Analysis and application of data reduction methods in decision support systems

Artsiom Maroz
Faculty of Applied Mathematics and Computer Science
Belarussian State University
Minsk, Belarus
marozAS@bsu.by

Alexander Valvachev
Faculty of Applied Mathematics and Computer Science
Belarussian State University
Minsk, Belarus
van_955@mail.ru

Abstract. **The paper addresses the problem of rapid growth of corporate data, which leads to a delay in decision-making and a decrease in their quality. A new approach to decision-making based on big data reduction is proposed, which reduces the time for synthesizing solutions while maintaining their quality and reducing the cost of processing.**

Keywords: **data reduction, principal component analysis, decision support system, big data, small data**

I. INTRODUCTION

The globalization of business has led to an avalanche-like increase in the amount of data needed for decision-making at all levels of distributed organizations . As a result, data flows from the lower levels of the hierarchy to the center began to overload the communication channels, which led to a delay in control decisions, a decrease in their quality and the loss of competitive advantages[1, 2]. Big data is a collection of large-scale, voluminous and multi-format data streams from autonomous data sources. The huge volume of big data causes heterogeneity of data and different dimensions of data sets. Therefore several approaches are used to solve problems with big data, including graph theory, reducing the dimension of big data, eliminating redundancy, data mining, neural networks [3]. The practical application of these approaches is complicated by the specific properties of distributed systems: this is due to the fact that at each level, different types and data structures, original software and reporting forms are used to support their own processes. Moreover, existing decision support systems are often focused on planned tasks, relatively stable global conditions and small amounts of information consumed.

Therefore rigorous “small data” paradigm that functions autonomously and collaboratively with big data analytics is also needed. By “small data“ we build on Estrin’s formulation and refer to the rigorous use of data collected to advance the goals of the specific N-of-1 unit for whom the data are about (i.e., a single person, clinic, hospital, healthcare system, community, city) [4, 5]. The purpose of using small is to achieve improved description, forecasting and control at the level of a specific unit. As part of this, the unit itself plays a role in determining the goals of data analysis. Thus, the transition from big data to small data allows you to get away from the general description of the system and come to a specific solution at each level.

The approaches listed above contribute to the elimination of delays in decision-making, as well as a full analysis of all available data, regardless of their volume and dimension, since all the information received is normalized.

II. PROBLEM ANALYSIS AND PROBLEM STATEMENT

This paper examines the problems of improving the quality of strategic management solutions based on corporate data in large-scale companies and distributed systems. The general scheme of data flow in such systems is shown in Fig. 1.

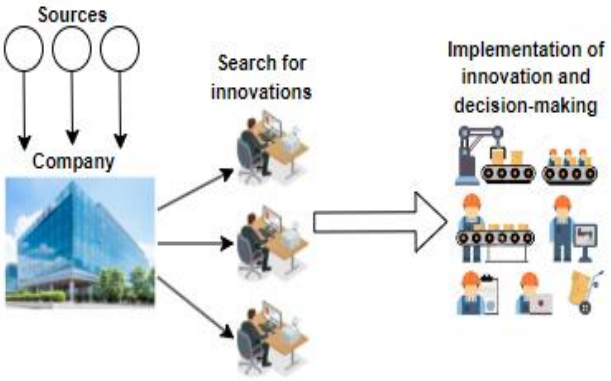

Fig.1 Data flow in distributed system

The main problems include: undeveloped means of solving operational tasks, overflow of communication channels when transmitting data from branches, violations of the data delivery schedule, violations of the schedule for synthesizing control decisions in the Center, delay in making operational decisions and their implementation. As a result, in this system, the amount of data for decision-making (A) and decision-making time (T) tends to infinity, and the quality and quantity of decisions (Q) made tends to 0: $(A \to \infty \; and \; T \to \infty), Q \to 0$. Our main goal is to change this trend and move to the scheme $(A \to 0 \; and \; T \to 0), Q \to \infty$. In other words, the amount of data required to obtain a relevant solution should be minimal. The time spent on making a decision is also minimal, and the quality of the decision is maximum.

