APPLYING BIG DATA ALGORITHMS FOR SALES DATA STORED IN SAP HANA

Judit Oláh, Edina Erdei, József Popp
University of Debrecen, Faculty of Economics and Business, Debrecen, Hungary
olah.judit@econ.unideb.hu
edina.erdei92@gmail.com
popp.jozsef@econ.unideb.hu

Abstract: *An increasing number of small and medium enterprises operate and supervise their financial, logistics, production, human resource and other activities with ERP (Enterprise Resource Planning) IT systems, which are capable of managing these processes in a uniform framework. It is our research objective to explore analytical and prediction methods which can provide solutions to various financial and logistics problems that are faced by nowadays' enterprises. In order to achieve this objective, we analysed a real database of an enterprise using SAP. This database covers the period between January 2013 – October 2016 and contains 7 million sales records of 15,674 different products. These products also include currently inactive items, however, they can still be important from the aspect of data analysis in relation to the examined period. The data structure was created using the recently introduced SAP HANA (High Performance Analytic Appliance) database management system which revolutionised data storage with its in-memory and column-oriented features. This new technology makes it possible to execute various transactions more effectively and more quickly. Following the preparation of the schematics of data needed for processing and the calculation of calculation fields, we used the tools provided by SAP Predictive Analytics which was introduced to the market in 2015. After filtering sales data for 15 quarters, we used k-means clustering for each period. After preparing and examining the clusters, we made observations which make it easier to perform stock management, logistics and pricing activities in the future, thereby contributing to the long-term increase of enterprise profit. Clustering was also performed in R programming language, which enabled us to illustrate the clustering results, i.e., each sales record in 3D, colouring them based on their associated cluster label. After inspecting these graphic outputs, it was concluded that certain products should be withdrawn from the market, while others should be either developed or their stock level increased. We used regression analysis on the cluster centroids to predict the movement of each cluster mainly in terms of time. As a result, we provided an estimation as to the direction that the products belonging to the sales records in each cluster will fluctuate to in accordance with the coordinates of the cluster centroids, thereby making recommendations related to the management of each item and group of items. As a next step, we performed a global F test for regression analysis to examine the correctness of our model. As a conclusion, we reject the null hypothesis which stated that our model is basically invalid.*

Keywords: *SAP HANA; SAP Predictive Analytics; R programming language; k-means clustering; regression analysis.*

JEL classification: *F47; M15; M49.*

1. Introduction

The economic success of companies and enterprises is based on how efficiently and successfully they can make use of their available resources in order to produce their products or services (Fenyves – Dajnoki, 2015). Nowadays, an increasing number of small and medium enterprises operate and supervise their financial, logistics, production, human resources and other activities with Enterprise Resource Planning (ERP) IT systems which are able to manage these processes in a uniform framework. SAP Business One is one of the main ERPs of SAP AG (System, Applications & Products in Data Processing) and it is tailored to help small and medium enterprises manage their business processes in a more simple way (Niefert, 2009).

The memory-based and column-oriented HANA is a relational database management system developed and distributed by SAP and it was introduced to the market in 2010. In the case of SMEs, HANA runs on an independent B1A (Business One Analytics) server, on which the whole SAP Business One database is loaded into the memory where queries are performed subsequently (Walker, 2013).

By using the in-memory technology, data become accessible without any difficulties (e.g. lack of capacity) in just seconds; therefore, the hard and time-consuming tasks of pre-processing and grouping the data is not necessary. “Hot data”, i.e., data that is needed to be accessed frequently, is entered into the memory (Plattner, 2012).

The column-oriented construction better fits analyses in which algorithmic operations (organising, sum, filtering, mean) are often necessary to be performed on homogeneous data in columns. As a result, it is not necessary to load the whole table during analysis and indexes are not needed to be modified either when inserting a new row. By using this system, reports which used to take even more than ten minutes to perform can now be compiled in only seconds without overburdening the administration system (Idreos et al., 2012).

It is possible to connect to HANA even with mobile devices, enabling more effective work, eventually leading to a positive impact on competitiveness. Therefore, managers are able to make decisions based on the most up-to-date information any time of the day (Silvia et al., 2015).

On-line Analytical Processing makes it possible for database management systems to enable its users to query important data very quickly (Abdullah, 2009). OLAP is a database technology optimised for using queries and reports instead of executing transactions (Boutkhoum et al., 2015). OLAP technology is described by more complex queries, commands which include more complex calculations, as well as larger amounts of data. The system stores the results of various pre-calculations and partial operations in advance (Thomas et al, 2001).

