt (J. Gama), pedro.correa@poli.usp.br (P.L.P. Corrêa), jorge.rady@usp.br (J. Rady).

http://dx.doi.org/10.1016/j.compag.2015.11.011
0168-1699/

Gama, 2013). Some complementary statistical methods were used, namely, the correlation matrix (Kazmier, 2004), ANOVA and the sliding window model (Datar et al., 2002). In addition, clusters analysis for pattern mining applied on spatio-temporal data were reviewed to consider by future implementation (Patel, 2005; Xiao, 2014).

Using a methodology based on the aforementioned data mining techniques, the idea is to understand the evolution of the time series of Grains price indexes over a time span of six years. In the case study presented it was used a real-world time series. Our main contribution is to propose a process methodology to identify, summarize and highlight past events and provide analysis methods to deal with multidimensional datasets.

This paper is organized as follows: In Section 2, we introduce the main concepts about the Tucker decomposition and data mining techniques for analyzing temporal data. After providing the background, we detail the proposed methodology in Section 3. The next section presents the application of the proposed methodology to multidimensional time series of grains price indexes. This paper ends with the related work, conclusions and suggestions for further research.

2. Tucker decomposition

Tucker decomposition is an unsupervised multiway data analysis method that is quite useful for data cleaning, data compression and visualization of the main structures of data in low-dimensional spaces. Tucker (1966) devised this method in order to extend the well-known PCA (Principal Component Analysis) to higher-order data representations, such as tensors. We can straightforward define a tensor as an extension of a matrix to three or more dimensions, or as an N-way data array, where N is the *order* of the tensor. The Data Mart analyzed in this paper can be arranged into a three-order tensor, by incorporating the temporal dimension. We resort to three-order tensors, instead of matrices, in order to explicitly account for the time dimension and, thus, avoid loss of information in the modeling process. The *order*, *ways* or *modes* of a tensor are synonyms and refer to the number of dimensions (in our case, we have three dimensions: products, market and time). For this specific type of tensors or, in other words, N-way data arrays for $N = 3$, the most appropriate Tucker decomposition model is the so-called *Tucker3 tensor decomposition* (Kolda and Bader, 2009), which performs the reduction of data in all three modes of the tensor.

The basic idea of the Tucker3 decomposition is to find a set of matrices (known as the *component matrices*) and a small tensor (known as the *core tensor*) that, in general, have less dimensionality than the original tensor, but are able to reconstruct the most important information contained in data.

The Tucker3 model can be formulated as the factorization of the original three-order tensor χ, such that

$$\chi_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr} a_{ip} b_{jq} c_{kr}$$

for $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$. Here, the coefficients a_{ip}, b_{jq} and c_{kr} represent the entries of the *component matrices* $\boldsymbol{A} \in \mathbb{R}^{I \times P}$, $\boldsymbol{B} \in \mathbb{R}^{J \times Q}$ and $\boldsymbol{C} \in \mathbb{R}^{K \times R}$. In turn, the coefficient g_{pqr} represents the entry of the so-called *core tensor* $\mathcal{G} \in \mathbb{R}^{P \times Q \times R}$. The number of entities in each mode are represented by letters I, J and K. The number of components (i.e., the number of columns of the matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$) in the first, second and third mode of the tensor are represented by letters P, Q and R, respectively. This decomposition is illustrated in Fig. 1.

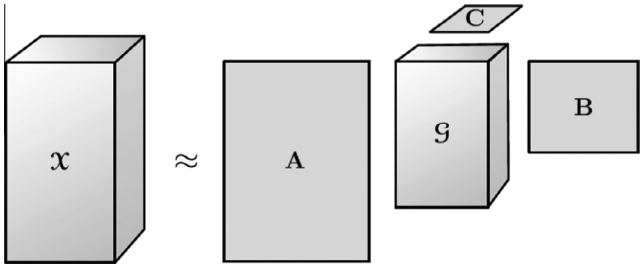

Fig. 1. The basic Tucker3 tensor decomposition (Kolda and Bader, 2009).

Tucker suggested that the *core tensor* $\mathcal{G}$ can be interpreted as describing the latent structure in the data and the *component matrices* ($\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$) as mapping this structure to give the observed data (Tucker, 1966). The *core tensor* can also be interpreted as a generalization of the eigenvalues of the SVD (Singular Value Decomposition) (Skillicorn, 2007). Detailed information about the Tucker3 technique can be found in Tucker (1966), Kolda and Bader (2009) and Kiers and Mechelen (2001).

This technique can be used to detect abnormal events and important milestones in the agribusiness data, by means of the projection of spatio-temporal trajectories in Tucker3 bi-dimensional subspaces. Spatio-temporal trajectories visually represent the movement of a given object in a plane. They can be formally defined as a function from the temporal dimension $I \subseteq \mathbb{R}$ to the geographical space $\mathbb{R}^2$ (*i.e.*, the 2D, or bi-dimensional, space) (Kiers and Mechelen, 2001). At each time point, the object occupies a given position in the 2D space. Each position is recorded in terms of (x, y) coordinates, which represent latent concepts, and associated with the corresponding time stamp. The temporally ordered sequence of an object's positions defines the trajectory of this object, which are often represented as (x, y, t) triples:

$$T = \{(x_1, y_1, t_1), (x_2, y_2, t_2), \ldots, (x_k, y_k, t_k)\}$$
$$\text{where } x_i, y_i, t_i \in \mathbb{R}\ (i = 1, \ldots, k) \text{ and } t_1 < t_2 < \ldots < t_k$$

These trajectories are graphically represented by a line that connects the coordinates of each position to the object's movement. The goal toward the use of spatio-temporal trajectories is the representation of time series in a way that is efficient to analyze. The analysis of trajectories allows us not only to understand the dynamics of an object's behavior (*e.g.*, the evolution of corn indexes with respect to a set of agro economic indicators) but also to understand large quantities of information in a concise way.

3. Methodology

3.1. Grain dataset specification

In order to make possible the analysis of the multidimensional databases, we had to do some procedures on the data, in order to retrieve time-ordered data in a format that can be used to generate the spatio-temporal trajectories. The multidimensional dataset was split into 3 dimensions, or *modes*. One of these dimensions is time. The time dimension can have several granularities. In our case, the time unit selected is the month. To obtain monthly data from years 2007 to 2012, six datasets were created, one for each year. Considering this division it was possible to label the trajectories' variables per year in the plot.

Each dataset that results from the application of the aforementioned technique will be called a *data cube* (i.e., a three-order tensor). The data cube used in this paper has 3 dimensions or modes, namely: products, market and time. The entities that belong to the products dimension are the collected price indexes