To solve this problem, we will use modifications of the principal component analysis (PCA) method together with the feature selection. This will create a fundamental approach to extracting valuable information from multidimensional data, will allow you to identify the necessary features and reduce the dimensionality of the data. This approach is integrated into the company's processes to improve efficiency, accuracy and interpretability in decision-making.

First step is using feature selection techniques to identify the most informative features. Techniques like recursive feature elimination or correlation analysis helps identify the most relevant and informative features while PCA further reduces the dimensionality of the remaining dataset. Using this statistical method, we transform the input data set consisting of correlated variables into a new set of uncorrelated variables. These components are linear combinations of the original variables arranged in descending order of variance. The first main component captures the maximum variance, followed by subsequent components with decreasing variance. PCA aims to minimize information loss

while reducing the dimensionality of data sets. The principle of operation is shown in Fig. 2.

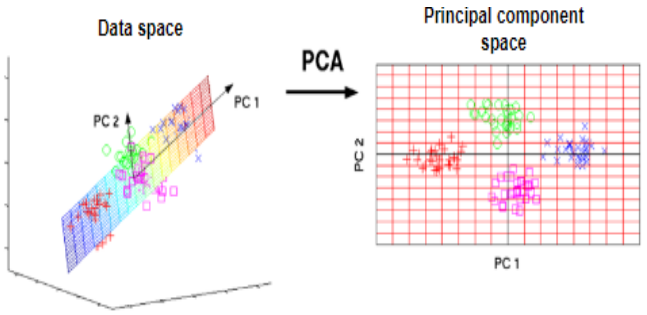

Fig.2 Workflow for PCA

When projecting data onto a new space of principal components, we save as much variance as possible, because this allows us to store more information about the data in a smaller number of components [6]. Understanding and managing variance helps to select the right number of principal components that carry the largest proportion of variance and, therefore, the most informative aspects of the source data to preserve. It is also necessary to standardize variables if they are at different scales. Dividing the values of each variable by the standard deviation of the corresponding variable is enough to eliminate the scale effect, and at the same time each variable will be centered [7].

It is evident that using this approach will minimize the amount of data that will be used for decision-making. It is also important to emphasize the use of a combination of feature selection and PCA. It is this variant that will lead to the best use of the input dataset and the transition from big data to small data.

III. The Process of Solving the Problem

Let's build a general schema of a distributed multi-level company. It is shown in Fig. 3.

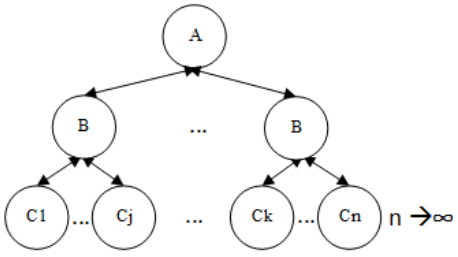

Fig.3 General schema of distributed multi-level company

The methodological approach and the software should be invariant to the number of objects at the third level. As a result, it is necessary to receive and process the data obtained at lower levels, make a decision in center A and distribute it to low levels B and C. Based on the previous provisions and the description of the problem. We have the following: the number of divisions is constantly increasing ($n \to \infty$). Accordingly, the amount of data ($vX \to \infty$) required for the synthesis of solutions (U) and the synthesis time (tU→∞) increases.

Based on this, we will build a decision-making model in a large-scale company. It is shown in Fig. 4.

In the center, during time t, the states of the divisions $S = f(vX)$ are determined and control solutions $U = g(S)$ are

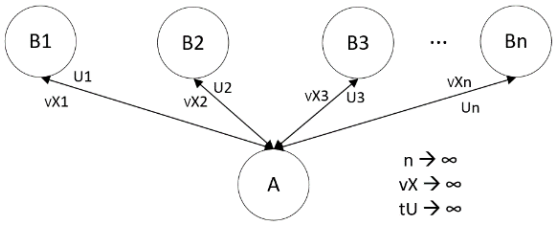

Fig.4 Decision-making model in large-scale company

synthesized. As we see the disadvantage of this scheme is that an increase in the number of branches (n) and therefore data (vX), entails an increase in decision-making time (tU), which will negatively affect the introduction of innovations. Let's build models that will help reduce the amount of data received and decision-making time, while improving their quality.