Implementing ERPs potentially leads to increased sales efficiency and improved interactions in warranty services results in customer satisfaction through providing lower quotations and increased awareness. Enhanced production and lower cost of inventories also improve the performance of organisations (Máté-Kárpáti, 2015).

It is the fundamental operation of OLAP to create the data cube. Other typical OLAP operations include slicing, dicing, roll-up, drill down and pivoting (Ceci et al., 2013).

SAP Predictive Analytics involves several analytical functions. The 36 different algorithms involve clustering, involving the K means clustering used in this research. In addition, naive Bayes and kNN classification techniques are also focused on.

Furthermore, the system makes it possible to search for outliers, to examine the correlation between data and to analyse time series (SAP SE, 2014).

2. Sample and methods

The enterprise deals with food trade and its partners include various hypermarkets, food industry chains and restaurants. The 2015 central financial report of the enterprise could have been more favourable if the previous data are properly analysed and the sales of the substitute and supplementary product is better performed. *Table 1* contains the most important enterprise-related information from the aspect of analyses.

Table 1: Data of the examined enterprise

Characteristic	Value
Number of sales records	7 052 888
Beginning of the examined period	01/01/2013
End of the examined period	01/10/2016
Number of sold products	15 674
Number of currently active products	8 392

Source: Own research, 2017

The dataset containing sales also includes significant transaction-related attributes. In order to perform the proper analyses, reports and predictions, the examination should be narrowed down to the following attributes: card code, item code, date of sales (calculated field: summarised per month), price gap, purchase price, quantity and line total.

Our analytical methods included k means clustering and density-based spatial clustering of applications with noise (DBSCAN). These clustering methods helped us find the correlations between sales data and items and we used regression calculation to predict (depending on time or other variables) the coordinates of the centroids of clusters and homogeneous groups identified this way. Regression calculation serves the purpose of modelling a correlation between two or more random variables. As a next step, a global F test was performed to analyse the goodness of our model. The performed analyses enabled us to decide whether the lines and polynomials created during regression analysis properly fit to the cluster centroids and, consequently, whether the examined variables altogether describe the dependent variable sufficiently.

The SAP HANA technology was used for data preparation. Both in the case of the used clustering methods and during regression calculation, we used the operators of Predictive Analytics, as well as the statistical programming language "*R*".

3. Results

3.1. Results of clustering

The SAP Predictive Analytics makes it possible to perform filtering by column and row. This way, the dataset was divided to quarterly periods. The following attributes

were selected during the parametrisation of clustering: *quantity, net row total, purchase price and price gap*. As a next step, we ignored the missing values in the input data and set the maximum number of iterations to 100. In order to speed up the calculation process, we ran two types of analyses simultaneously, based on the available technology.

Clustering was tested with $k \in [1, 10]$ values and 5 clusters were created for each period, as the other parameters either resulted in excessively heterogeneous groups or the high number of clusters would have been difficult to interpret. The initial centroids of clusters is a difficult question, but observations show that the best result can be obtained if initial centroids are determined randomly. In addition, Predictive Analytics provides various options as to how the new coordinates of centroids can be calculated during the clustering process.

The results of the clustering process referring to the last quarter (01/07/2016 – 01/10/2016) show that the table was extended with two new columns: the number of cluster which represents which of the five clusters the given record belongs to and distance which shows the Euclidean distance between the given row and the centroid of its own cluster.

After finishing the algorithm, clusters can be classified into groups such as very good, less good, neither/nor, less bad, very bad. The first cluster contains the least number of items and the price gap coordinate of its centroid is the lowest (*Figure 1*). Therefore, it can be concluded that there were few sales in which purchase price was higher than selling price. For example, this phenomenon can be observed in the case of products with low shelf life, as it may occur that an enterprise does not pay proper attention to the appropriate management of products whose shelf life is shorter than a few months. There are relatively few products in the cluster with the highest price gap centroid coordinate (8978.69) (*Figure 1*), which relates to the fact that relatively few products were sold with high profit.

Algorithm Summary

Summary:

Overview

Model Building Date : 10/03/17 18:34 PM
Number of clusters :5

The size of each cluster:

Cluster1	:21
Cluster2	:30
Cluster3	:49908
Cluster4	:1697
Cluster5	:130863

Sum of all clusters :182519

Cluster Centers

		Arres
1	:	-628.047619047618
2	:	8978.699538795805
3	:	265.71665871890696
4	:	1024.8347494683708
5	:	68.933937853437186

Figure 1: Clustering algorithm summary
Source: Own construction, 2017.

In addition to the operators of the SAP Predictive Analytics, clusterings were also performed in *R*, which separated the quarterly data relatively successfully by presenting them in a 3D cube (*Figure 3*). It is recommended to revise the items belonging to the records in the cluster on the left with very bad centroid coordinate and they should be withdrawn from the product range of the enterprise due to their high variability and low net row total. It is suggested for the enterprise to pay increased attention to the items belonging in the groups in the middle of the cube (less bad, neither/nor, less good), since these provide relatively stable trade and income performance, while they can easily represent the elements of the very good cluster in the future, as this group shows the most successful sales data of the enterprise.

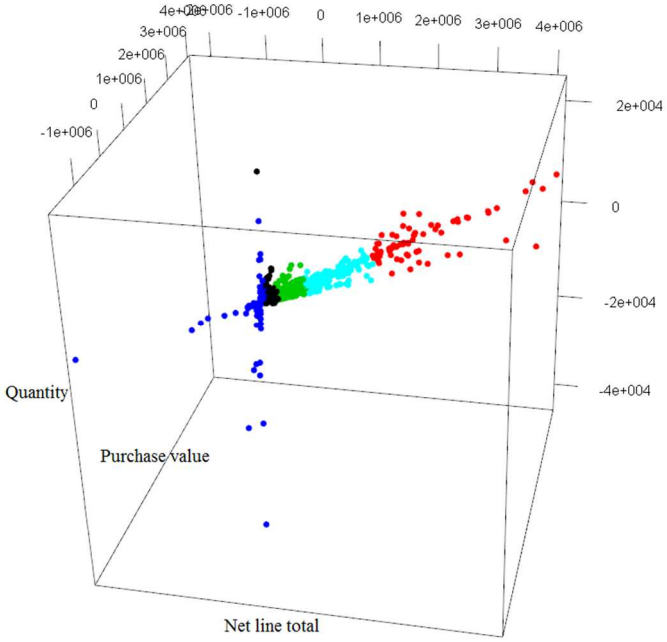


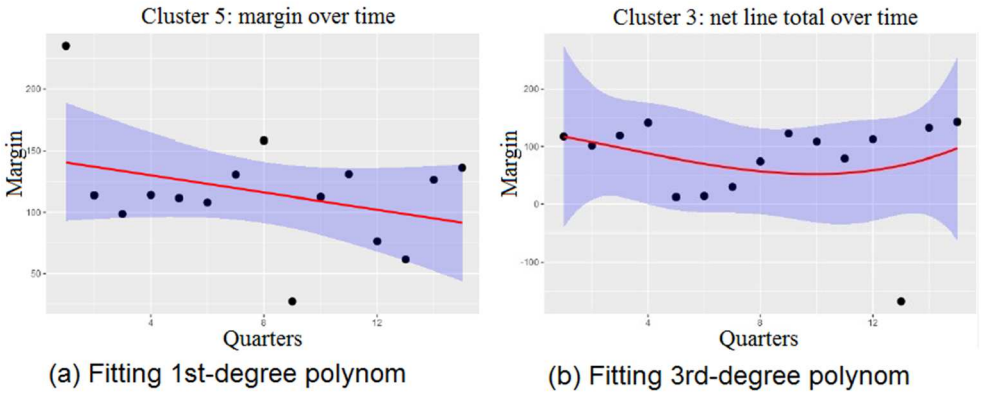
Figure 3: Clustering results in a 3D graph
Source: Own construction, 2017.