We will represent the distributed company as a whole as a tuple: $DC = (A, B, C, com)$. A, B, C – actors of the highest, middle and lower levels; com – communications between them.

Describe the models of actors according to a unified scheme: $mAct = (xAct, uxAct\ rxAct, com)$.

$xAct$ – data from sensors, $rxAct$ – resources to ensure the implementation of solutions, $uxAct$ – the control decision that is made in relation to this level com – communication channels for communication between factors of different levels. Thus, it is enough to apply a function for data reduction (f), which will determine the state ($sxAct$) of the actor at any level. In the future, one control solution from the set is put in accordance with this state. The state model is described as follows: $sxAct = f(xAct, rxAct)$. Hence the control solution is equal to $uxAct = g(sxAct)$, the function of finding a relevant solution depending on the state. Then the final model of the company looks like this: $mAct = (g(f(xAct, rxAct)), com)$.

Solution algorithm:

Step 1. Clarification by a top manager of a distributed company of a new operational task.

Step 2. Development functions f and g for each actor, which include reducing the amount of data, building the state of the company and making a management decision.

Step 3. Integration functions f and g into the decision support system lifecycle.

Step 4. Regular survey of sensors and the use of functions f and g for the synthesis of states (S) for all actors at level C. The survey period is set depending on the importance of the task and the scale of the consequences of the missed situation.

Step 5. Sending a package of states (S) and data from actors C to actors B or A, depending on the degree of violation of restrictions C.

Step 6. Assessment of the state (S). In case of a correct state, go to step 7. In case of an incorrect state, go to step 4.

Step 7. Making decisions (U) at levels C *or* B *or* A.

Step 8. Go to step 4.

The constructed models and algorithms make it possible to construct a system using the principle of data reduction to a level that ensures decision-making without loss of their quality.

IV. Practical Example Of Data Reduction Methods In Decision Support System

Example of the work of a decision-making system will consider in the framework of simulation modeling. Let the task be to analyze the financial condition of the company's divisions based on various statistical data. It is necessary to determine the financial condition of each division and make an appropriate management decision regarding this division.

Select required attributes using features selection method and apply to it PCA. Schematic view of the PCA workflow is shown in Fig. 5.

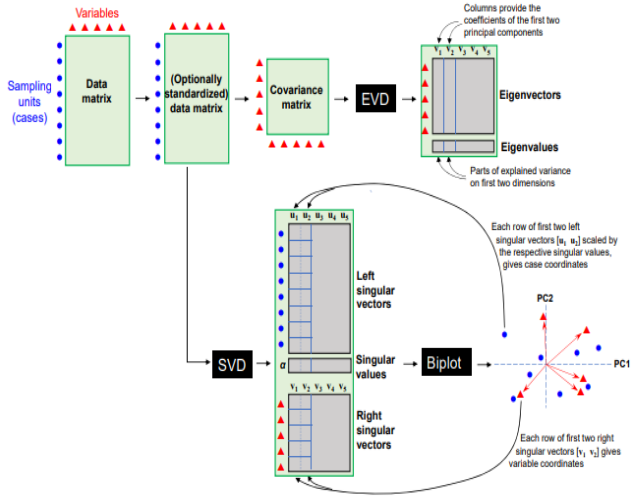

Fig.5 Schematic view of the PCA workflow for company

The determination of the principal components (PCs) can be obtained using the eigenvalue decomposition (EVD) of the covariance matrix of variables, as well as the use of singular values (SVD), which lead directly to the determination of the states of the company's branches.

Schematic view of dimension reduction in PCA in Fig.6.