After the clustering process, we also performed the analyses of substitute products in the case of each item group, since one cannot necessarily recommend a product from a different item group instead of a given product. For example, it is not an option to recommend customers to buy products classified as dry bakery products instead of an item belonging to the group of chilled pork.
We analysed the item groups of clusters belonging to different categories based on two price gaps in the given period and concluded that the examined enterprise could recommend an item belonging to a cluster with a “very good” price gap centroid instead of products in the cluster with “less good” price gap centroid. Consequently, the enterprise may increase its profit based on items with higher price gap. For example, one may recommend customers to buy “chilled fresh chicken breast fillet” instead of “fresh chicken breast”. The former product has a higher price gap, while

the latter one has a lower price gap. The examined enterprise should withdraw its products with price gaps belonging to the "very bad" category, since their storage and removal may incur considerable costs.
In our research, we mostly used the k means algorithm, but we also tested other methods, the most important of which was DBSCAN. This clustering method is based on density, but the number of created clusters is not determined in advance; therefore, the variation of two parameters (epsilon (*eps*) and minimum points (*minPts*)) needed for the definition of density caused a problem related to the quantity of the resulting clusters. The number of resulting clusters was around 50-100. This method found a significant amount of sales records to be outliers. Finally, this procedure needed significantly more resources than the k means procedure, since it used around 9 GB memory in the case of *eps*=1000 and *minPts*=10.

3.2. Results of predictions

Clusters were compared to each other not only within a quarter, but also between different quarters in order to examine the change of clusters in a time series.
"Much" data is needed to draw statistically substantiated conclusions; therefore, we used the quarterly clustering results of nearly four years. Furthermore, it is very important to decide how to determine the associated cluster in each quarter. We strived to select a refined method; therefore, we decided to use the Magyar method. This approach called for nothing else than to calculate the distance of the cluster centroids of the two current quarters from each other, which means that the 5-5 clusters will result in a $D \in \mathbb{R}^{5\times5}$matrix. The Magyar method was used for the matrix obtained this way. The output of this method described how to pair the clusters of the previous and the subsequent quarter. Formally: Magyar method $(D) \rightarrow M$, where the pair of $M \in \mathbb{N}^{5\times2}$ and $\forall\{i,j\} \in M$ is the pairing of the elements of each cluster set (e.g. if *i=3 and j=4*, the 4th cluster of the subsequent quarterly analyses is associated as the 3rd cluster of the previous quarter). When used in each quarter, this procedure is an optimum solution of the problem, resulting in time series comprised of the people matching each other the most.
The examples below demonstrate the results which were obtained with regression analysis performed for the time series created using the above described method. Due to the strong fluctuation of cluster centroids, the linear regression could properly represent the correlation between time and the given coordinate only rarely; therefore, the given time series could be modelled more realistically using a third degree polynomial.
Figure 4 shows the results of the regression calculation.
The squared error of lines based on a few data points is relatively high, as it is very difficult to establish connection between two variables using such a simple model. *Figure 4a* shows the price gap coordinate of the 5th cluster against time, which, according to the model, shows a decreasing trend, while the net row total increases (*Figure 4b*), as it was expected. The enterprise should develop a better pricing strategy for the products belonging to the sales records of this cluster, as the profit realised on this product will be minimised in a few quarters. The reduction of the price gap can be explained by the more significant increase of purchase price and partially the fluctuation of products leaving a cluster and entering another. For this reason, the enterprise should withdraw all riskier products in this group and they should stabilise the less fluctuating ones.



(a) Fitting 1st-degree polynom (b) Fitting 3rd-degree polynom

Figure 4: Results of the regression model
Source: Own research, 2017

It is important to make the results of predictions, i.e., the fitted polynomials fit the best possible way to the time series developed from the centroid coordinates of clusters. In order to test fitting, we also performed the well-known F test for each polynomial and observed whether the obtained parameters are good, i.e., whether the respective variables are good independent variables in the regression model. Based on the analysis of a third degree polynomial fitting (*Figure 4b*), it can be concluded that the polynomial fits the centroid coordinates of clusters better, due to the lower outliers. The results of the global F test performed for the model shown in *Figure 4b* are presented in *Figure 5*.

	Sum of Squares	df	Mean Square	F	Sig.
Regression	6,778E+13	3	2,259E+13	2,992	,077
Residual	8,306E+13	11	7,551E+12		
Total	1,508E+14	14			

Figure 5: ANOVA results for the net value coordinate of the 3rd cluster against time
Source: Own research, 2017

Consequently, we reject the null hypothesis of the F test at a significance level of 10% that is the model is wrong as a whole. Therefore, we managed to find a regression polynomial which explains the net row total coordinate of the 3rd cluster at the proper level of significance, depending on time.
Furthermore, we examined the sales quantity of the items of two different subsequent periods and concluded that the sales transaction of each item increases. As a result, the examined enterprise is recommended to invest in a larger stock at a promotional price.

4. Conclusions and recommendations

We created a dataset based on a novel technology related to SAP HANA and analysed the sales records. After the clustering process, we identified substitute products within each item group of the enterprise. If the enterprise pays more attention to the correlation between products of this kind and they focus more intensively on products with higher price gap in their strategy, they can realise higher income in a short term. In addition to substitute products, there were also seasonal products in the clusters which were characterised by the various sales quantities of each period. The enterprise can prepare for such fluctuations by purchasing a large quantity of these items before the increasing period at a promotional price. Furthermore, by classifying each cluster in a separate category, there were also groups with "very bad" price gaps. The items in these clusters or products with low shelf life are the so-called "frozen" stocks. The enterprise should pay more attention to the coordination of these items, i.e., they are recommended to use "warning procedures" using the SAP Business One system.

The examined enterprise can develop a marketing strategy related to the top 10 customers which maintains and even improves their purchasing habits. A possible example of how this goal can be reached is to provide customers who purchased the most in the first two weeks of the month with a previously determined discount from the net total; therefore, customers purchase increasing amounts of products.

Clustering was also performed with the programming language "R"; therefore, the clustering results and each sales record, coloured on the basis of the cluster label, were depicted also in 3D form. As regards these graphic outputs, it was concluded that there are products which are recommended to be withdrawn from the market, while others should be developed or the minimum level of their stocks should be increased.

During regression analysis, the movement of each cluster was predicted mainly depending on time and both with using simple regression techniques and more complex models. It was concluded that the opportunities in the presented technologies and data mining algorithms have an infinite storehouse. Although both data storage technologies and the science of data mining underwent significant development during the past years, it is still not possible to use these methods at the desired safety and accuracy level among small and medium enterprises.

It was concluded that if these recommendations are accepted, the enterprise can proceed with its existing successful competition strategy in parallel with the constant development of technology, with special regard to its stockpiling, logistics and forwarding resources.

References

1. Abdullah, A. (2009): *Analysis of mealybug incidence on the cotton crop using ADSS-OLAP Online Analytical Processing tool. Computers and Electronics in Agriculture*, Amsterdam: Elsevier, pp. 59-72.

2. Boutkhoum, O., Hanine, M., Tikniouine, A. and Agouti, T. (2015): *Multi-criteria Decisional Approach of the OLAP Analysis by Fuzzy Logic: Green Logistics as a Case Study,* Arabian Journal for Science and Engineering, Vol. 40., No. 8, pp. 2345-2359.

3. Ceci, M., Cuzzocrea, A. and Malerba, D. (2015): *Effectively and efficiently supporting roll-up and drill-down OLAP operations over continuous dimensions via hierarchical clustering.* Journal of Intelligent Information Systems, Vol. 44, No. 3, pp. 309-333.
4. Fenyves V. and Dajnoki K. (2015): Controlling opportunities in area of the human resources management. Analele Universitatii Din Oradea Fasciola Management Si Inginerie Tehnologica / Annals Of The University Of Oradea Fascicle Of Management And Technological Engineering. Vol. 24, No. 1, pp. 137-142.
5. Idreos, S., Groffen, F., Nes, N., Manegold, S., Mullender, S. and Kersten, M. (2012): *MonetDB: Two Decades of Research in Column-oriented Database Architectures*, IEEE Data Engineering Bulletin, Vol. 35, No. 1, pp. 40-45.
6. Máté, D. and Kárpáti T. J. 2015: How can Enterprise Resource Planning (ERP) Systems Impact on Labour Productivity, Annals Of The University Of Oradea Economic Science XXIV:(2) pp. 623-626. (2015)
7. Niefert, W. (2009): *SAP Business One Implementation.* Pack Publishing Ltd., Birmingham
8. Plattner, H. (2009): *A Common Database Approach for OLTP and OLAP Using an In-Memory Column Database.* University of Potsdam.
9. SAP SE (2016): *SAP HANA Predictive Analysis Library (PAL).* SAP affiliate company, Waldorf, Document Version: 1.1
10. Silvia, P., Frye, R. and Berg. B. (2015): *SAP HANA An Introduction.* Pack Publishing Ltd., Birmingham
11. Thomas, H. and Datta, A. (2001): *A Conceptual Model and Algebra for On-Line Analytical Processing in Decision Support Databases.* Information Systems Research, Vol. 12, No. 1, March, pp. 83-102.
12. Walker, M. (2013): *Software Development on the SAP HANA Platform.* Pack Publishing Ltd., Birmingham