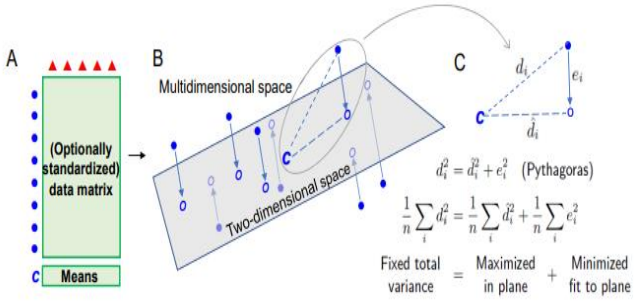

Fig.6 Schematic view of dimension reduction in PCA for company

The distances between the projected points in a PCA are approximating the Euclidean distances between the points in the full space. PCA shorts the data variance into the major features in the data on the leading dimensions and what is considered random noise on the minor dimensions [8].

Based on the data obtained, we visualize the results obtained Fig.7.

As a result, we get the following management decisions: close the red branches, send additional resources to the yellow ones and continue monitoring the green ones. The overall financial condition of the company is at an average level, frequent monitoring of branches and a change in the concept of work is required.

TABLE I. THE RESULT OF THE ANALYSIS OF THE FINANCIAL CONDITION OF THE COMPANY'S BRANCHES

Type of income	Principal Components		
	PC1	*PC2*	*PC3*
stock income	0.825	-0.295	0.303
revenue from sales	0.862	-0.269	0.002
donations	0.764	0.285	0.178

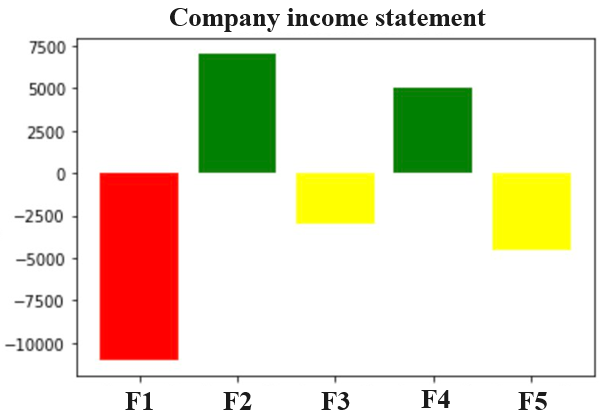

Fig.7 Results for company branches

V. Conclusion

The paper discusses a combined method of using the method of analysis of the main components and the selection of features. Reduction of the initial big data at the object level made it possible to ensure the prompt response of the Center to negative changes at the object level and to supplement the range of planned tasks with a range of operational tasks.

At all levels, the time lag of management decisions was reduced and a holistic view of the company as a whole was quickly formed and strategic decisions were made reasonably.

References

[1] V. V. Krasnoproshin, H. E. R. M. Vissia, Decision-Making and Big Data, Las Nuevas Areas del Poder Economico Mundial. XII Acto Internacional de la Real Academia de Ciencias Economicas y Financieras, Barcelona, Royal Academy of Economy and Financial Sciences, 2017, pp. 105– 120

[2] V. Krasnoproshin, V. Obraztsov, H. Vissia, Solution of applied problems: formalization, methodology and justification, World Scientific Proceeding Series on Computer Engineering and Information Science, vol. 3. "Computational Intelligence in Business and Economics", London, World Scientific, 2010, pp. 57–64.

[3] B. Escofier, Multiple factor analysis, Computer Stat Data analysis, 1994, pp. 121–140

[4] E. Hekler, Why we need a small data paradigm, vol.10, "BMC Medicine", 2019, pp.1–9.

[5] D. Estrin, Small data, where N = me, vol.57, Commun ACM, 2014, 4–32

[6] H. Abdi, L. Williams, Principal component analysis, WIREs Comp. Stat Data analysis, 2010, 433–459

[7] J. Hosse, F. Husson, Selecting the number of principal components analysis using cross-validation approximations, Computer Stat Data analysis, 2012, pp. 1869–1879

[8] L. Platt, N. Malcolm, The Decision Intelligence Handbook: Practical Steps for Evidence-Based Decisions in a Complex World, O'Reilly Media, 2023, pp. 127–14